HUTECH
Đại học Công nghệ Tp.HCM

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

ĐỒ ÁN CHUYÊN NGÀNH

ỨNG DỤNG KHAI THÁC TẬP PHỔ BIẾN TỐI ĐẠI CẢI THIỆN KẾT QUẢ PHÂN CỤM DỮ LIỆU CÁC CỬA HÀNG TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ TIKI

Ngành: **KHOA HỌC DỮ LIỆU**
Chuyên ngành: **KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: **TS. BÙI DANH HƯỜNG**
Sinh viên thực hiện: Nguyễn Văn Đạt
MSSV: 2186400229 Lớp: 21DKHA1

TP. Hồ Chí Minh, 2024

HUTECH
Đại học Công nghệ Tp.HCM

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

ĐỒ ÁN CHUYÊN NGÀNH

ỨNG DỤNG KHAI THÁC TẬP PHỔ BIẾN TỐI ĐẠI CẢI THIỆN KẾT QUẢ PHÂN CỤM DỮ LIỆU CÁC CỬA HÀNG TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ TIKI

Ngành: **KHOA HỌC DỮ LIỆU**
Chuyên ngành: **KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: **TS. BÙI DANH HƯỜNG**
Sinh viên thực hiện: Nguyễn Văn Đạt
MSSV: 2186400229 Lớp: 21DKHA1

TP. Hồ Chí Minh, 2024

LỜI CAM ĐOAN

Tôi, Nguyễn Văn Đạt xin cam đoan rằng:

Mọi thông tin và nghiên cứu được trình bày trong bài báo cáo này là trung thực và khách quan được thu thập và phân tích một cách cẩn thận dựa trên các nguồn chính thống và đáng tin cậy.

Bất kỳ thông tin hoặc ý kiến nào được trích dẫn từ các nguồn khác đều được nêu rõ nguồn gốc và được trích dẫn theo đúng quy định. Tôi cam đoan rằng không có bất kỳ sự sao chép hoặc sử dụng thông tin không đúng đắn nào từ các nguồn khác.

Bài báo cáo này là công trình nghiên cứu độc lập của tôi chưa từng được công bố ở bất kỳ nơi nào khác. Tôi cam đoan rằng đã tuân thủ đầy đủ các quy tắc và quy định của môn học bao gồm cả việc tham khảo và sử dụng công cụ nghiên cứu.

Tôi hy vọng rằng bài báo cáo này sẽ cung cấp một cái nhìn tổng quan rõ ràng và toàn diện về chủ đề “Ứng dụng khai thác tập phổ biến tối đại cải thiện kết quả phân cụm dữ liệu các cửa hàng trên sàn thương mại điện tử Tiki" và sẽ đóng góp một phần nhỏ vào lĩnh vực nghiên cứu này.

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

TPHCM, Ngày..... tháng..... năm 2024

Giáo viên hướng dẫn

(Ký tên, đóng dấu)

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

Chương 1. TỔNG QUAN

1.1. Giới thiệu về đề tài

Trong thời đại thương mại điện tử phát triển mạnh mẽ như hiện nay, việc tối ưu hóa hiệu quả kinh doanh của các cửa hàng trực tuyến trở thành yếu tố quan trọng không thể thiếu đối với mọi doanh nghiệp. Mỗi cửa hàng trên sàn thương mại điện tử đều có những đặc điểm riêng, và việc hiểu rõ những yếu tố này sẽ giúp các cửa hàng định hướng phát triển phù hợp, từ đó nâng cao hiệu quả hoạt động và khả năng cạnh tranh.

Tuy nhiên, để có thể đưa ra các chiến lược phát triển chính xác, các doanh nghiệp cần phải hiểu rõ về hành vi, đặc điểm và hiệu suất của các cửa hàng khác nhau trên nền tảng thương mại điện tử. Đây là lý do việc phân tích và phân cụm dữ liệu của các cửa hàng trực tuyến đóng vai trò quan trọng trong việc xác định các nhóm cửa hàng có những đặc điểm tương đồng. Các nhóm này sẽ giúp chỉ ra những yếu tố thành công chung hoặc những điểm yếu cần khắc phục.

Đề tài "Ứng Dụng Khai Thác Tập Phổ Biến Tối Đại Cải Thiện Kết Quả Phân Cụm Dữ Liệu Các Cửa Hàng Trên Sàn Thương Mại Điện Tử Tiki" được lựa chọn nhằm phát triển một phương pháp khoa học để cải thiện kết quả phân cụm các cửa hàng trực tuyến. Cụ thể, phương pháp kết hợp khai thác tập phổ biến tối đại và phân cụm sẽ giúp làm giàu dữ liệu và nâng cao chất lượng phân cụm cửa hàng. Điều này sẽ giúp doanh nghiệp dễ dàng nhận diện các yếu tố quan trọng ảnh hưởng đến hiệu suất kinh doanh, từ đó đưa ra chiến lược phát triển chính xác hơn và cải thiện khả năng cạnh tranh trên thị trường.

