**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA KẾ TOÁN – KIỂM TOÁN**

---------\*\*\*---------



**BÁO CÁO MÔN HỌC**

**KINH DOANH THÔNG MINH**

**ĐỀ TÀI**

**PHÂN TÍCH HÀNH VI MUA SẮM TRỰC TUYẾN DỰA TRÊN MÔ HÌNH HỌC MÁY**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **TS.Trần Hùng Cường** |
| **Lớp** | **20232IT6088001** |
| **Sinh viên thực hiện** | **Nguyễn Lệ Hằng** |

**Hà Nội, 5/2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA KẾ TOÁN – KIỂM TOÁN**

---------\*\*\*---------



**BÁO CÁO MÔN HỌC**

**KINH DOANH THÔNG MINH**

**ĐỀ TÀI**

**PHÂN TÍCH HÀNH VI MUA SẮM TRỰC TUYẾN DỰA TRÊN MÔ HÌNH HỌC MÁY**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **TS.Trần Hùng Cường** |
| **Sinh viên thực hiện**  Nguyễn Lệ Hằng | **MSV**  2021606898 |

**Hà Nội, 5/2024**

# LỜI CẢM ƠN

Trước hết, em xin bày tỏ lòng biết ơn sâu sắc đến các thầy cô khoa Kế toán - Kiểm toán và các thầy cô khoa Công nghệ thông tin đã hỗ trợ và giúp đỡ em trong suốt quá trình thực hiện và hoàn thành báo cáo này. Em đặc biệt cảm ơn Thầy Trần Hùng Cường, người đã tận tình hướng dẫn, cung cấp tài liệu và góp ý quý báo cho em trong suốt quá trình nghiên cứu và hoàn thành báo cáo. Sự hỗ trợ và định hướng của thầy là nguồn động lực lớn lao giúp em vượt qua những khó khăn và đặt được kết quả tốt nhất.Tuy nhiên, vì kiến thức chuyên môn vẫn còn nhiều hạn chế cùng với việc em chưa có nhiều kinh nghiệm thực tế nên bài báo cáo vẫn còn nhiều điều thiếu sót. Kính mong nhận được sự góp ý và chỉ bảo của quý thầy cô và mọi người.

Em xin chân thành cảm ơn!

# MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc170814529)

[MỤC LỤC 2](#_Toc170814530)

[DANH MỤC HÌNH ẢNH 4](#_Toc170814531)

[DANH MỤC BẢNG BIỂU 5](#_Toc170814532)

[LỜI MỞ ĐẦU 6](#_Toc170814533)

[CHƯƠNG 1 - TỔNG QUAN VỀ KINH DOANH THÔNG MINH 8](#_Toc170814534)

[**1.1. Khái niệm kinh doanh thông minh 8**](#_Toc170814535)

[**1.2. Lịch sử của Kinh doanh thông minh 8**](#_Toc170814536)

[**1.3. Lợi ích của Kinh doanh thông minh 8**](#_Toc170814537)

[**1.4. Mô hình kinh doanh thông minh 9**](#_Toc170814538)

[**1.5. Mối liên kết giữa Kinh doanh Thông minh và Phân tích Hành vi Mua sắm Trực tuyến 11**](#_Toc170814539)

[CHƯƠNG 2 - CƠ SỞ LÝ THUYẾT VÀ MÔ HÌNH NGHIÊN CỨU 12](#_Toc170814540)

[**2.1. Cơ sở lý thuyết 12**](#_Toc170814541)

[**2.1.1. Các công trình nghiên cứu trước đó 12**](#_Toc170814542)

[**2.1.2. Tổng quan về mua sắm trực tuyến 14**](#_Toc170814543)

[**2.1.3. Các nhóm khách hàng trực tuyến 14**](#_Toc170814544)

[**2.1.4. Quy trình mua sắm trực tuyến 15**](#_Toc170814545)

[**2.2. Kỹ thuật phân cụm trong Khai phá dữ liệu 16**](#_Toc170814546)

[**2.2.1. Khái niệm phân cụm 16**](#_Toc170814547)

[**2.2.2. Phân loại các kỹ thuật phân cụm 17**](#_Toc170814548)

[**2.3. Tổng quan về thuật toán K-means 17**](#_Toc170814549)

[**2.3.1. Khái niệm về thuật toán K-means 17**](#_Toc170814550)

[**2.3.2. Cấu trúc K-means 17**](#_Toc170814551)

[**2.3.3. Quy trình xây dựng K-means 18**](#_Toc170814552)

[**2.4. Tổng quan về DBSCAN 20**](#_Toc170814553)

[**2.4.1. Khái niệm về DBSCAN 20**](#_Toc170814554)

[**2.4.2 Cấu trúc của DBSCAN 21**](#_Toc170814555)

[**2.4.2. Quy trình xây dựng DBSCAN 22**](#_Toc170814556)

[CHƯƠNG 3 - CHƯƠNG 3. MÔ HÌNH PHÂN TÍCH HÀNH VI MUA SẮM TRỰC TUYẾN 24](#_Toc170814557)

[**3.1. Thu thập Dữ liệu 24**](#_Toc170814558)

[**3.2. Tiền xử lý Dữ liệu 24**](#_Toc170814559)

[**3.2.1. Xử lí các giá trị thiếu 24**](#_Toc170814560)

[**3.2.2. PCA 25**](#_Toc170814561)

[**3.3. Kết quả Mô hình 28**](#_Toc170814562)

[**3.3.1. Mô hình K-means 28**](#_Toc170814563)

[**3.3.2. Mô hình DBSCAN 34**](#_Toc170814564)

[**3.3.3. So sánh và Đánh giá 38**](#_Toc170814565)

[**3.4. Ứng dụng kết quả mô hình K-means 40**](#_Toc170814566)

KẾT LUẬN .44

TÀI LIỆU THAM KHẢO …45

# DANH MỤC HÌNH ẢNH

[Hình 2.1 Ý định mua sắm trực tuyến 12](#_Toc170821143)

[Hình 2.2 Lý thuyết hành vi có hoạch định (TPB) 13](#_Toc170821144)

[Hình 2.3 Mô hình nghiên cứu 13](file:///C:\Users\Admin\Documents\Zalo%20Received%20Files\Bao_cao_KPDLL%20(1).docx#_Toc170821145)

[Hình 2.4 Quy trình bán lẻ trực tuyến 16](#_Toc170821146)

[Hình 2.5 Phân loại quá trình phân cụm dữ liệu 17](#_Toc170821147)

[Hình 2.6 Mô tả thuật toán K-means 20](#_Toc170821148)

Hình 2.7 Hình vẽ mô phỏng thể hiện ba loại điểm trong DBSCAN 22

[Hình 2.8 Sơ đồ thuật toán DBSCAN cơ bản 23](#_Toc170821149)

[Hình 3.1 Xử lyí các giá trị thiếu 24](file:///C:\Users\Admin\Documents\Zalo%20Received%20Files\Bao_cao_KPDLL%20(1).docx#_Toc170821150)

[Hình 3.2 Biểu đồ PCA hiển thị phần tram phương sai 25](file:///C:\Users\Admin\Documents\Zalo%20Received%20Files\Bao_cao_KPDLL%20(1).docx#_Toc170821151)

[Hình 3.3 Biểu đồ % Explained Variance 26](#_Toc170821152)

[Hình 3.4 Sử dụng phương pháp Elbow 28](#_Toc170821153)

[Hình 3.5 Biểu đồ sử dụng phương pháp Elbow 29](#_Toc170821154)

[Hình 3.6 Đánh giá phân cụm 30](#_Toc170821155)

[Hình 3.7 Biểu đồ thanh mô ta sự phân bổ khách hàng 30](#_Toc170821156)

[Hình 3.8 Phân tích hồ sơ khách hàng bằng biểu đồ Radar 33](#_Toc170821157)

[Hình 3.9 Biểu đồ K-distance 34](file:///C:\Users\Admin\Documents\Zalo%20Received%20Files\Bao_cao_KPDLL%20(1).docx#_Toc170821158)

[Hình 3.10 Xây dựng mô hình DBSCAN 35](#_Toc170821159)

[Hình 3.11 Kết quả mô hình DBSCAN 37](file:///C:\Users\Admin\Documents\Zalo%20Received%20Files\Bao_cao_KPDLL%20(1).docx#_Toc170821160)

[Hình 3.12 Top 10 sản phẩm bán chạy nhất 41](#_Toc170821161)

[Hình 3.13 Top 10 sản phẩm có doanh số cao nhất 41](#_Toc170821162)

[Hình 3.14 Số lượng sản phẩm và doanh thu qua các tháng 42](#_Toc170821163)

[Hình 3.15 Top 10 khách hàng phổ biến 43](#_Toc170821164)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1 Các chỉ số đánh giá 31](#_Toc170821181)

[Bảng 4.2 Bảng so sánh giữa hai mô hình 38](file:///C:\Users\Admin\Documents\Zalo%20Received%20Files\Bao_cao_KPDLL%20(1).docx#_Toc170821182)

# LỜI MỞ ĐẦU

**1.Tính cấp thiết của đề tài**

Mua sắm trực tuyến hoặc mua sắm qua internet đang tăng lên từng ngày. Với sự phát triển của công nghệ hiện đại, thương mại điện tử ngày càng phát triển. Mọi người ngày nay thích mua sắm trực tuyến vì nó tiết kiệm thời gian, năng lượng và tiền bạc. Nhờ sự phát triển của Internet mà thương mại điện tử đã ra mắt và ngày càng phát triển. Sự phát triển của TMĐT đã mang lại nhiều lợi ích cho người tiêu dùng, doanh nghiệp và nền kinh tế. Đối với người tiêu dùng, TMĐT mang lại sự tiện lợi, đa dạng về lựa chọn và giá cả cạnh tranh. Đối với doanh nghiệp, TMĐT mở ra cơ hội tiếp cận thị trường rộng lớn hơn và gia tăng doanh thu. Vì vậy, trước sự phát triển ngày càng tăng của thương mại điện tử, cần tìm hiểu hành vi mua sắm trực tuyến và thúc đẩy doanh số bán hàng. Nâng cao hiệu quả của các chiến lược tiếp thị đóng vai trò quan trọng trong thúc đẩy doanh số bán hàng. Đối với nền kinh tế, TMĐT góp phần thúc đẩy tăng trưởng kinh tế, tạo việc làm và nâng cao năng suất lao động.

Trong dự án này, tôi đi sâu vào lĩnh vực bán lẻ trực tuyến đang phát triển mạnh bằng cách phân tích tập dữ liệu giao dịch từ một doanh nghiệp có trụ sở tại Vương quốc Anh, có sẵn tại Kho lưu trữ máy học của UCI. Tập dữ liệu này ghi lại tất cả các giao dịch từ năm 2010 đến năm 2011.

Thúc đẩy doanh số bán hàng là mục tiêu cốt lõi của mọi doanh nghiệp, nhằm gia tăng lợi nhuận và mở rộng thị phần. Theo một nghiên cứu của McKinsey, các doanh nghiệp áp dụng hệ thống đề xuất sản phẩm dựa trên dữ liệu đã ghi nhận mức tăng trưởng doanh số trung bình từ 10% đến 30% . Bằng cách áp dụng các kỹ thuật phân tích dữ liệu tiên tiến và hệ thống đề xuất sản phẩm, doanh nghiệp có thể gợi ý chính xác các sản phẩm phù hợp với từng khách hàng, từ đó tăng tỷ lệ chuyển đổi và giá trị đơn hàng. Đồng thời, việc cá nhân hóa trải nghiệm mua sắm và triển khai các chiến dịch tiếp thị hiệu quả giúp xây dựng lòng trung thành của khách hàng, khuyến khích họ quay lại mua sắm thường xuyên hơn. Nhờ đó, doanh số bán hàng không ngừng tăng trưởng, góp phần nâng cao vị thế cạnh tranh của doanh nghiệp trên thị trường.

Dựa trên điều này, tôi dự định phát triển nâng cao hiệu quả tiếp thị và chất lượng sản phẩm từ đó hiểu được hành vi khách hàng.

**2. Đối tượng và nội dung phân tích**

**2.1 Đối tượng phân tích**

Đề tài tập trung phân tích hành vi mua sắm trực tuyến dựa trên mô hình học máy

**2.2 Nội dung phân tích**

Báo cáo tập trung phân tích hiệu quả của các chiến lược tiếp thị và chất lượng sản phẩm của từng phân khúc khách hàng thông qua hai thuật toán K-means và DBSCAN để hiểu được hành vi mua sắm trực tuyến của từng phân khúc khách hàng đó.