1.2. Nhiệm vụ của đề tài

Đề tài "Ứng Dụng Khai Thác Tập Phổ Biến Tối Đại Cải Thiện Kết Quả Phân Cụm Dữ Liệu Các Cửa Hàng Trên Sàn Thương Mại Điện Tử Tiki" tập trung áp dụng các kỹ thuật xử lý dữ liệu và thuật toán Machine Learning để phân nhóm các cửa hàng trực tuyến có đặc điểm tương đồng. Kết hợp khai thác tập phổ biến tối đại (maximal frequent itemset mining) giúp phát hiện mối quan hệ tiềm ẩn giữa các yếu tố, từ đó nâng cao hiệu quả phân cụm. Kết quả phân tích sẽ hỗ trợ doanh nghiệp hiểu rõ thị trường, tối ưu hóa chiến lược marketing, cải thiện dịch vụ khách hàng, nâng cao trải nghiệm người dùng và hiệu quả kinh doanh.

1.2.1. Tính cấp thiết của đề tài

Thương mại điện tử phát triển mạnh mẽ và cạnh tranh ngày càng khốc liệt, đặc biệt trên các sàn như Tiki, đòi hỏi doanh nghiệp phải liên tục tối ưu hóa hoạt động để phát triển. Phân cụm cửa hàng giúp nhận diện và phân tích các nhóm có đặc điểm và hiệu suất tương đồng, hỗ trợ xây dựng chiến lược kinh doanh hiệu quả. Việc kết hợp khai thác tập phổ biến tối đại không chỉ làm giàu dữ liệu mà còn cải thiện chất lượng phân cụm, giúp doanh nghiệp đưa ra quyết định chính xác hơn.

Cụ thể, phân cụm các cửa hàng giúp doanh nghiệp:

- Tối ưu hóa chiến lược marketing và bán hàng: Thiết kế chiến lược phù hợp với từng nhóm cửa hàng, tối ưu hóa chi phí và hiệu quả.
- Cải thiện dịch vụ khách hàng và trải nghiệm người dùng: Hiểu rõ nhu cầu của khách hàng, nâng cao dịch vụ và trải nghiệm người dùng.
- Tối ưu hóa hoạt động kinh doanh: Tập trung vào các nhóm cửa hàng có hiệu suất cao, từ đó tối ưu hóa quản lý kho, phân phối sản phẩm.
- Nâng cao khả năng cạnh tranh: Hiểu rõ và đáp ứng nhanh các thay đổi thị trường, giữ vững vị thế cạnh tranh.
- Ra quyết định chiến lược dựa trên dữ liệu: Sử dụng dữ liệu để đưa ra quyết định chính xác và hiệu quả hơn.

1.2.2. Ý nghĩa khoa học và thực tiễn của đề tài

Về mặt khoa học, đề tài đóng góp vào lĩnh vực phân tích dữ liệu và học máy thông qua việc áp dụng các thuật toán phân cụm tiên tiến như K-Means, Agglomerative, Gaussian Mixture Model kết hợp khai thác tập phổ biến tối đại, giúp nâng cao hiệu quả phân nhóm và cung cấp cái nhìn sâu sắc về hành vi người tiêu dùng, hiệu suất kinh doanh.

Về thực tiễn, đề tài hỗ trợ doanh nghiệp ra quyết định chính xác, xây dựng chiến lược quản lý và marketing phù hợp theo từng nhóm cửa hàng. Kết quả nghiên cứu có thể áp dụng trên các nền tảng thương mại điện tử khác, nâng cao chất lượng ngành và hỗ trợ các nhà quản lý thiết kế chính sách phát triển bền vững.

1.3. Mục tiêu, đối tượng và phạm vi nghiên cứu

1.3.1. Mục tiêu

Đề tài hướng đến việc xây dựng phương pháp phân tích và phân cụm các cửa hàng trực tuyến trên Tiki, nhận diện các nhóm cửa hàng có đặc điểm và hiệu suất

kinh doanh tương đồng. Bằng cách kết hợp khai thác tập phổ biến tối đại (maximal frequent itemset mining), đề tài phát hiện mối quan hệ tiềm ẩn trong dữ liệu, từ đó nâng cao chất lượng phân cụm.

Các bước thực hiện bao gồm thu thập, tiền xử lý dữ liệu, khám phá đặc điểm cơ bản và áp dụng các mô hình phân cụm (K-Means, Agglomerative, Gaussian Mixture Model). Quá trình phân cụm được đánh giá qua các chỉ số Silhouette Score, Davies-Bouldin Index, và Calinski–Harabasz Index. Kết quả phân tích sẽ đề xuất chiến lược phát triển và tối ưu hóa kinh doanh, giúp nâng cao năng lực cạnh tranh.

1.3.2. Đối tượng và phạm vi

Đối tượng nghiên cứu là các cửa hàng trực tuyến trên Tiki. Phạm vi tập trung vào việc thu thập, xử lý và phân tích dữ liệu cửa hàng, áp dụng phân cụm trước và sau khi kết hợp khai thác tập phổ biến tối đại để đánh giá hiệu quả, từ đó đề xuất giải pháp kinh doanh tối ưu.