**3. Mục tiêu phân tích**

Mục tiêu chính của tôi là phát triển nâng cao hiệu quả của các chiến lược tiếp thị và chất lượng sản phẩm từ đó hiểu được hành vi khách hàng thông qua phân khúc khách hàng. Điều này không chỉ giúp nhà bán lẻ tiếp cận đúng đối tượng mà còn tối ưu hóa các nguồn lực, giảm thiểu lãng phí và tăng cường sự hài lòng của khách hàng

**4. Phương pháp phân tích**

Tôi lựa chọn sử dụng mô hình K-means và mô hình DBSCAN chia khách hàng thành các nhóm có tính đồng nhất dựa trên các đặc điểm và hành vi mua sắm của họ. Quá trình này giúp tôi hiểu rõ hơn về nhu cầu và ưu tiên của từng nhóm khách hàng.

**5. Cấu trúc báo cáo**

Ngoài phần mở đầu, kết luận, Tài liệu tham khảo, và Phụ lục, nội dung của báo cáo được trình bày trong 3 chương:

Chương 1. Tổng quan về kinh doanh thông minh

Chương 2. Cơ sở lý thuyết và mô hình nghiên cứu

Chương 3. Mô hình phân tích hành vi mua sắm trưc tuyến

# TỔNG QUAN VỀ KINH DOANH THÔNG MINH

## Khái niệm kinh doanh thông minh

**Kinh doanh thông minh (Business Intelligence)** là một tập hợp các quy trình, kiến trúc và công nghệ chuyển đổi dữ liệu thô thành thông tin có ý nghĩa, nhằm thúc đẩy các hoạt động kinh doanh có lợi. Kinh doanh thông minh kết hợp giữa phân tích kinh doanh, khai thác dữ liệu, trực quan hóa dữ liệu, sử dụng công cụ dữ liệu và cơ sở hạ tầng cũng như các phương pháp hay nhất để giúp các tổ chức đưa ra quyết định dựa trên dữ liệu nhiều hơn. Kinh doanh thông minh bao gồm một bộ phần mềm và sản phẩm dịch vụ để chuyển đổi dữ liệu thành thông tin và kiến thức hữu ích.

## Lịch sử của Kinh doanh thông minh

Thuật ngữ “kinh doanh thông minh” đã tồn tại trong hàng thập niên, nhưng nó được dùng lần đầu với ý nghĩa của ngày hôm nay vào năm 1988 bởi Howard Dresner.Ngày nay, Kinh doanh thông minh được định nghĩa bởi Forrester là một tập có phương pháp, quy trình, kiến trúc và công nghệ mà chuyển dữ liệu thô thành thông tin hữu ích và có giá trị được sử dụng để nâng cao các hiểu biết sâu sắc và đưa ra quyết định kinh doanh về hoạt động, chiến lược và chiến thuật của doanh nghiệp.

## Lợi ích của Kinh doanh thông minh

Mang lại nhiều lợi ích cho các tổ chức bằng cách cho phép họ đưa ra quyết định dựa trên dữ liệu, tối ưu hóa hoạt động và duy trì lợi thế cạnh tranh. Một số lợi ích chính của BI bao gồm:

* **Cải thiện việc ra quyết định:** Bằng cách cung cấp thông tin chi tiết kịp thời và chính xác, kinh doanh thông minh giúp những người ra quyết định đưa ra những lựa chọn sáng suốt hơn, giảm sự phụ thuộc vào cảm tính hoặc phỏng đoán.
* **Nâng cao hiệu quả hoạt động:** Các công cụ kinh doanh thông minh có thể xác định sự thiếu hiệu quả, tắc nghẽn hoặc hạn chế tài nguyên, cho phép các tổ chức hợp lý hóa các quy trình, giảm chi phí và tối ưu hóa phân bổ tài nguyên.
* **Tăng doanh thu và lợi nhuận:** Với thông tin chi tiết về sở thích của khách hàng, xu hướng thị trường và hiệu suất bán hàng, các doanh nghiệp có thể điều chỉnh các dịch vụ, chiến lược giá và chiến dịch tiếp thị của mình, cuối cùng là thúc đẩy tăng trưởng doanh thu và lợi nhuận cao hơn.
* **Hiểu khách hàng tốt hơn:** Cho phép các tổ chức phân tích dữ liệu khách hàng, xác định các mẫu và xu hướng giúp điều chỉnh các sản phẩm, dịch vụ và nỗ lực tiếp thị để đáp ứng nhu cầu của khách hàng và nâng cao sự hài lòng của khách hàng.

## Mô hình kinh doanh thông minh

Mô hình kinh doanh thông minh là một tập hợp các quy trình và công nghệ giúp tổ chức thu thập, tích hợp, phân tích và trình bày dữ liệu để hỗ trợ ra quyết định và cải thiện hiệu suất. Mô hình BI điển hình bao gồm các thành phần chính sau:

**1. Data Sources (Nguồn dữ liệu)**

Nguồn dữ liệu là nơi thông tin được thu thập để phân tích và ra quyết định. Các nguồn dữ liệu phổ biến bao gồm:

**Cơ sở dữ liệu quan hệ (RDBMS)**: MySQL, PostgreSQL, Oracle, SQL Server.

**Hệ thống ERP và CRM**: SAP, Salesforce.

**Dữ liệu web và mạng xã hội**: Google Analytics, Facebook, Twitter.

**Dữ liệu giao dịch**: Từ hệ thống POS, hệ thống thanh toán.

**Dữ liệu từ đối tác và bên thứ ba**: Dữ liệu thị trường, dữ liệu công khai, dữ liệu nghiên cứu.

**2. Data Integration Tools (Công cụ tích hợp dữ liệu)**

Công cụ tích hợp dữ liệu giúp kết nối và hợp nhất dữ liệu từ nhiều nguồn khác nhau:

**ETL (Extract, Transform, Load)**: Trích xuất dữ liệu từ các nguồn, chuyển đổi dữ liệu để phù hợp với phân tích và tải dữ liệu vào kho dữ liệu.

**Công cụ ETL**: Talend, Informatica, Apache NiFi.

**Data Virtualization**: Kết hợp dữ liệu từ nhiều nguồn mà không cần di chuyển hoặc sao chép dữ liệu.

**Công cụ Data Virtualization**: Denodo, TIBCO Data Virtualization.

**3. Data Warehouse (Kho dữ liệu)**

Kho dữ liệu là hệ thống lưu trữ dữ liệu lớn từ nhiều nguồn, được tổ chức và tối ưu hóa cho việc phân tích và báo cáo:

**Công nghệ Data Warehouse**: Amazon Redshift, Google BigQuery, Snowflake.

**Data Marts**: Phân vùng của kho dữ liệu, chứa dữ liệu liên quan đến một bộ phận hoặc chức năng cụ thể.

**4. OLAP Cubes and Dicing (Khối OLAP và phân tích)**

OLAP (Online Analytical Processing) cung cấp khả năng phân tích dữ liệu đa chiều, cho phép người dùng xem dữ liệu từ nhiều góc độ khác nhau:

**OLAP Cubes**: Các cấu trúc dữ liệu đa chiều, tối ưu hóa cho việc truy vấn và phân tích nhanh.

**Dicing**: Cắt nhỏ và phân tích dữ liệu theo các chiều khác nhau để khám phá các mẫu và xu hướng.

**5. Business Analytics Tools (Công cụ phân tích kinh doanh)**

Công cụ phân tích kinh doanh giúp người dùng khám phá, phân tích và trực quan hóa dữ liệu:

**Công cụ phân tích dữ liệu**: R, Python, SAS.

**Data Visualization**: Tableau, Power BI, Tạo các báo cáo, biểu đồ và dashboard để trực quan hóa dữ liệu.

## Mối liên kết giữa Kinh doanh thông minh và Phân tích hành vi mua sắm trực tuyến

Kinh doanh thông minh và phân tích hành vi mua sắm trực tuyến có mối liên hệ mật thiết, trong đó kinh doanh thông minh cung cấp các công cụ và phương pháp để thu thập, phân tích và diễn giải dữ liệu từ các hoạt động mua sắm trực tuyến. Mối liên kết này giúp doanh nghiệp hiểu rõ hơn về hành vi khách hàng, tối ưu hóa trải nghiệm mua sắm và nâng cao hiệu quả kinh doanh.

Việc hiểu được kinh doanh thông minh có thể cung cấp những thông tin quan trọng cho phân tích hành vi mua sắm trực tuyến.

**Hiểu rõ hơn về khách hàng**:

Cung cấp cái nhìn sâu sắc về hành vi và sở thích của khách hàng thông qua phân tích dữ liệu từ các hoạt động trực tuyến.

Thông tin này giúp doanh nghiệp tùy chỉnh các chiến dịch marketing, cải thiện sản phẩm và dịch vụ để đáp ứng nhu cầu của khách hàng.

**Tối ưu hóa trải nghiệm mua sắm**:

Giúp xác định các điểm khó khăn trong quá trình mua sắm và đưa ra các biện pháp cải thiện trải nghiệm người dùng.

Ví dụ, dữ liệu về thời gian tải trang, tỉ lệ bỏ giỏ hàng và các hành vi khác có thể được sử dụng để tối ưu hóa website và ứng dụng di động.

**Dự đoán xu hướng và nhu cầu thị trường**:

Sử dụng các mô hình dự đoán để xác định các xu hướng mua sắm trong tương lai, giúp doanh nghiệp chuẩn bị sẵn sàng cho các biến động của thị trường.

Dự đoán nhu cầu sản phẩm và xu hướng tiêu dùng giúp doanh nghiệp quản lý kho hàng hiệu quả hơn và tối ưu hóa chuỗi cung ứng.

**Kết luận:** Chương 1 của nghiên cứu đã cung cấp cái nhìn tổng thể về khái niệm, lịch sử phát triển, lợi ích và các mô hình liên quan đến kinh doanh thông minh. Ngoài ra còn cung cấp cái nhìn tổng quát về mối liên kết giữa Kinh doanh thông minh và Hành vi mua sắm trực tuyến.

# CƠ SỞ LÝ THUYẾT VÀ MÔ HÌNH NGHIÊN CỨU

## Cơ sở lý thuyết

### Các công trình nghiên cứu trước đó

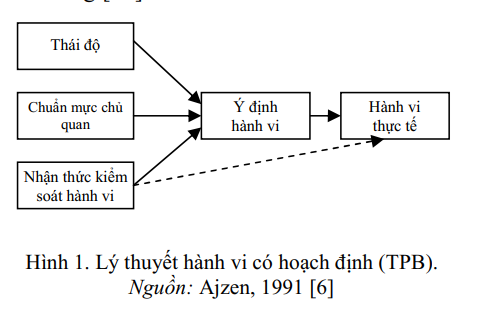
Nghiên cứu của hai tác giả Nguyễn Thị Bảo Châu & Lê Nguyễn Xuân Đào (2014), các nhân tố niềm tin, sự đa dạng về lựa chọn hàng hóa, giá cả, tính thích hợp của trang web, sự thoải mái, sự thân thiện, các nhân tố rủi ro về thời gian, rủi ro về sản phẩm, rủi ro về tài chính có ảnh hưởng đến quyết định mua sắm trực tuyến của người tiêu dùng. Nghiên cứu chỉ ra rằng nhân tố sự thoải mái có vai trò quan trọng nhất đến quyết định mua sắm trực tuyến. Hai nhóm đối tượng chưa từng mua sắm trực tuyến, hai nhóm chịu tác động mạnh bởi nhân tố sự đa dạng trong việc lựa chọn hàng hóa, và đây là nhân tố phân biệt hai nhóm đối tượng mua sắm trực tuyến.

Trong nghiên cứu của tác giả Nguyễn Thị Kim Hiệp (2022) đã xây dựng mô hình nghiên cứu dựa trên các mô hình lý thuyết, mô hình về hành vi mua sắm trực tuyến, ý định mua hàng của người tiêu dùng, cùng với sự tham khảo các nghiên cứu khoa học của các tác giả trước đây và cả những ý kiến đóng góp để nghiên cứu các yếu tố ảnh hưởng đến ý định mua sắm trực tuyến của người tiêu dùng thế hệ Z trên địa bàn thành phố Biên Hòa (Hình …)

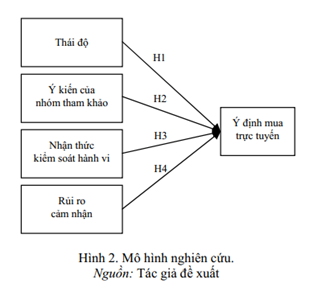
Hình 2.1 Ý định mua sắm trực tuyến

Kết quả nghiên cứu này cho thấy có 3 yếu tố trong mô hình lý thuyết có ảnh hưởng đến ý định mua sắm trực tuyến của người tiêu dùng thế hệ Z trên địa bàn thành phố Biên Hòa được sắp xếp theo trình tự mức độ quan trọng từ cao xuống thấp, đó là: Mức độ Uy tín, Cảm nhận về chất lượng và An toàn.