1.4. Phương pháp nghiên cứu

1.4.1. Phương pháp nghiên cứu sơ bộ

Trước khi tiến hành thu thập dữ liệu, một nghiên cứu sơ bộ sẽ được thực hiện nhằm nắm bắt tổng quan về lĩnh vực thương mại điện tử, các yếu tố ảnh hưởng đến hiệu suất kinh doanh của cửa hàng trực tuyến, và các phương pháp phân tích dữ liệu phổ biến. Nghiên cứu sơ bộ giúp xác định rõ các vấn đề cần giải quyết và đề xuất phương pháp nghiên cứu phù hợp.

1.4.2. Phương pháp nghiên cứu tài liệu

Đề tài sẽ thu thập và phân tích tài liệu liên quan đến các phương pháp phân cụm và khai thác tập phổ biến tối đại trong lĩnh vực học máy và thương mại điện tử. Qua việc đánh giá các nghiên cứu trước đây và công trình khoa học, phương pháp nghiên cứu hiệu quả nhất sẽ được chọn để áp dụng vào dữ liệu thực tế, đảm bảo tính chính xác và phù hợp.

1.4.3. Phương pháp nghiên cứu thống kê

Trong phân tích dữ liệu, các kỹ thuật thống kê như phân tích đơn biến, đa biến, phân tích phương sai, và kiểm tra độ tương quan sẽ được sử dụng để mô tả và đánh giá các biến số. Những phân tích này sẽ giúp nhận diện các yếu tố ảnh hưởng đến hiệu suất kinh doanh của các cửa hàng trực tuyến, cung cấp nền tảng cho việc áp dụng các phương pháp phân cụm.

1.4.4. Phương pháp thực nghiệm

Thực nghiệm được tiến hành trên dữ liệu thu thập từ sàn thương mại điện tử Tiki. Quá trình này bao gồm tiền xử lý dữ liệu, áp dụng thuật toán K-Means, Agglomerative, Gaussian Mixture Model để phân cụm các cửa hàng, và kết hợp khai thác tập phổ biến tối đại nhằm nâng cao hiệu quả phân cụm. Kết quả sẽ được đánh giá qua các chỉ số như Silhouette Score, Davies-Bouldin Index, Calinski–Harabasz Index và phân tích hiệu suất kinh doanh của từng cụm để đề xuất chiến lược kinh doanh cụ thể.

1.4.5. Phương pháp đánh giá

Phương pháp đánh giá sẽ được sử dụng để kiểm tra tính hiệu quả của các mô hình phân cụm và chiến lược kinh doanh đề xuất. Việc đo lường và so sánh các chỉ số hiệu suất kinh doanh giữa các cụm sẽ đảm bảo rằng các phương pháp được áp dụng mang lại giá trị thực tiễn cao.

1.5. Những đóng góp nghiên cứu của đề tài

1.5.1. Trong lĩnh vực học thuật

Đề tài đóng góp vào việc ứng dụng và phát triển các phương pháp phân tích dữ liệu tiên tiến, đặc biệt là kết hợp các thuật toán phân cụm K-Means, Agglomerative, Gaussian Mixture Model và khai thác tập phổ biến tối đại trong ngữ cảnh thương mại điện tử. Đây là một hướng tiếp cận mới, tạo ra các mô hình phân cụm cụ thể để phân tích dữ liệu các cửa hàng trực tuyến. Bên cạnh đó, đề tài xây dựng một bộ dữ liệu đa dạng từ sàn thương mại điện tử Tiki, cung cấp nguồn tài nguyên giá trị cho các nghiên cứu tiếp theo trong lĩnh vực thương mại điện tử và học máy.

1.5.2. Trong thực tiễn kinh doanh

Đề tài mang lại giá trị thực tiễn bằng cách cung cấp các chiến lược kinh doanh phù hợp dựa trên việc nhận diện các nhóm cửa hàng có đặc điểm và hiệu suất kinh doanh tương đồng. Điều này giúp doanh nghiệp tối ưu hóa chiến lược marketing, cải thiện dịch vụ khách hàng, và nâng cao hiệu quả kinh doanh. Các chiến lược cụ thể cho từng nhóm cửa hàng không chỉ giúp doanh nghiệp gia tăng doanh thu mà còn nâng cao khả năng cạnh tranh trên thị trường thương mại điện tử đang phát triển mạnh mẽ.

Chương 2. CƠ SỞ LÝ THUYẾT

2.1. API Scraping

2.1.1. Giới thiệu về trích xuất dữ liệu từ API (API Scraping)

API scraping [1] là một kỹ thuật để trích xuất dữ liệu từ các trang web bằng cách sử dụng API, cung cấp quyền truy cập dữ liệu có cấu trúc và có tổ chức. Nó rất hữu ích để trích xuất dữ liệu từ các nền tảng truyền thông xã hội và các trang web thương mại điện tử.

Quá trình này bao gồm ba bước chính:

- Xác định API endpoint: Đây là URL mà yêu cầu sẽ được gửi tới để truy xuất dữ liệu.
- Gửi yêu cầu: Tạo yêu cầu HTTP đến API endpoint, thường sử dụng các phương thức như GET, POST, PUT, DELETE.
- Xử lý phản hồi: Nhận phản hồi từ API, thường được trả về dưới dạng dữ liệu cấu trúc như JSON hoặc XML và sau đó được xử lý bằng các ngôn ngữ lập trình như Python, JavaScript, hoặc Ruby.

2.1.2. Ưu điểm và hạn chế

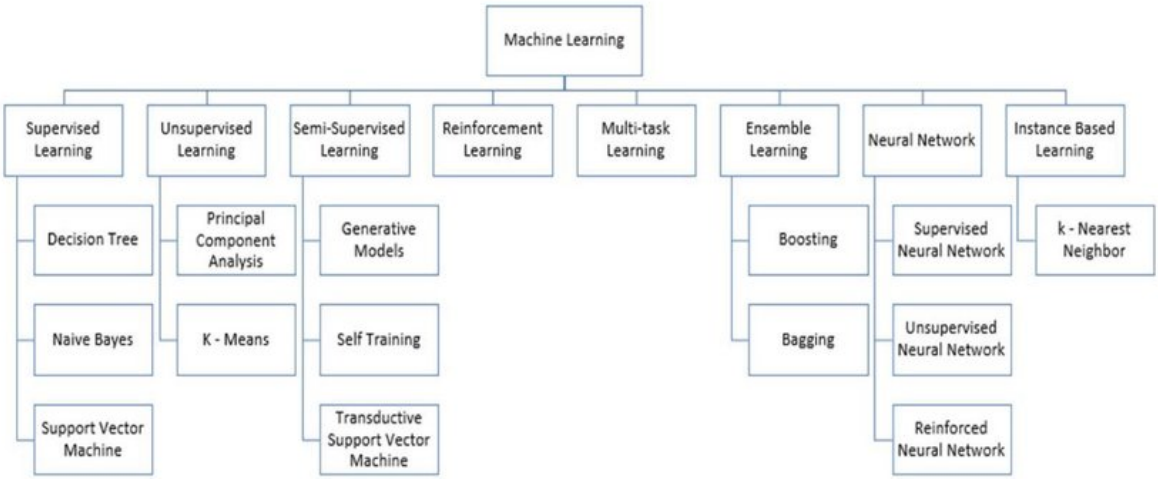
Ưu điểm: Dữ liệu có cấu trúc (JSON, XML) dễ xử lý. Độ chính xác cao, cập nhật từ nguồn chính. Truy vấn nhanh hơn so với phân tích HTML.

Hạn chế: API có thể giới hạn số truy vấn và chỉ cung cấp một phần dữ liệu, gây khó khăn khi thu thập thông tin. Một số API yêu cầu xác thực phức tạp và tính phí, đặc biệt với khối lượng dữ liệu lớn. Ngoài ra, việc phụ thuộc vào nhà cung cấp API tiềm ẩn rủi ro khi dịch vụ thay đổi hoặc ngừng mà không báo trước.

2.2. Machine Learning

Machine Learning [2] là một nhánh của trí tuệ nhân tạo tập trung vào việc phát triển và nghiên cứu các kỹ thuật giúp hệ thống có thể tự học từ dữ liệu để giải quyết các vấn đề cụ thể. Theo Arthur Samuel Machine learning được định nghĩa là lĩnh vực nghiên cứu mang lại cho máy tính khả năng học hỏi mà không cần lập trình rõ ràng. Trong đề tài này, tôi tập trung vào sử dụng các thuật toán Unsupervised Learning trong **Hình 2.1** để giải quyết bài toán phân cụm.

2.2.1. Unsupervised Learning



Hình 2.1: Tổng quan về Machine Learning

Unsupervised Learning [3] là một nhánh của Machine Learning nhằm tìm ra một mô hình mà phù hợp với các quan sát. Các thuật toán Unsupervised Learning học một số tính năng từ dữ liệu. Khi dữ liệu mới được đưa vào, nó sẽ sử dụng các tính năng đã học trước đó để nhận dạng lớp dữ liệu. **Hình 2.2** dưới đây mô tả tổng quan về học không giám sát.

2.2.2. Clustering

Clustering [4] là một nhánh của Unsupervised Learning được sử dụng để nhóm các điểm dữ liệu có đặc tính tương tự vào các nhóm cụm (clusters) khác nhau. Mục tiêu của phân cụm là tìm ra cấu trúc ẩn trong dữ liệu mà không cần sự giám sát của nhãn. **Hình 2.3** thể hiện tổng quan về phân cụm.