Hạn chế của nghiên cứu này là chưa kiểm định được mối quan hệ tương tác giữa các yếu tố ảnh hưởng đến ý định mua sắm trực tuyến của người tiêu dùng thành phố Hồ Chí Minh trong giai đoạn Covid-19. Để kiểm định mối quan hệ giữa các yếu tố trong mô hình nghiên cứu, nên sử dụng công cụ kiểm định mô hình cấu trúc tuyến tính SEM sẽ cho kết quả nghiên cứu chính xác cao hơn.

Nghiên cứu Các yếu tố ảnh hưởng đến ý định mua sắm trực tuyến của người tiêu dùng Việt Nam: Nghiên cứu mở rộng thuyết hành vi có hoạch định của hai tác giả Hà Ngọc Thắng & Nguyễn Thành Độ (2006)

Hình 2.2 Lý thuyết hành vi có hoạch định (TPB)

Đóng góp chính của nghiên cứu này là đã phát triển lý thuyết TPB bằng cách bổ sung yếu tố rủi ro cảm nhận trong nghiên cứu ý định mua sắm trực tuyến của người tiêu dùng.

Hình 2.3 Mô hình nghiên cứu

Kết quả nghiên cứu cho thấy, ý định mua sắm trực tuyến của người tiêu dùng bị ảnh hưởng bởi thái độ của người tiêu dùng đối với trang web, nhận thức kiểm soát hành vi và rủi ro cảm nhận.

### Tổng quan về mua sắm trực tuyến

#### Khái niệm Thương mại điện tử

Thương mại điện tử (E-Commerce) là hình thức kinh doanh trực tuyến sử dụng nền tảng công nghệ thông tin với sự hỗ trợ của Internet để thực hiện các giao dịch mua bán, trao đổi, thanh toán trực tuyến . Với sự thay đổi nhanh chóng trong hành vi của người tiêu dùng, thương mại điện tử (TMĐT) đã trở thành một phương thức hiệu quả để phát triển và mở rộng kinh doanh. Hiện nay, nhiều doanh nghiệp đã tích cực triển khai mô hình kinh doanh trên nền tảng thương mại điện tử. Mặc dù thương mại điện tử bắt đầu xuất hiện dưới hình thức giao dịch B2B (Business-to-Business) vào cuối những năm 1970, nhưng phân khúc B2C (Business-to-Customer) lại thu hút nhiều sự quan tâm hơn. Thương mại điện tử đang trải qua giai đoạn phát triển và thu hút sự chú ý đặc biệt tại Việt Nam.

#### Khái niệm mua sắm trực tuyến

Mua sắm trực tuyến (Online shopping) là một trong những hình thức của TMĐT, ngày càng trở nên quan trọng trong phát triển thương mại điện tử B2C từ cuối thế kỳ XX.

Mua sắm trực tuyến là quá trình người tiêu dùng trực tiếp mua hàng hóa, dịch vụ từ một người bán trong thời gian xác định thông qua Internet mà không có một dịch vụ trung gian nào. Mua sắm trực tuyến cũng là một tiến trình dùng để liệt kê hàng hóa và dịch vụ cùng với hình ảnh kèm theo được hiển thị từ xa thông qua các phương tiện điện tử. Khi sản phẩm hoặc dịch vụ được chọn, giao dịch sẽ được thực hiện một cách tự động bằng việc thanh toán trực tuyến hoặc thanh toán bằng tiền mặt.

### Các nhóm khách hàng trực tuyến

Các khách hàng trực tuyến thường được chia thành năm nhóm, gồm: Người mua sắm ngẫu hứng, Khách hàng tiềm năng, Khách hàng - nhà sản xuất, Người săn lùng ưu đãi và khách hàng trung thành.

* **Người mua sắm ngẫu hứng (Impulse Shoppers)**

Hành vi mua hàng ngẫu hứng là việc mua sắm mà người tiêu dùng thực hiện đột ngột và không có kế hoạch trước. Trong tình huống này, người mua quyết định mua sản phẩm hoặc dịch vụ mà họ thường không có ý định mua, thường là dưới tác động của cảm xúc hoặc sự kích thích tại thời điểm đó.

* **Khách hàng tiềm năng (Potential Customer)**

Khách hàng tiềm năng là những cá nhân hoặc nhóm người có sự quan tâm đến sản phẩm/ dịch vụ mà doanh nghiệp cung cấp, đồng thời có khả năng chi trả cho chúng. Những khách hàng này thường đã có mong muốn, nhu cầu về sản phẩm, dịch vụ nhưng chưa quyết định mua. Hay nói cách khác, họ là những người hoặc tổ chức có khả năng trở thành khách hàng của doanh nghiệp trong tương lai gần.

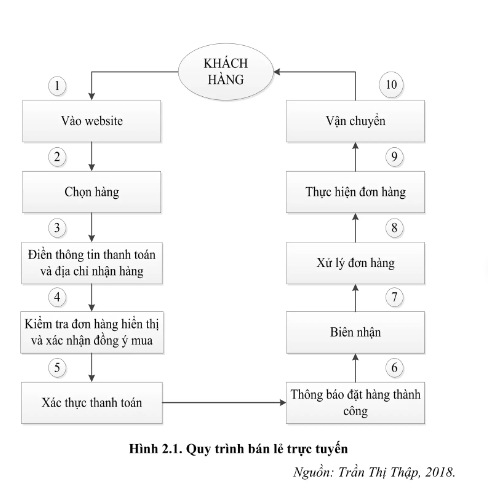
* **Người săn lùng ưu đãi (Bargain Hunters)**

Nhóm khách hàng này chủ yếu tìm kiếm các chương trình khuyến mãi, giảm giá, hoặc các ưu đãi đặc biệt. Họ có xu hướng mua hàng khi có các chương trình giảm giá hấp dẫn và luôn chú trọng đến việc tiết kiệm tiền. Những người săn lùng ưu đãi thường xuyên cập nhật thông tin về các đợt giảm giá, sự kiện bán hàng lớn như Black Friday, Cyber Monday, hoặc các chương trình khuyến mãi theo mùa.

* **Khách hàng trung thành (Key online Customers)**

Khách hàng trung thành là những người gắn bó với một hoặc vài thương hiệu cụ thể và thường xuyên quay lại mua sắm tại những nơi họ có trải nghiệm tốt. Họ không chỉ tìm kiếm sản phẩm hoặc dịch vụ chất lượng cao mà còn đánh giá cao sự nhất quán và đáng tin cậy từ phía nhà cung cấp. Khách hàng trung thành thường có mối quan hệ lâu dài với thương hiệu.

### Quy trình mua sắm trực tuyến

Việc phân tích hành vi khách hàng trực tuyến khi sử dụng website để phân biệt các nhóm khách hàng khác nhau. Khi một khách hàng truy cập vào trang web, người đó sẽ trải qua nhiều giai đoạn khác nhau trong hành vi mua. Càng có nhiều kinh nghiệm, các đánh giá của họ về sản phẩm/ dịch vụ điện từ càng sắc sảo. Đồng thời, cùng với sự hỗ trợ và quản trị đối thoại chuyên nghiệp thì lòng tin đối với nhà cung cấp trực tuyến sẽ tăng lên; xác suất mua hàng sẽ cao hơn. Nếu công việc phân phối và giao hàng thành công thi những rào cản tâm lý cuối cùng sẽ được gỡ bỏ và cơ hội thu hút khách hàng trung thành sẽ tăng lên. Quy trình mua sắm trực tuyến xem xét quan hệ thương mại giữa nhà bán lẻ và người tiêu dùng. Một quy trình mua sắm trực tuyến điển hình được mô tả trong **Hình 2.4** như sau:

Hình 2.4 Quy trình bán lẻ trực tuyến

Các bước mua hàng đã nêu ở trên áp dụng chủ yếu cho hàng hóa hữu hình. Trong trường hợp mua dịch vụ như nội dung số, mã điện tử đặt phòng khách sạn, hoặc mã vé máy bay, quy trình có thể được điều chỉnh. Ví dụ, nếu khách hàng đã đăng nhập với tài khoản có sẵn, các bước như đăng nhập hoặc đăng ký tài khoản và nhập thông tin giao hàng có thể được bỏ qua.

## Kỹ thuật phân cụm trong Khai phá dữ liệu (Clustering Techniques in

### Khái niệm phân cụm

Phân cụm dữ liệu (Data clustering) nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định. Bài toán phân cụm nhằm phân tách một tập dữ liệu gồm n phần tử thành k cụm sao cho các phần tử trong mỗi cụm là tương tự nhau và các phần tử nằm khác cụm thì sẽ không tương tự nhau theo một nghĩa nào đó.

**Phân cụm dữ liệu** thuộc lớp bài toán **học không giám sát** (unsupervised learning) trong học máy (machine learning) ở đó các mẫu chưa được gán nhãn khác loại với các thuật toán học có giám sát chẳng hạn như mạng Nơron hay phương pháp Support Vector Machine

### Phân loại các kỹ thuật phân cụm

Hình 2.5 Phân loại quá trình phân cụm dữ liệu

## Tổng quan về thuật toán K-means

### Khái niệm về thuật toán K-means

Thuật toán phân cụm K-means được giới thiệu năm 1957 bởi Lloyd K-means và là phương pháp phổ biến nhất cho việc phân cụm, dựa trên việc phân vùng dữ liệu. Tư tưởng chính của thuật toán K-means là tìm cách gom nhóm các đối tượng đã cho và K cụm (K là số các cụm được xác định trước K nguyên, dương) sao cho khoảng cách từ đối tượng đến tâm (centroid) của các nhóm là nhỏ nhất.

### Cấu trúc K-means

Cấu trúc của K-means bao gồm các thành phần chính sau:

**Centroids (Tâm cụm):** Đại diện cho trung tâm của mỗi cụm. Các centroids được cập nhật lặp đi lặp lại để giảm thiểu khoảng cách giữa các điểm dữ liệu trong cụm và tâm cụm.

**Clusters (Cụm):** Tập hợp các điểm dữ liệu gần nhau về mặt khoảng cách với một centroid nhất định.

### Quy trình xây dựng K-means

**Input:** Tập X = {x1, x2, …, xn} và số cụm K

**Output:** Các cụm Cj (j = 1, 2, …, K)

(thỏa mãn các điều kiện: Cj ≠ Ø ∀j, Cj ∩ Ck = Ø với ∀j ≠ k và )  
**Bước 1:** Khởi tạo

Chọn K đối tượng mj (j = 1, 2, …, k) làm tâm của K cụm (có thể lựa chọn K mẫu đầu tiên, chọn K mẫu ngẫu nhiên, …).

**Bước 2:** Xác định cụm mới

Đối với mỗi đối tượng (1≤ i ≤n), tính khoảng cách từ nó tới trọng tâm của mỗi cụm với j = 1, 2, …, K

Gán đối tượng về cụm mà khoảng cách từ đến tâm cụm đó là nhỏ nhất:

Nếu thì

*Trong đó:* là khoảng cách Euclidean từ tới tâm cụm ở bước lặp t với ∀j = 1, 2, …, K và j ≠ s.

Công thức tính khoảng cách Euclidean:

**Bước 3:** Cập nhật lại trọng tâm

Trọng tâm của mỗi cụm ở lần lặp t + 1 được tính theo công thức:

Với là số phần tử của cụm ở bước lặp t.

**Bước 4:** Kiểm tra tính hội tụ

Lặp bước 2 và bước 3 cho đến khi thỏa mãn điều kiện hội tụ.

Trong đó điều kiện hội tụ của thuật toán có thể là:

(i). Tâm cụm không đổi: tâm cụm ở bước lặp t + 1 chính là tâm cụm ở

bước lặp t

= với j = 1,2, ..., K.

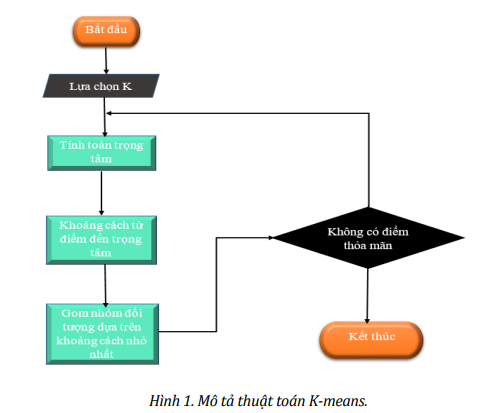
(ii). Đạt được số bước lặp tối đa: t > Maxlt, ở đó t là số lần lặp còn

MaxIt là hằng số quy định số lần lặp tối đa.

(iii). Thỏa mãn điều kiện hàm lỗi, chẳng hạn với khoảng cách Euclidean, ta có thể sử dụng hàm lỗi bình phương (square - error). Lỗi của cụm Cj là tổng khoảng cách Euclidean giữa các mẫu trong Cj và tâm của cụm:

Khi đó độ lỗi đối với K cụm là:

Cần tìm một phân hạch chứa K cụm sao cho



Hình 2.6 Mô tả thuật toán K-means

## Tổng quan về DBSCAN

### Khái niệm về DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise tạm dịch là Phân cụm không gian dựa trên mật độ các ứng dụng với nhiễu) là một thuật toán phân cụm không giám sát, được phát triển bởi Martin Ester, Hans-Peter Kriegel, Jörg Sander và Xiaowei Xu vào năm 1996. DBSCAN không yêu cầu xác định số lượng cụm trước như K-means mà dựa vào khái niệm mật độ điểm dữ liệu để xác định cụm.