2.3. Silhouette Score

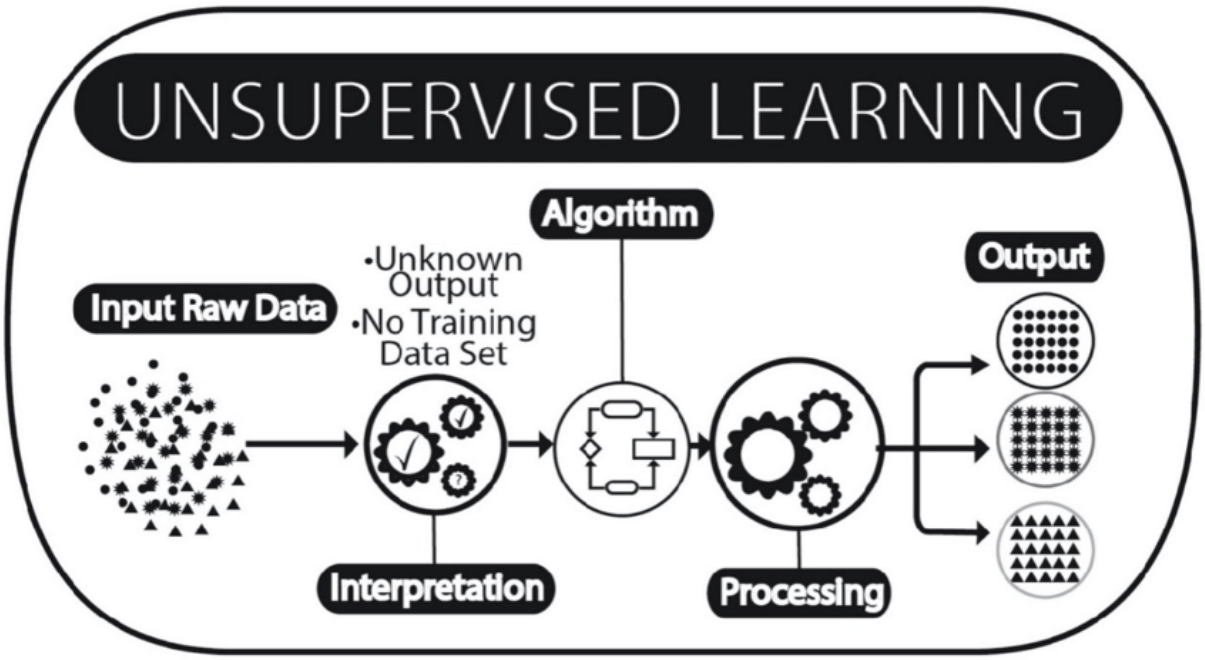
2.3.1. Giới thiệu về Silhouette Score

Chỉ số Silhouette [5] là một phương pháp đánh giá hiệu suất phân cụm. Nó cung cấp một phương pháp đánh giá đối với từng điểm dữ liệu trong một cụm bằng cách tính toán độ tương tự của điểm đó với các điểm trong cụm và độ khác biệt với các điểm trong các cụm khác.

2.3.2. Nền tảng toán học

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{1}$$

Trong đó:



Hình 2.2: Tổng quan về Unsupervised Learning

- $a(i)$ Khoảng cách trung bình từ điểm i đến các điểm khác trong cùng cụm.
- $b(i)$ Khoảng cách trung bình từ điểm i đến các điểm trong cụm gần nhất.

Silhouette score cho mỗi điểm dữ liệu nằm trong khoảng [-1, 1]. Giá trị gần 1 cho thấy điểm dữ liệu đó nằm trong cụm thích hợp, giá trị gần -1 cho thấy điểm dữ liệu đó có thể được phân loại sai, giá trị gần 0 cho thấy điểm dữ liệu đó nằm gần biên của hai cụm.