Đây là một trong những phương pháp phân cụm phổ biến trong học máy và khai phá dữ liệu, đặc biệt là khi làm việc với dữ liệu không có hình dạng xác định rõ ràng và có nhiễu.

Thuật toán sử dụng hai thông số chính:

**ε (epsilon):** Bán kính lân cận mà trong đó điểm được coi là gần nhau.

**MinPts:** Số lượng điểm tối thiểu cần thiết để tạo thành một cụm.

**Định nghĩa 1:** Vùng lân cận epsilon (Eps-neighborhood) của một điểm dữ liệu **P** là tợp hợp tất cả các điểm dữ liệu nằm trong phạm vi bán kính epsilon (kí hiệu ) xung quanh điểm **P**.

Trong đó: D là tập hợp tất cả các điểm dữ liệu của tập huấn luyện.

**Định nghĩa 2:** Khả năng tiếp cận trực tiếp mật độ (directly density-reachable) đề cập tới việc một điểm có thể tiếp cận trực tiếp tới một điểm dữ liệu khác. Cụ thể là một điểm **Q** được coi là có thể tiếp cận trực tiếp bởi điểm **P** tương ứng với tham số epsilon và minPts nếu như nó thoả mãn hai điều kiện:

**Q** nằm trong vùng lân cận epsilon của **P**: **Q**

Số lượng các điểm dữ liệu nằm trong vùng lân cận epsilon tối thiểu là minPts: **| minPts**

Như vậy một điểm dữ liệu có thể tiếp cận được trực tiếp tới một điểm khác không chỉ dựa vào khoảng cách giữa chúng mà còn phụ thuộc vào mật độ các điểm dữ liệu trong vùng lân cận epsilon phải tối thiểu bằng minPts.

**Định nghĩa 3:** Khả năng tiếp cận mật độ (density-reachable) liên quan đến cách hình thành một chuỗi liên kết điểm trong cụm. Cụ thể là trong một tập hợp chuỗi điểm mà nếu như bất kì một điểm nào cũng đều có thể tiếp cận trực tiếp mật độ (định nghĩa 2) bởi theo tham số epsilon và minPts thì khi đó ta nói điểm **P =**  có khả năng kết nối mật độ tới điểm **Q = .**

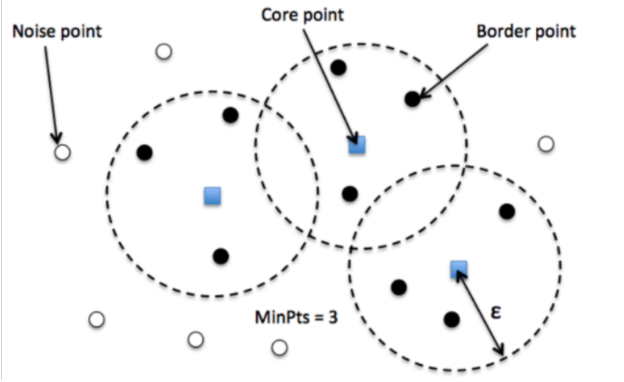
### Cấu trúc của DBSCAN

DBSCAN phân chia các điểm dữ liệu thành ba loại chính:

**Điểm cốt lõi (Core Points):** Các điểm có ít nhất MinPts điểm trong bán kính ε.

**Điểm biên (Border Points):** Các điểm nằm trong vùng ε của một core point nhưng có ít hơn MinPts điểm trong bán kính ε.

**Điểm nhiễu (Noise Points):** Các điểm không phải là core points và không nằm trong vùng ε của bất kỳ core point nào.



***Hình 2.7. Hình vẽ mô phỏng thể hiện ba loại điểm trong DBSCAN***

Cấu trúc của DBSCAN bao gồm các cụm hình thành từ các core points liên thông và các border points xung quanh chúng.

### Quy trình xây dựng DBSCAN

Thuật toán sẽ thực hiện lan truyền để mở rộng dần phạm vi của cụm cho tới khi chạm tới những điểm biên thì thuật toán sẽ chuyển sang một cụm mới và lặp lại tiếp quá trình trên.

**Bước 1 Khởi tạo:**

Bắt đầu với một điểm dữ liệu ngẫu nhiên P.

**Bước 2 Tìm kiếm lân cận:**

Xác định các điểm trong bán kính ε của điểm dữ liệu hiện tại.

**Bước 3 Xác định loại điểm:**

Nếu số lượng điểm trong bán kính ε lớn hơn hoặc bằng MinPts, điểm hiện tại là core point và một cụm mới bắt đầu được hình thành.

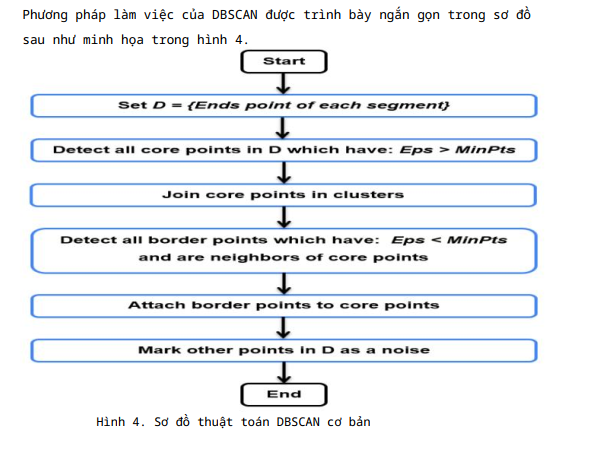
Nếu không, điểm hiện tại là noise hoặc border point.

**Bước 4 Mở rộng cụm:**

Với mỗi core point, mở rộng cụm bằng cách lặp lại quá trình tìm kiếm lân cận cho tất cả các điểm trong vùng lân cận ε và thêm các điểm đủ điều kiện vào cụm.

**Bước 5 Tiếp tục:**

Lặp lại quy trình cho đến khi tất cả các điểm dữ liệu đều được xử lý (được gán vào một cụm hoặc xác định là noise).



Hình 2.8 Sơ đồ thuật toán DBSCAN cơ bản

**Kết luận:** Chương 2 của nghiên cứu đã cung cấp cái nhìn toàn diện về hành vi mua sắm trực tuyến của người tiêu dùng, các nhóm khách hàng và quy trình mua sắm trực tuyến. Nhấn mạnh tầm quan trọng của việc chọn phương pháp phân cụm phù hợp với đặc điểm của dữ liệu, trong đó K-Means thích hợp cho dữ liệu đồng đều, còn DBSCAN phù hợp với dữ liệu có cấu trúc phức tạp và nhiều nhiễu.

# MÔ HÌNH PHÂN TÍCH HÀNH VI MUA SẮM TRỰC TUYẾN

## Thu thập Dữ liệu

Nguồn dữ liệu: Kaggle

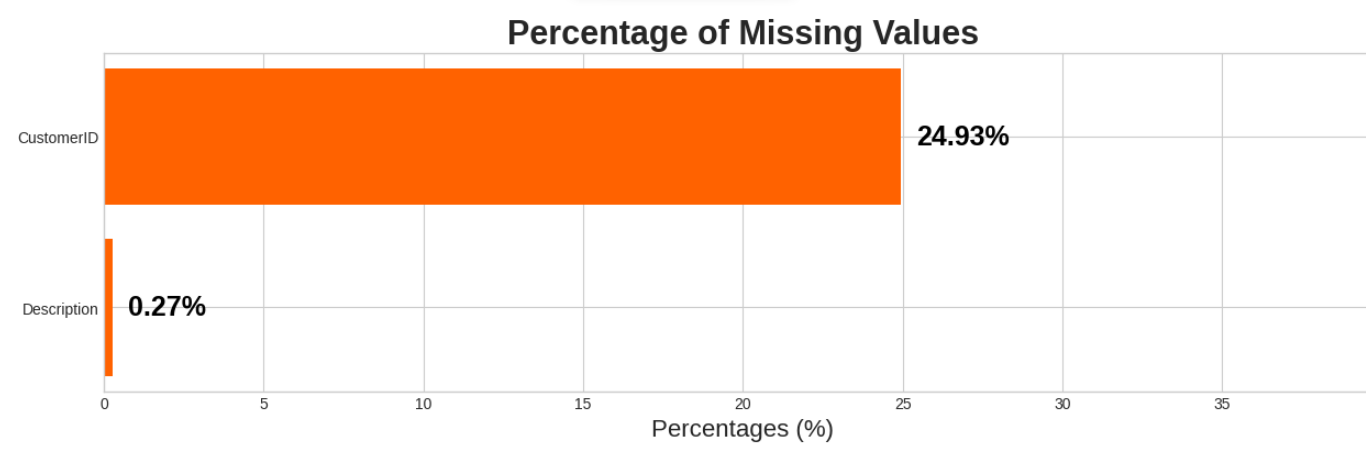
Biến đặc trưng:

* **CustomerID:** ID của mỗi khách hàng
* **recency:** Số ngày đã trôi qua từ lần mua cuối cùng của khách hàng
* **frequency:** Tần suất mua hàng của khách hàng
* **total\_product\_purchased:** Tổng số lượng sản phẩm đã được khách hàng mua trong tất cả các giao dịch
* **monetary:** Tổng số tiền khách hàng đã mua trong tất cả các giao dịch
* **avg\_spend\_a\_trans:** Giá trị trung bình các giao dịch của khách hàng
* **#\_unique\_product\_purchased:** Số lượng sản phẩm khác nhau mà khách hàng đã mua
* **freq\_hour\_buy:** Giờ trong ngày khách hàng thích mua sắm hơn (theo mức 24h)
* **freq\_day\_buy:** Ngày trong tuần khách hàng thích mua sắm hơn (Kí hiệu: 2: thứ 2, 3:thứ 3, ..., 8: CN)
* **is\_uk:** Biến cho biết cửa hàng ở Anh Quốc không?(1: Có, 2: Không)
* **#\_cancelled:** Tổng số đơn hàng mà khách hàng đã hủy.

## Tiền xử lý Dữ liệu

### Xử lí các giá trị thiếu

Đầu tiên tôi sẽ thực hiện phân tích sơ bộ để hiểu cấu trúc và các loại cột dữ liệu:



Hình 3.1 Xử lý các giá trị thiếu

**CustomerID (thiếu 24,93% giá trị)**

Cột CustomerID chứa gần 1/4 dữ liệu bị thiếu. Cột này rất cần thiết để phân nhóm khách hàng và tạo hệ thống đề xuất. Việc đưa ra một tỷ lệ lớn các giá trị bị thiếu như vậy có thể gây ra sai lệch hoặc nhiễu đáng kể trong phân tích.

Hơn nữa, vì việc phân cụm dựa trên hành vi và sở thích của khách hàng nên điều quan trọng là phải có dữ liệu chính xác về số nhận dạng khách hàng. Do đó, việc loại bỏ các hàng bị thiếu CustomerID có vẻ là cách tiếp cận hợp lý nhất để duy trì tính toàn vẹn của các cụm và phân tích.