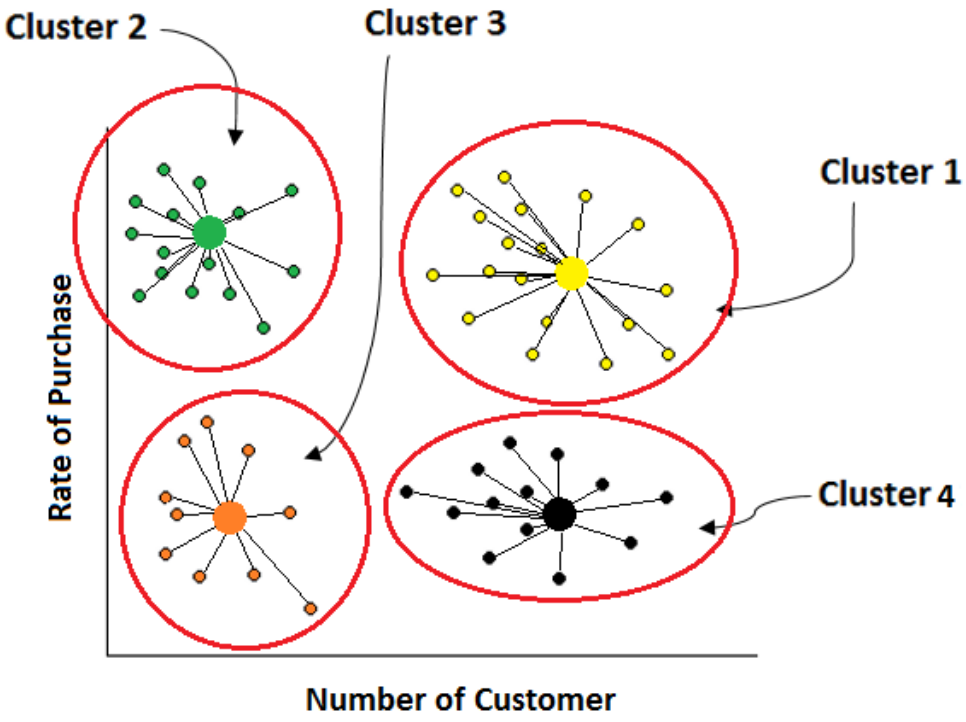
2.3.3. Diễn giải thuật toán

Silhouette Score đo chất lượng phân cụm qua mức độ gắn kết trong cụm và tách biệt giữa các cụm. Mỗi điểm được tính khoảng cách trung bình đến các điểm trong cùng cụm (gắn kết) và cụm gần nhất (tách biệt). Giá trị từ -1 đến 1, với gần 1 là cụm tốt, 0 là ranh giới, và âm là phân cụm sai. Chỉ số trung bình giúp đánh giá và chọn số cụm tối ưu.

2.3.4. Ưu điểm và hạn chế

Ưu điểm: cung cấp một phương pháp đánh giá đối với chất lượng của phân cụm mà không cần biết trước số lượng cụm.

Hạn chế: Cần tính toán khoảng cách giữa mỗi cặp điểm dữ liệu, làm tăng độ phức tạp tính toán. Không hiệu quả đối với dữ liệu có cấu trúc phức tạp hoặc cụm có kích thước không đồng đều.



Hình 2.3: Tổng quan về Clustering

2.4. Davies-Bouldin Index

2.4.1. Giới thiệu về Davies-Bouldin Index

Davies-Bouldin Index (DBI) [6] là chỉ số đánh giá chất lượng phân cụm dựa trên mức độ gắn kết trong cụm và tách biệt giữa các cụm. Mức độ gắn kết đo sự gần nhau của các điểm trong một cụm, trong khi mức độ tách biệt đo khoảng cách giữa các cụm. Davies-Bouldin Index phản ánh sự cân bằng giữa hai yếu tố này, với giá trị nhỏ hơn cho thấy phân cụm chất lượng tốt hơn.

2.4.2. Nền tảng toán học

Cho một tập dữ liệu được phân thành k cụm, công thức tính Davies-Bouldin Index được định nghĩa như sau:

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j} R_{ij} \tag{2}$$

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

$$S_i = \frac{1}{n_i}\sum_{x\in C_i}\|x - c_i\|$$

$$M_{ij} = \|c_i - c_j\|$$

Trong đó:

- R_{ij} để đánh giá sự tương đồng giữa hai cụm i và j.
- Giá trị $\max_{i \neq j} R_{ij}$ thể hiện cụm i nào gần cụm j nhất. DBI trung bình qua tất cả các cụm cho phép đánh giá tổng thể chất lượng phân cụm.
- S_i là độ phân tán của cụm i thường được tính bằng trung bình khoảng cách Euclidean giữa các điểm trong cụm i và tâm cụm c_i.
- n_i là số điểm trong cụm i và C_i là tập hợp các điểm thuộc cụm i.
- M_{ij} là khoảng cách giữa tâm cụm c_i và c_j.

2.4.3. Diễn giải thuật toán

Davies-Bouldin Index đánh giá chất lượng phân cụm bằng cách đo mức độ nhỏ gọn của từng cụm và sự tách biệt giữa các cụm. Tâm cụm được xác định, sau đó so sánh độ nhỏ gọn với khoảng cách đến các cụm khác để chọn giá trị đại diện cho mỗi cụm. Trung bình các giá trị này tạo thành chỉ số DBI, với giá trị nhỏ cho thấy các cụm nhỏ gọn và phân tách rõ ràng, trong khi giá trị lớn phản ánh sự chồng lấn. DBI hỗ trợ đánh giá, cải thiện phân cụm và chọn số cụm tối ưu.

2.4.4. Ưu điểm và hạn chế

Ưu điểm: Dễ tính toán nhờ sử dụng các khái niệm toán học đơn giản như khoảng cách và trung bình. Phổ biến, được áp dụng với nhiều thuật toán phân cụm khác nhau. Tự động đánh giá chất lượng phân cụm, giá trị DBI nhỏ thể hiện cụm tách biệt và nhỏ gọn tốt.

Hạn chế: Phụ thuộc vào số lượng cụm được chọn, dễ sai lệch nếu số cụm không phù hợp. Nhạy cảm với giá trị ngoại lai, khiến outlier ảnh hưởng đến kết quả đánh giá. Không phù hợp với dữ liệu phi tuyến tính hoặc các phân cụm có hình dạng phức tạp.

2.5. Calinski–Harabasz Index

2.5.1. Giới thiệu về Calinski–Harabasz index

Calinski–Harabasz Index (CH Index) [7], còn được gọi là Variance Ratio Criterion, là một chỉ số được sử dụng phổ biến để đánh giá chất lượng của các cụm trong bài toán phân cụm không giám sát. CH Index đo lường tỷ lệ giữa tổng phương sai giữa các cụm và tổng phương sai trong cụm, nhằm đánh giá mức độ tách biệt giữa

các cụm và sự nhỏ gọn trong từng cụm. Giá trị CH càng cao, phân cụm càng được đánh giá tốt.

2.5.2. Nền tảng toán học

Công thức toán học của Calinski–Harabasz Index được định nghĩa như sau:

$$CH = \frac{\text{trace}(B_k)}{\text{trace}(W_k)} \cdot \frac{n-k}{k-1} \tag{3}$$

$$\text{trace}(B_k) = \sum_{i=1}^{k} n_i \cdot \|c_i - c\|^2$$

$$\text{trace}(W_k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$$

Trong đó:

- $\frac{\text{trace}(B_k)}{\text{trace}(W_k)}$ là tỷ lệ giữa phương sai giữa các cụm và phương sai trong cụm.
- $\frac{n-k}{k-1}$ là hệ số điều chỉnh phụ thuộc vào số điểm dữ liệu n và số cụm k.
- $\text{trace}(B_k)$ là tổng phương sai giữa các cụm.
- $\text{trace}(W_k)$ là tổng phương sai trong cụm.
- n_i là số điểm dữ liệu trong cụm i.
- c_i là tâm của cụm i.
- c là tâm của toàn bộ tập dữ liệu.
- x là một điểm dữ liệu trong cụm i.
- C_i là tập hợp các điểm dữ liệu trong cụm i.
- n là tổng số điểm dữ liệu trong tập dữ liệu.
- k là số cụm.

x là một điểm dữ liệu trong cụm i, C_i là tập hợp các điểm dữ liệu trong cụm i, n là tổng số điểm dữ liệu trong tập dữ liệu, k là số cụm.

2.5.3. Diễn giải thuật toán

Calinski–Harabasz Index được tính toán để đánh giá chất lượng phân cụm dựa trên sự nhỏ gọn trong cụm và mức độ tách biệt giữa các cụm. Đầu tiên, tập dữ liệu được chia thành các cụm, mỗi cụm có một tâm đại diện. Thuật toán đo lường sự nhỏ gọn bằng khoảng cách giữa các điểm trong cụm và tâm của nó, đồng thời đánh giá sự tách biệt dựa trên khoảng cách giữa tâm cụm và tâm toàn bộ tập

dữ liệu. Cuối cùng, chỉ số Calinski–Harabasz được tính bằng tỷ lệ giữa sự tách biệt và sự nhỏ gọn, có điều chỉnh theo số lượng cụm và số điểm dữ liệu. Giá trị chỉ số càng cao chứng tỏ các cụm được tách biệt rõ ràng và nhỏ gọn.

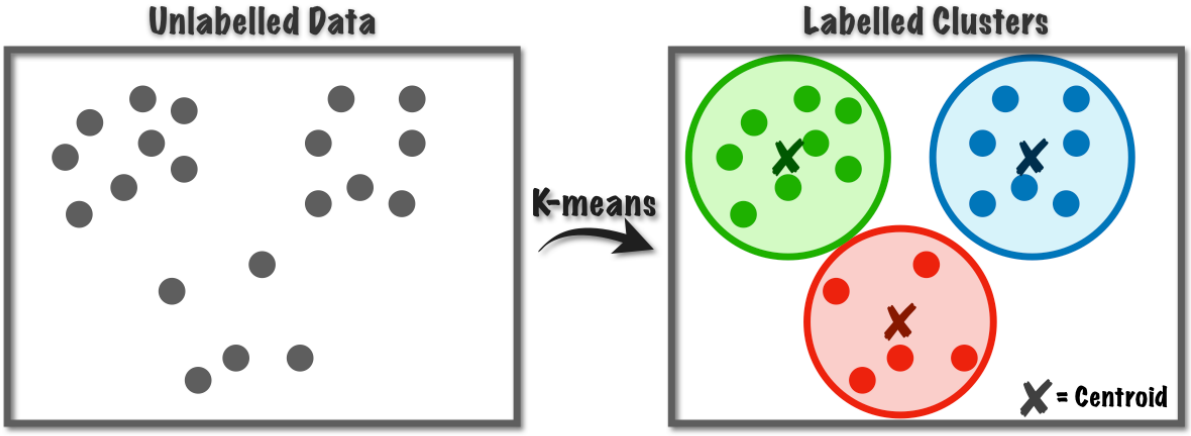
2.5.4. Ưu điểm và hạn chế

Ưu điểm: Đơn giản và dễ tính toán, áp dụng được trên nhiều thuật toán phân cụm. Cho phép so sánh hiệu quả phân cụm giữa các thuật toán hoặc số lượng cụm khác nhau. Giá trị phản ánh rõ sự tách biệt và sự nhỏ gọn của các cụm, hỗ trợ đánh giá chất lượng.