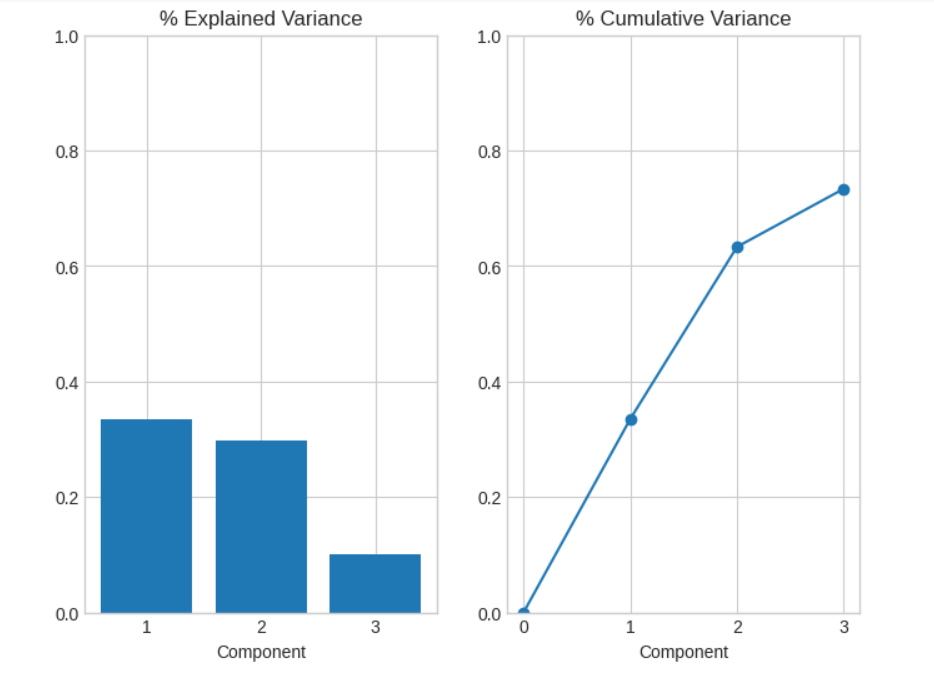
**Description (thiếu 0,27% giá trị)**

Cột Description có một tỷ lệ phần trăm nhỏ các giá trị bị thiếu. Tuy nhiên, người ta nhận thấy rằng có sự mâu thuẫn trong dữ liệu trong đó cùng một mã sản phẩm không phải lúc nào cũng có cùng mô tả. Điều này cho biết các vấn đề về chất lượng dữ liệu và các lỗi tiềm ẩn trong mô tả sản phẩm. Do những mâu thuẫn này, việc đưa ra các mô tả còn thiếu dựa trên Stock Code có thể không đáng tin cậy

### PCA

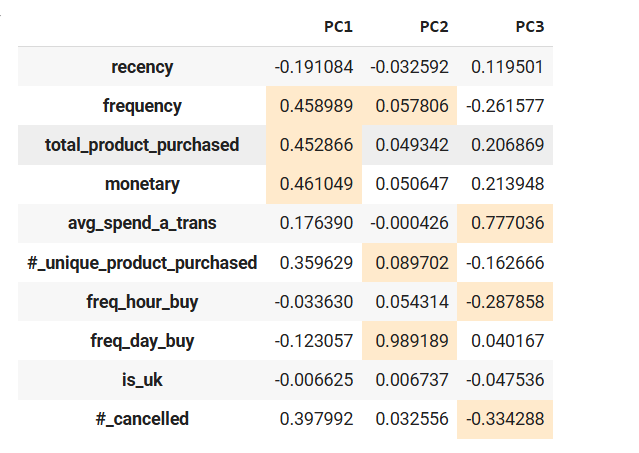
Hình 3.2 Biểu đồ PCA hiển thị phần trăm phương sai

Biểu đồ hiển thị phần trăm phương sai giải thích bởi mỗi thành phần chính. Chúng ta lựa chọn điểm “elbow point" (điểm khuỷu tay) trên biểu đồ, là điểm mà phần trăm phương sai giải thích bắt đầu giảm chậm lại tại n = 3.



Hình 3.3 Biểu đồ % Explained Variance

Biểu đồ % Explained Variance cho thấy các thành phần chính đầu tiên có khả năng giải thích càng nhiều phương sai của dữ liệu gốc. Cụ thể, thành phần chính đầu tiên giải thích khoảng hơn 37% phương sai, thành phần chính thứ 2 giải thích thêm khoảng 26% phương sai, và thành phần chính thứ 3 giải thích khoảng 17% phương sai. Khi xem xét biểu đồ % Cumulative Variance, biểu đồ đường này thể hiện sự tích lũy phương sai theo từng thành phần chính. Ta thấy rằng sử dụng 3 thành phần chính có thể giải thích khoảng 80% tổng phương sai của dữ liệu gốc. Điều này cho thấy 3 thành phần chính là một lựa chọn hợp lý để làm giảm số chiều dữ liệu mà vẫn duy trì được phần lớn thông tin quan trọng. Dựa vào các biểu đồ, ta có thể xác định được số lượng thành phần chính tối ưu cần giữ lại sau khi thực hiện PCA. Trong trường hợp này, 3 thành phần chính là lựa chọn hợp lý, vì chúng có thể giải thích khoảng 70-80% phương sai tổng thể.

***Bảng 3.1 Bảng kết quả PCA***

Kết quả PCA được hiển thị dưới dạng biểu đồ thể hiện các hệ số tương quan giữa từng thuộc tính (như "recency", "frequency", "total\_product\_purchased", "monetary",...) và các thành phần chính (PC1, PC2, PC3).

**Phân tích ma trận tải trọng thành phần**

Thuộc tính "monetary" có tương quan mạnh (0.461049) với PC1, cho thấy thuộc tính này đóng góp đáng kể cho PC1.Thuộc tính "avg\_spend\_a\_trans" có tương quan mạnh hơn với PC3 (0.777036) so với PC1, cho thấy thuộc tính này đóng góp nhiều hơn cho PC3.

PC1 có thể liên quan đến tổng quan về hành vi mua sắm của khách hàng, bao gồm số tiền chi tiêu, tần suất mua hàng và số lượng sản phẩm mua. PC3 có thể liên quan đến các khía cạnh cụ thể hơn của hành vi mua sắm, như mức chi tiêu trung bình cho mỗi giao dịch hoặc thói quen mua sắm theo thời gian.

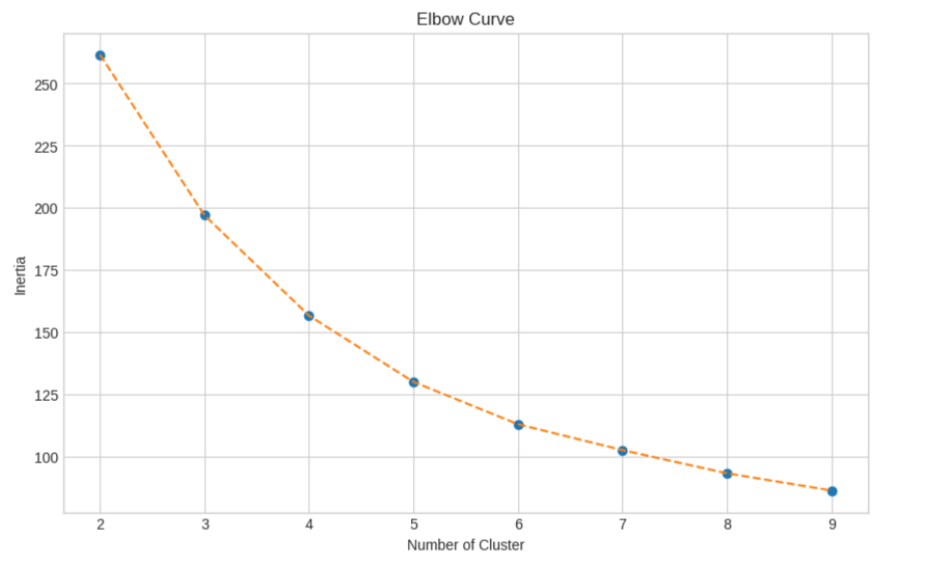
## Kết quả Mô hình

### Mô hình K-means

#### Xác định số lượng cụm tối ưu

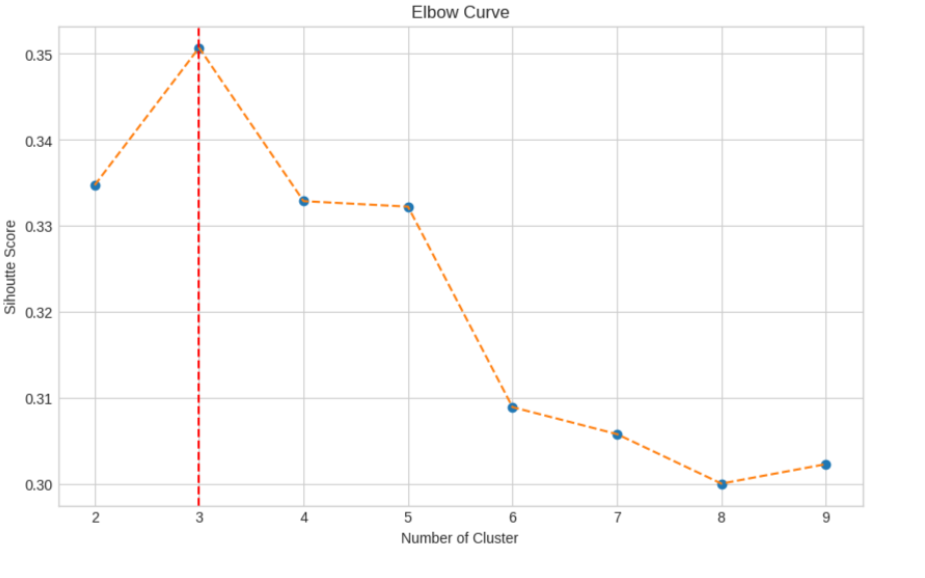
Để xác định số cụm (k) tối ưu cho việc phân khúc khách hàng, tôi sẽ khám phá hai phương pháp: phương pháp Elbow và phương pháp Silhouette để chứng thực kết quả:

\*Phương pháp Elbow



Hình 3.4 Sử dụng phương pháp Elbow

Bằng cách sử dụng phương pháp Elbow (Khuỷu tay), tôi nhận thấy rằng tại 3 điểm k=3, k=4 và k=5 đồ thị có sự thay đổi điểm inertia (tổng bình phương khoảng cách mỗi điểm dữ liệu đến tâm) thay đổi chậm lại.



Hình 3.5 Biểu đồ sử dụng phương pháp Elbow

Sau khi xem xét cẩn thận các đồ thị hình bóng, việc chọn ( k = 3 ) là lựa chọn tốt hơn, nơi có điểm silhouette cao nhất. Lựa chọn này mang lại cho tôi các cụm được kết hợp đồng đều hơn và được xác định rõ ràng hơn, giúp giải pháp phân cụm của tôi mạnh mẽ và đáng tin cậy hơn.

#### Đánh giá phân cụm

Sau khi xác định số cụm tối ưu k=3 tôi chuyển sang bước đánh giá để đánh giá chất lượng của các cụm được hình thành. Bước này nhằm xác nhận tính hiệu quả của việc phân cụm và đảm bảo rằng các cụm được mạch lạc và tách biệt tốt. Các số liệu đánh giá và kỹ thuật trực quan hóa mà tôi sử dụng:

1. Trực quan hóa 3D của các PC hàng đầu

2. Trực quan hóa phân phối cụm

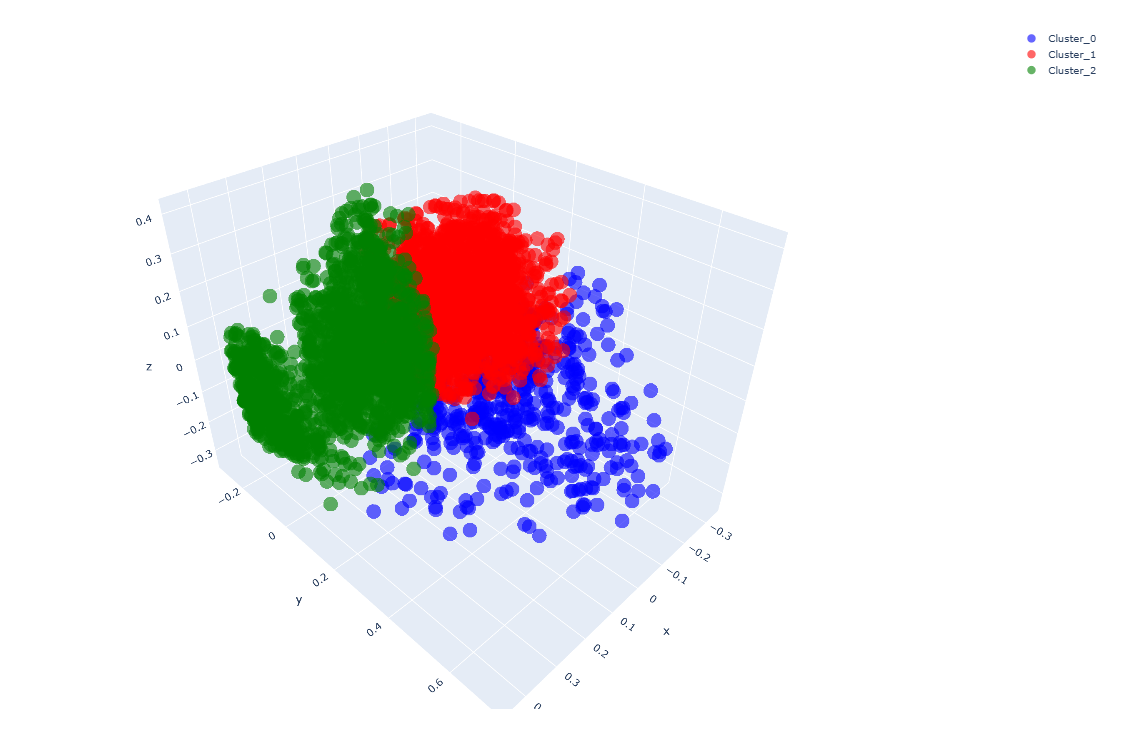
3. Các số liệu đánh giá:

Điểm Silhouette

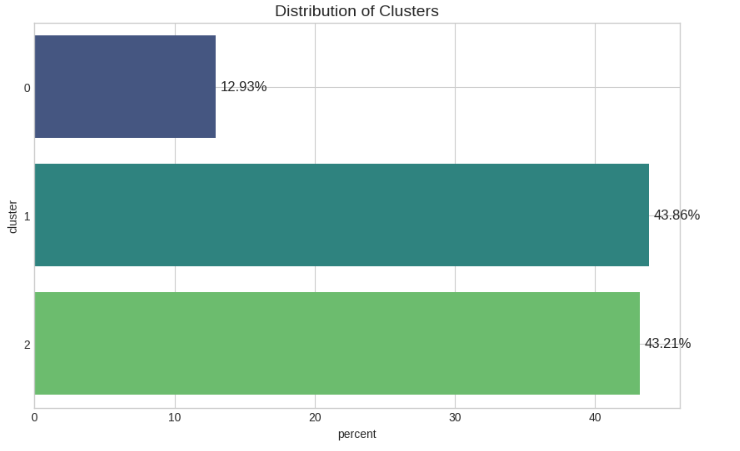
Điểm Calinski Harabasz

Điểm Davies Bouldin

Đầu tiên tôi sẽ chọn 3 PC hàng đầu (thu thập được nhiều sự khác biệt nhất trong dữ liệu) và sử dụng chúng để tạo hình ảnh 3D. Điều này sẽ cho phép tôi kiểm tra trực quan chất lượng phân tách và gắn kết của các cụm ở một mức độ nào đó:



Hình 3.6 Đánh giá phân cụm

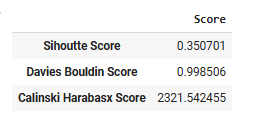
Tiếp theo tôi sẽ sử dụng biểu đồ thanh để trực quan hóa tỷ lệ phần trăm khách hàng trong mỗi cụm, điều này giúp hiểu được liệu các cụm có cân bằng và có ý nghĩa hay không:

***Hình 3.7 Biểu đồ thanh mô ta sự phân bổ khách hàng***

Sự phân bổ khách hàng trên các cụm, như được mô tả bằng biểu đồ thanh, cho thấy sự phân bổ khá cân bằng với các cụm 0 nắm giữ khoảng 12.93% số khách hàng mỗi cụm, cụm 1 chứa khoảng 43.86% số khách hàng, cụm 2 chứa khoảng 43.21% số khách hàng. Sự phân bổ cân bằng này cho thấy rằng quá trình phân cụm của tôi phần lớn đã thành công trong việc xác định các mẫu có ý nghĩa trong dữ liệu, thay vì chỉ phân nhóm nhiễu hoặc các ngoại lệ. Nó ngụ ý rằng mỗi cụm đại diện cho một phân khúc cơ sở khách hàng đáng kể và riêng biệt, từ đó cung cấp những hiểu biết sâu sắc có giá trị cho các chiến lược kinh doanh trong tương lai.

Cuối cùng để xem xét kỹ lưỡng hơn nữa chất lượng phân cụm tôi sẽ sử dụng các số liệu sau:

* Điểm Silhouette: Một thước đo để đánh giá khoảng cách tách biệt giữa các cụm. Giá trị cao hơn cho thấy khả năng phân tách cụm tốt hơn. Nó dao động từ -1 đến 1.
* Điểm Davies Bouldin: Nó đánh giá mức độ tương tự trung bình giữa mỗi cụm và cụm tương tự nhất của nó. Giá trị thấp hơn cho thấy khả năng phân tách cụm tốt hơn.
* Điểm Calinski Harabasz: Điểm này được sử dụng để đánh giá sự phân tán giữa và trong các cụm. Điểm cao hơn cho thấy các cụm được xác định tốt hơn.

Bảng 3.1 Các chỉ số đánh giá

Suy luận:

Điểm Silhouette xấp xỉ 0.35, mặc dù không gần bằng 1, nhưng vẫn cho thấy mức độ tách biệt khá lớn giữa các cụm. Nó gợi ý rằng các cụm có phần khác biệt nhưng có thể có sự chồng chéo nhẹ giữa chúng.

Điểm Davies Bouldin là 0.9 là điểm hợp lý, cho thấy mức độ tương đồng vừa phải giữa mỗi cụm và cụm tương tự nhất. Điểm thấp hơn thường tốt hơn vì nó cho thấy ít sự tương đồng hơn giữa các cụm và do đó, điểm số của tôi ở đây cho thấy sự tách biệt hợp lý giữa các cụm.

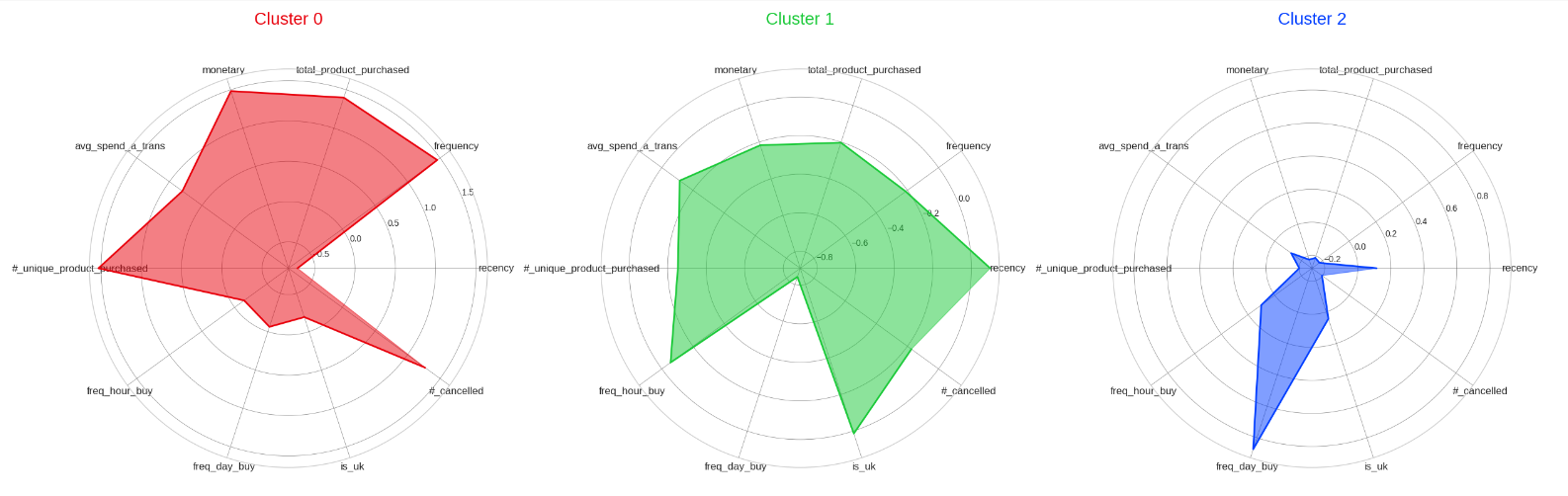
Điểm Calinski Harabasz là 2322, cao đáng kể, cho thấy các cụm được xác định rõ ràng. Điểm cao hơn trong số liệu này thường báo hiệu các định nghĩa cụm tốt hơn, do đó ngụ ý rằng việc phân cụm của tôi đã tìm được cấu trúc quan trọng trong dữ liệu.

=> Tóm lại, các số liệu cho thấy việc phân cụm có chất lượng tốt, với các cụm được xác định rõ ràng và khá tách biệt. Tuy nhiên, vẫn có thể cần tối ưu hóa hơn nữa để tăng cường phân tách và định nghĩa cụm, có thể bằng cách thử các thuật toán phân cụm và giảm kích thước khác.

#### Phân tích và lập hồ sơ cụm

Trong phần này, tôi sẽ phân tích đặc điểm của từng cụm để hiểu hành vi và sở thích riêng biệt của các phân khúc khách hàng khác nhau, đồng thời lập hồ sơ cho từng cụm để xác định các đặc điểm khách hàng chính trong mỗi cụm.

Tôi sẽ tạo biểu đồ radar để trực quan hóa các giá trị trọng tâm của từng cụm trên các đối tượng địa lý khác nhau. Điều này có thể đưa ra sự so sánh trực quan nhanh chóng về cấu hình của các cụm khác nhau.



Hình 3.8 Phân tích hồ sơ khách hàng bằng biểu đồ Radar

Phân tích hồ sơ khách hàng bắt nguồn từ biểu đồ Radar:

Cluster 0 (Red chart)

Khách hàng trong nhóm này có xu hướng chi tiêu nhiều, với số lượng giao dịch và sản phẩm mua cũng cao hơn. Số khách hàng không phải là khách hàng gần đây mua. khách hàng không có xu hướng mua hàng theo giờ, theo ngày cụ thể. Tỉ lệ hủy ở nhóm khách hàng này khá cao nhưng không phân bố đồng đều.

Cluster 1 (Green chart)

Khách hàng trong nhóm này có xu hướng chi tiêu ít hơn, với số lượng giao dịch và sản phẩm mua ít hơn. Có một số khách hàng là khách hàng gần đây mua xu hướng chi tiêu của họ tương đối ổn định nhưng ở mức thấp hơn. Khách hàng không có xu hướng mua hàng theo giờ, theo ngày cụ thể. Những khách hàng này thực hiện nhiều lần hủy, cho thấy tần suất và tỷ lệ hủy cao.

Cluster 2 (Blue chart)

Khách hàng trong nhóm này có xu hướng chi tiêu ít, với số lượng giao dịch và sản phẩm mua rất ít. Số khách hàng có xu hướng mua hàng theo ngày cụ thể. Những khách hàng này chưa thực hiện nhiều lần hủy và tỷ lệ hủy thấp. Có thể thấy là đây không phải nhóm khách hàng tiềm năng vì họ chi tiêu ít và đa số mua hàng vào các ngày cố định.

### Mô hình DBSCAN

#### Lựa chọn tham số cho mô hình DBSCAN

Hình 3.9 Biểu đồ K-distance

Biểu đồ K-distance hiển thị khoảng cách từ mỗi điểm dữ liệu đến điểm lân cận thứ k trong dữ liệu, sau khi các khoảng cách này được sắp xếp theo thứ tự tăng dần. Đây là một cách phổ biến để xác định giá trị Epsilon (ε) cho DBSCAN.

Trong biểu đồ K-distance:

**Trục x**: Đại diện cho các điểm dữ liệu, được sắp xếp theo khoảng cách lân cận thứ k của chúng.

**Trục y**: Đại diện cho khoảng cách lân cận thứ k (epsilon).

**Epsilon (ε):** Từ biểu đồ, ε khoảng 0.06 là một giá trị hợp lý. Đây là khoảng cách tối đa mà một điểm được xem là lân cận của một điểm khác trong cùng cụm.

**MinPts:** Tôi chọn MinPts = 6 cho bài toán cho phép mô hình phát hiện nhiều cụm nhỏ hơn và cải thiện khả năng phân tích dữ liệu chi tiết. Tuy nhiên, cần cân nhắc kỹ về mức độ nhiễu và tối ưu hóa các tham số để đạt được hiệu suất tốt nhất cho mô hình phân cụm.

Với các giá trị này, tôi có thể triển khai mô hình DBSCAN trên tập dữ liệu của mình để xác định các cụm và phân loại các điểm nhiễu. Điều này sẽ giúp khám phá các cấu trúc tiềm ẩn trong dữ liệu mà không cần giả định về hình dạng cụm trước.

#### Xây dựng mô hình DBSCAN

Hình 3.10 Xây dựng mô hình DBSCAN

**Số Lượng Cụm:**

Hình ảnh hiển thị một biểu đồ phân cụm 3D với 7 cụm khác nhau, được biểu diễn bằng các màu khác nhau. Đây là biểu đồ phân tích thành phần chính giúp giảm chiều dữ liệu xuống để trực quan hóa một cách dễ hiểu hơn.

**Phân Bố Cụm:**

Các cụm được phân bố khá rõ ràng trong không gian 3D.

**Cluster\_0 (Xanh dương)**: Cụm này tập trung ở phần trên của biểu đồ, có vẻ hơi tách rời so với các cụm khác. Điều này có thể cho thấy rằng các điểm dữ liệu trong cụm này có sự khác biệt rõ rệt so với các cụm còn lại.

**Cluster\_1 (Cam)**: Phân bố rộng khắp và có xu hướng xen lẫn vào các cụm khác. Điều này có thể ám chỉ rằng cụm này không có sự tách biệt rõ ràng và có nhiều điểm dữ liệu giao thoa với các cụm khác.

**Cluster\_2 (Xanh lá cây)**: Tập trung ở trung tâm của biểu đồ và có mật độ khá cao. Sự tập trung này cho thấy rằng cụm này có nhiều điểm dữ liệu tương đồng nằm gần nhau.

**Cluster\_3 (Tím)**: Phân bố gần trung tâm nhưng có mật độ ít hơn so với Cluster\_2. Cụm này có thể có một số điểm tương đồng với các cụm khác nhưng vẫn có sự phân biệt riêng.

**Cluster\_4 (Nâu)**: Cụm này có sự tập trung cao ở một khu vực nhất định và có thể chứa một số điểm dữ liệu ngoại lai hoặc nhiễu.

**Cluster\_5 (Đen)**: Được phân bố rộng và có vẻ như nằm rìa của biểu đồ. Cụm này có thể chứa nhiều điểm nhiễu hoặc dữ liệu không thuộc về các cụm chính.

**Cluster\_6 (Hồng)**: Nằm ở rìa của biểu đồ và có mật độ điểm dữ liệu thấp. Điều này có thể cho thấy rằng cụm này chứa các điểm dữ liệu ngoại lai hoặc có sự khác biệt lớn với các cụm khác.

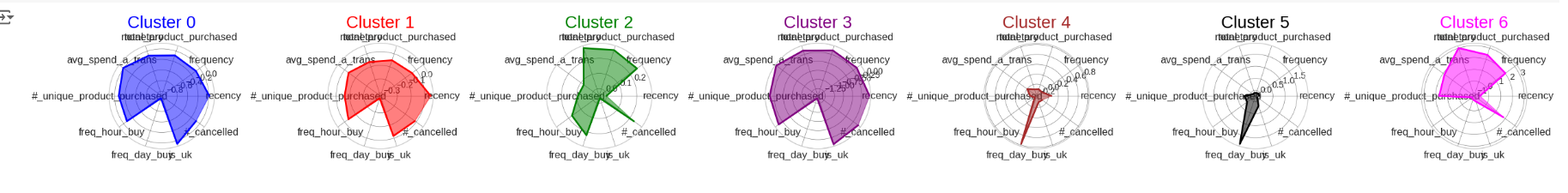
**Cụm Nhiễu (Noise):**

**Cluster\_4 (Nâu):** Cụm này có mật độ điểm dữ liệu cao trong một khu vực nhỏ, điều này có thể chỉ ra sự hiện diện của nhiễu hoặc các điểm dữ liệu không đặc trưng.

**Cluster\_5 (Đen):** Với sự phân bố rộng rãi và nằm rìa biểu đồ, cụm này có khả năng chứa các điểm dữ liệu nhiễu hoặc không có sự liên kết chặt chẽ với các cụm chính khác.

**Cluster\_6 (Hồng):** Với mật độ điểm dữ liệu thấp và vị trí nằm rìa biểu đồ, cụm này có thể chứa các điểm dữ liệu ngoại lai hoặc nhiễu.

#### Kết quả mô hình DBSCAN

****

Hình 3.11 Kết quả mô hình DBSCAN

**Phân Bố Khách Hàng:**

**Cluster 0, Cluster 1,** **Cluster 2** và **Cluster 6** là bốn nhóm chính với tần suất giao dịch và chi phí giao dịch cao. Trong đó, khách hàng mua rất đa dạng các loại sản phẩm và trung bình giá một đơn hàng khá cao. Chứng tỏ đây là bốn nhóm khách hàng có nguồn kinh tế ổn định.

**Cluster 4** và **Cluster 5** đại diện cho các khách hàng mua sắm cố định theo ngày. Hai nhóm khách hàng này tần suất giao dịch rất thấp, kéo theo tổng chi phí cho mua sắm ít.

**Phân Khúc Khách Hàng:**

**Cluster 0**: Nhóm khách hàng thường xuyên mua hàng. Có thể là khách hàng trung thành với nhiều giao dịch có giá trị cao.

**Cluster 1**: Nhóm khách hàng thường xuyên mua hàng. Có thể là khách hàng trung thành với nhiều giao dịch có giá trị tầm trung.

**Cluster 2**: Nhóm khách hàng thường xuyên mua hàng. Có thể là khách hàng trung thành với nhiều giao dịch có giá trị nhỏ.

**Cluster 3**: Nhóm khách hàng thường xuyên mua hàng. Có thể là khách hàng trung thành với nhiều giao dịch có giá trị cao.

**Cluster 4**: Nhóm khách hàng ít giao dịch. Có thể là nhóm khách hàng săn lùng ưu đãi.

**Cluster 5**: Nhóm khách hàng ít giao dịch. Có thể là nhóm khách hàng săn lùng ưu đãi

**Cluster 6**: Nhóm khách hàng thường xuyên mua hàng. Có thể là nhóm khách hàng tiềm năng.

Phân cụm bằng DBSCAN đã giúp xác định rõ các nhóm khách hàng với các đặc điểm khác nhau. Mỗi cụm cần các chiến lược quản lý và tiếp thị riêng biệt để tối ưu hóa giá trị và tần suất giao dịch, tăng cường sự hài lòng và duy trì khách hàng.

### So sánh và Đánh giá

So sánh kết quả của mô hình

Dựa trên các chỉ số: Silhouette Score, Davies Bouldin Score, Calinski Harabasz

A screenshot of a computer

Description automatically generated

Bảng 3.2 Bảng so sánh giữa hai mô hình

**Silhouette Score:** Silhouette Score đo lường mức độ tương đồng của mỗi điểm dữ liệu với cụm của nó so với các cụm khác. Giá trị dao động từ -1 đến 1, giá trị càng cao thì cụm càng tốt.

K-Means: 0.35

DBSCAN: 0.14

**Nhận xét**: K-Means có Silhouette Score cao hơn đáng kể so với DBSCAN, cho thấy rằng các cụm được tạo bởi K-Means có sự phân biệt rõ ràng hơn và điểm dữ liệu trong cùng một cụm có xu hướng tương đồng hơn.

**Davies-Bouldin Score:** Davies-Bouldin Score đo lường sự gần gũi giữa các cụm, giá trị càng thấp thì cụm càng tốt.

**K-Means**: 0.998506

**DBSCAN**: 2.262253

**Nhận xét**: K-Means có Davies-Bouldin Score thấp hơn, chỉ ra rằng các cụm của K-Means cách xa nhau hơn và khác biệt rõ ràng hơn so với DBSCAN.

**Calinski-Harabasz Score:** Calinski-Harabasz Score đo lường tỷ lệ giữa tổng độ phân tán giữa các cụm và tổng độ phân tán trong cụm. Giá trị càng cao thì cụm càng tốt.

**K-Means**: 2321.6

**DBSCAN**: 905.256

**Nhận xét**: K-Means có Calinski-Harabasz Score cao hơn nhiều so với DBSCAN, cho thấy rằng các cụm của K-Means có độ phân tán tốt hơn và các điểm dữ liệu trong cùng cụm gần nhau hơn.

Dựa trên các biểu đồ radar trực quan của các cụm được tạo bởi K-Means và DBSCAN

**K-Means Clusters:**

Các cụm được phân tách rõ ràng và đồng đều.

Các đặc trưng như monetary, total\_product\_purchased, frequency và recency có giá trị đặc trưng khác nhau cho mỗi cụm.

Các cụm có hình dạng gần như đối xứng và các giá trị của các đặc trưng được phân bố đều.

**DBSCAN Clusters:**

Có nhiều cụm hơn so với K-means, điều này cho thấy DBSCAN có khả năng phát hiện các cụm nhỏ và phức tạp hơn.

Một số cụm có hình dạng không đối xứng và có sự biến đổi lớn giữa các đặc trưng.

DBSCAN cũng phát hiện các điểm nhiễu (nếu có) và không gán chúng vào bất kỳ cụm nào, điều này hữu ích cho việc loại bỏ các dữ liệu bất thường.

**Đánh giá mô hình**

Dựa trên mục tiêu nâng cao hiệu quả của các chiến lược tiếp thị, tăng hiệu suất bán hàng và thúc đẩy doanh số bán hàng thông qua phân khúc khách hàng, K-Means là lựa chọn phù hợp hơn vì:

Các chỉ số Silhouette Score, Davies Bouldin Score, Calinski Harabasz tốt hơn mô hơn mô hình DBSCAN

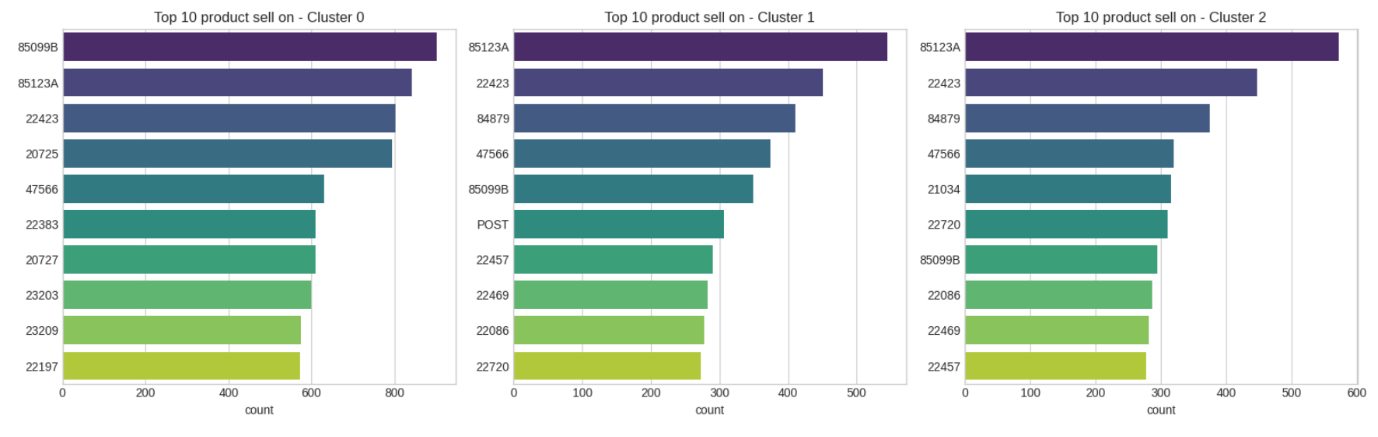
Các cụm được phân tách rõ ràng, đồng đều và dễ hiểu.

Dễ dàng triển khai các chiến lược tiếp thị cho từng nhóm khách hàng cụ thể dựa trên các cụm này.

Nó có vẻ là lựa chọn tốt hơn để phân khúc khách hàng, vì nó cung cấp các cụm rõ ràng và dễ dàng xác định các nhóm khách hàng có giá trị cao, từ đó nâng cao hiệu quả của các chiến lược tiếp thị và thúc đẩy doanh số bán hàng.

## Ứng dụng kết quả mô hình K-means

Để giúp doanh nghiệp tối ưu hóa các chiến lược tiếp thị và nâng cao chất lượng sản phẩm. Tôi có thể dựa trên kết quả phân cụm của mô hình k-means, từ đó có thể cung cấp những thông tin quý giá để nâng cao hiệu quả kinh doanh.

* **Top 10 sản phẩm bán chạy nhất theo từng cụm (Cluster) sau khi áp dụng mô hình phân cụm K Means.**

Hình 3.12 Top 10 sản phẩm bán chạy nhất

**Sản phẩm phổ biến:**

85123A là sản phẩm phổ biến nhất, xuất hiện trong top bán chạy nhất trong 3 cụm

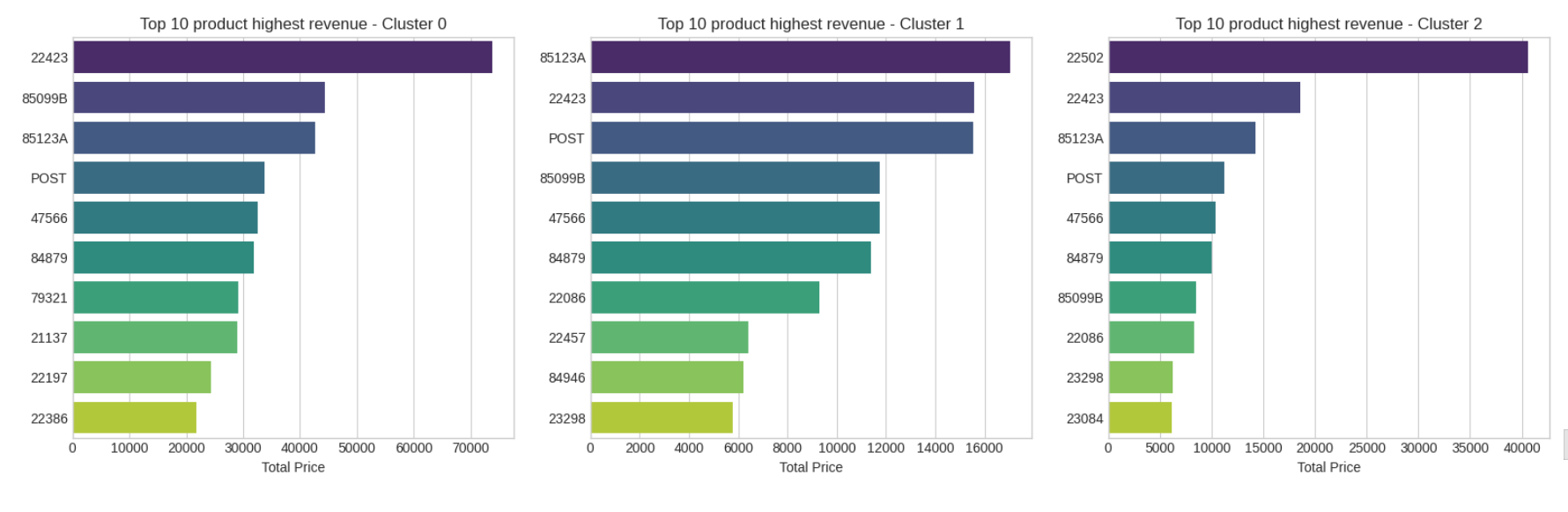
85099B cũng là sản phẩm phổ biến, đứng đầu trong cụm 0

**Phân bố số lượng bán:**

Cluster 0, Cluster 1 và Cluster 2 có sự phân bố số lượng bán khá đồng đều.

Biểu đồ này giúp nhận diện các sản phẩm chủ lực trong từng cụm khách hàng, từ đó có thể điều chỉnh chiến lược marketing và quản lý hàng tồn kho phù hợp để tối ưu hóa doanh thu và lợi nhuận.

* **Top 10 sản phẩm có doanh số cao nhất theo từng cụm(Cluster) sau khi áp dụng mô hình phân cụm K Means**



Hình 3.13 Top 10 sản phẩm có doanh số cao nhất

**Sản phẩm phổ biến:**

22423 là sản phẩm có doanh số cao nhất trong 2 cụm (0, 2).

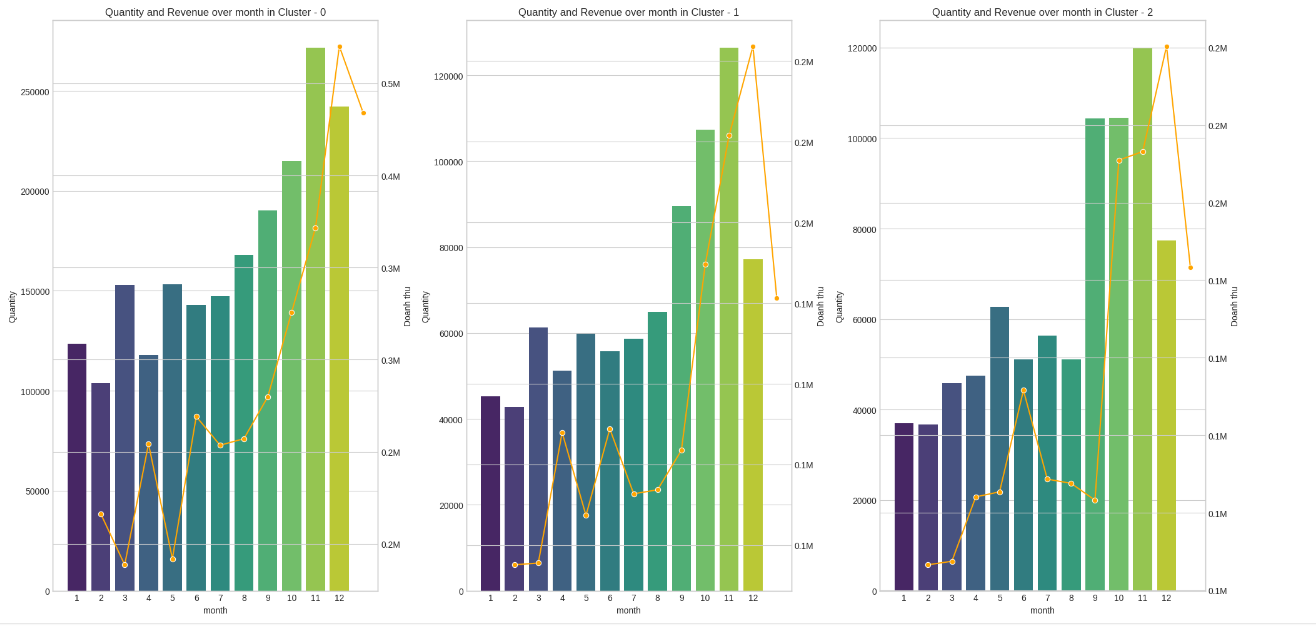
85123A và POST cũng xuất hiện nhiều lần trong top 10 của các cụm khác nhau.

**Phân bố doanh thu:**

Cluster 0 và Cluster 1 có sản phẩm dẫn đầu với doanh thu gần 40,000 cho thấy sự chênh lệch rõ rệt về doanh thu so với các sản phẩm khác.

Cluster 2 có tổng doanh thu thấp hơn so với các cụm còn lại, cho thấy có thể đây là nhóm khách hàng ít mua sắm hơn hoặc sản phẩm ít phổ biến.

Biểu đồ này giúp nhận diện các sản phẩm chủ lực đem lại doanh thu cao trong từng cụm khách hàng, từ đó có thể điều chỉnh chiến lược marketing tận dụng bán các sản phẩm doanh thu cao tặng kèm các sản phẩm doanh thu thấp. Từ đó dễ dàng quản lý hàng tồn kho phù hợp để tối ưu hóa doanh thu.

* **Số lượng sản phẩm và doanh thu qua các tháng từng cụm(Cluster) sau khi áp dụng mô hình phân cụm K Means**.

Hình 3.14 Số lượng sản phẩm và doanh thu qua các tháng

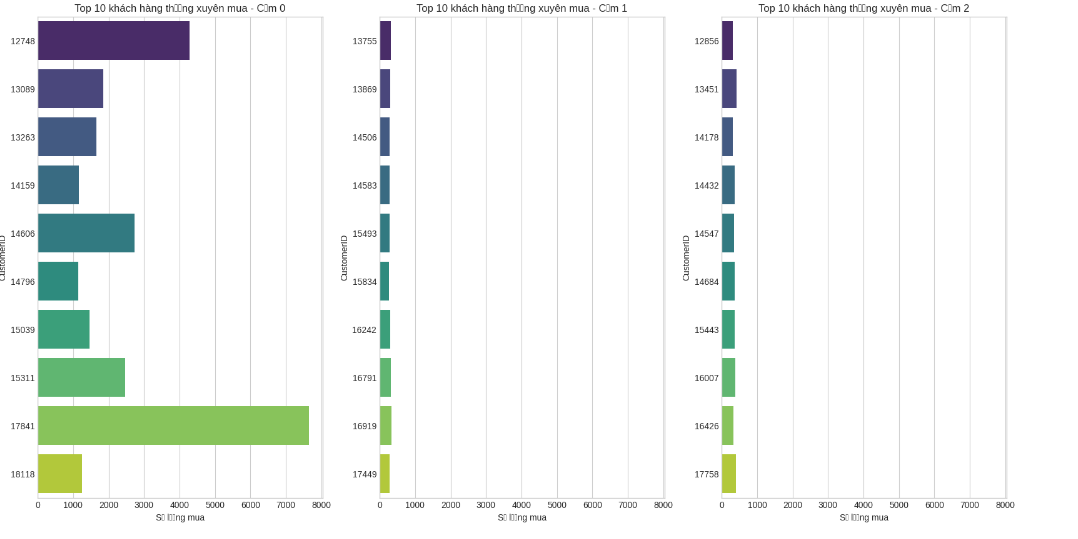
Nhận xét chung:

Tất cả các cụm đều có xu hướng tăng số lượng sản phẩm bán ra và doanh thu từ giữa năm trở đi, đặc biệt là từ tháng 9 đến tháng 11 hoặc 12.

Cụm 2 có sự biến động lớn hơn trong số lượng sản phẩm bán ra và doanh thu so với cụm 0 và 1.

Doanh thu thường đạt đỉnh vào tháng 11 hoặc 12 trong tất cả các cụm, cho thấy sự tăng trưởng mạnh vào cuối năm, có thể do các sự kiện mua sắm lớn như Black Friday, Cyber Monday, và mùa mua sắm Giáng sinh.

Biểu đồ này giúp chúng ta nhận ra các xu hướng theo mùa trong doanh thu và số lượng sản phẩm bán ra của từng cụm khách hàng, từ đó có thể điều chỉnh các sản phầm phù hợp theo mùa. Giúp khách hàng dễ dàng tiếp cận và mua sản phẩm.

* **Top 10 Khách hàng mua thường xuyên trong mỗi cụm(Cluster) sau khi áp dụng mô hình phân cụm K Means.**

Hình 3.15 Top 10 khách hàng phổ biến

**Khách hàng phổ biến:**

Khách hàng trong các cụm khác nhau có hành vi mua sắm khá đồng nhất, với top khách hàng mua thường xuyên nhất có số lượng mua tương đương nhau (dao động từ 2000 đến 8000 đơn).

CustomerID 17841 trong Cluster 3 và CustomerID 14071 trong Cluster 0 là những khách hàng có số lượng mua cao nhất.

**Phân bố tần suất mua:**

Mỗi cụm có ít nhất một khách hàng có số lượng mua gần 8000, cho thấy có một số khách hàng rất trung thành và mua hàng thường xuyên.

Các khách hàng còn lại trong top 10 của mỗi cụm có số lượng mua tương đối đồng đều, dao động từ 2000 đến 6000.

**Khách hàng đặc biệt:**

Các khách hàng trong Cluster 0 có số lượng mua cao hơn so với các cụm khác, cho thấy có thể đây là nhóm khách hàng có giá trị cao và cần được chăm sóc đặc biệt.

Việc phân tích tần suất mua hàng của khách hàng trong từng cụm giúp tôi hiểu rõ hơn về hành vi mua sắm của khách hàng và từ đó đưa ra các chiến lược kinh doanh

**Kết luận:** Chương 3 của nghiên cứu đã sử dụng hai mô hình K-Means và DBSCAN để phân tích hành vi mua sắm của khách hàng, đưa ra các kết luận quan trọng. Giúp xác định các yếu tố ảnh hưởng lớn đến quyết định mua hàng của khách hàng, như giá cả, đánh giá sản phẩm và các chương trình khuyến mãi, qua đó điều chỉnh chiến lược kinh doanh phù hợp để tăng cường sự gắn kết với khách hàng.

# KẾT LUẬN

Tôi tập trung vào việc ứng dụng mô hình K-Means để dự đoán hành vi mua sắm trực tuyến của người tiêu dùng. Qua quá trình nghiên cứu và phân tích dữ liệu, mô hình K-Means đã chứng minh là một công cụ mạnh mẽ trong việc phân loại và xác định các nhóm khách hàng có hành vi mua sắm tương đồng.

Kết quả từ mô hình này không chỉ giúp tôi hiểu rõ hơn về nhu cầu và sở thích của từng nhóm khách hàng mà còn hỗ trợ doanh nghiệp trong việc phát triển các chiến lược tiếp thị và dịch vụ cá nhân hóa, nâng cao trải nghiệm người dùng. Hơn nữa, việc sử dụng K-Means để phân tích hành vi mua sắm trực tuyến đã mở ra những cơ hội mới cho việc cải thiện hiệu quả hoạt động và tối ưu hóa quy trình kinh doanh của doanh nghiệp.

# TÀI LIỆU THAM KHẢO

Trần Hùng Cường, Trần Thanh Hùng. (2017). Giáo trình khai phá dữ liệu. Trường Đại học Công nghiệp Hà Nội.

PhamDinhKhanh. Phương pháp phân cụm dựa trên mật độ (Density-Based Clustering)

Lê Kim Dung. “Nghiên cứu hành vi mua sắm trực tuyến của người tiêu dùng Việt Nam”.2020. Luận án tiến sĩ, Học viện khoa học xã hội.

Jasiński, M., Sikorski, T., Kaczorowska, D., Rezmer, J., Suresh, V., Leonowicz, Z., Kostyła, P., Szymańda, J., Janik, P., Bieńkowski, J., & Prus, P. (2021). A case study on data mining application in a virtual power plant: Cluster analysis of power quality measurements. *Energies*, *14*(4), 974.

Kanagala, Hari Krishna, and VV Jaya Rama Krishnaiah. "A comparative study of K-Means, DBSCAN and OPTICS." 2016 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2016.

Fawzia Omer, Albasheer, et al. "Big data mining using K-Means and DBSCAN clustering techniques." Big Data Analytics and Computational Intelligence for Cybersecurity. Cham: Springer International Publishing, 2022. 231-246.

Ahmed, Mohammed A., Hanif Baharin, and Puteri NE Nohuddin. "Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses." International Journal of Advanced Computer Science and Applications 11.8 (2020): 248-254.

Limwattanapibool, Onapa, and Somjit Arch‐int. "Determination of the appropriate parameters for K‐means clustering using selection of region clusters based on density DBSCAN (SRCD‐DBSCAN)." Expert Systems 34.3 (2017): e12204.