Hạn chế: Có xu hướng ưu tiên số cụm lớn, có thể dẫn đến kết quả không tối ưu. Không phù hợp với dữ liệu phi tuyến hoặc cụm có hình dạng phức tạp. Nhạy cảm với kích thước không đồng đều giữa các cụm, ảnh hưởng đến độ chính xác.

2.6. Phân cụm K-Means

2.6.1. Giới thiệu về thuật toán K-Means



Hình 2.4: Tổng quan về K-Means. Nguồn: [9]

K-Means [8] là một thuật toán học không giám sát đơn giản và phổ biến, được sử dụng rộng rãi để giải quyết các bài toán phân cụm. Thuật toán này đặc biệt hiệu quả trong việc xử lý các tập dữ liệu lớn, giúp nó trở thành một công cụ lý tưởng cho các nhiệm vụ khai thác dữ liệu. **Hình 2.4** thể hiện 1 cách rõ nét về K-Means.

2.6.2. Nền tảng toán học

1. Tạo các trung tâm ngẫu nhiên:

$$c^{(0)} = \left(m_1^{(0)}, m_2^{(0)}, \ldots, m_k^{(0)}\right) \tag{2.2}$$

2. Gán các điểm dữ liệu vào các cụm.
   Với mỗi điểm dữ liệu, ta sẽ tính khoảng cách của nó tới các trung tâm (bằng Khoảng cách Euclid). Ta sẽ gán chúng vào trung tâm gần nhất. Tập hợp các điểm được gán vào cùng 1 trung tâm sẽ tạo thành cụm.

$$S_j^{(t)} = \left\{x_p : \|x_p - m_j^{(t)}\|^2 \leq \|x_p - m_i^{(t)}\|^2, \forall j, i \leq k\right\} \tag{2.3}$$

   Trong đó:

   - $S_j^{(t)}$: Tập hợp các điểm dữ liệu được gán vào cụm j tại bước thứ t.
   - x_p: Một điểm dữ liệu trong tập dữ liệu.
   - $m_j^{(t)}$: Trung tâm của cụm j tại bước thứ t.
   - $\|x_p - m_i^{(t)}\|^2$: Là bình phương của khoảng cách Euclide giữa điểm dữ liệu x_p và trung tâm của cụm i tại vòng lặp thứ t.
   - $\|x_p - m_j^{(t)}\|^2$: Là bình phương của khoảng cách Euclide giữa điểm dữ liệu x_p và trung tâm của cụm j tại vòng lặp thứ t.

3. Cập nhật trung tâm: Với mỗi cụm đã tìm được ở bước 2, trung tâm mới sẽ là trung bình cộng của các điểm dữ liệu trong cụm đó.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{2.4}$$

   Trong đó:

   - $m_i^{(t+1)}$: Đây là trung tâm (centroid) mới của cụm i tại vòng lặp $(t+1)$. Sau khi cập nhật, $m_i^{(t+1)}$ trở thành trung tâm mới của cụm i.
   - $S_i^{(t)}$: Đây là tập hợp các điểm dữ liệu thuộc cụm i tại vòng lặp t. Nói cách khác, $S_i^{(t)}$ là kích thước của tập hợp $S_i^{(t)}$, chứa tất cả các điểm dữ liệu được gán vào cụm i ở vòng lặp t.
   - $\frac{1}{\left|S_i^{(t)}\right|}$: Đây là phần tử nhân lợi cho việc tính trung bình của các điểm dữ liệu trong tập hợp $S_i^{(t)}$. Nói cách khác, công thức này tính trung bình cộng của tất cả các điểm dữ liệu trong cụm i ở vòng lặp t.

2.6.3. Diễn giải thuật toán

K-Means Clustering dựa trên nguyên tắc tối ưu hóa tổng khoảng cách bình phương từ các điểm dữ liệu đến trung tâm cụm của chúng. Các bước chính bao gồm: Khởi tạo các centroid: Chọn ngẫu nhiên K điểm làm trung tâm ban đầu của các cụm. Phân công cụm: Gán mỗi điểm dữ liệu vào cụm có centroid gần nhất. Cập nhật centroid: Tính toán lại vị trí centroid bằng cách lấy trung bình tất cả các điểm dữ liệu trong cụm. Lặp lại: Tiếp tục quá trình phân công và cập nhật cho đến khi các centroid không thay đổi hoặc thay đổi rất ít giữa các lần lặp.

2.6.4. Ưu điểm và hạn chế

Ưu điểm: Dễ triển khai và hiệu quả đối với dữ liệu lớn, Cho phép phân cụm dựa trên khoảng cách Euclidean giữa các điểm dữ liệu.

Hạn chế: Yêu cầu biết trước số lượng cụm k. Nhạy cảm với trọng tâm ban đầu, có thể dẫn đến kết quả khác nhau. Không hiệu quả với các cụm có kích thước hoặc hình dạng không đồng nhất.

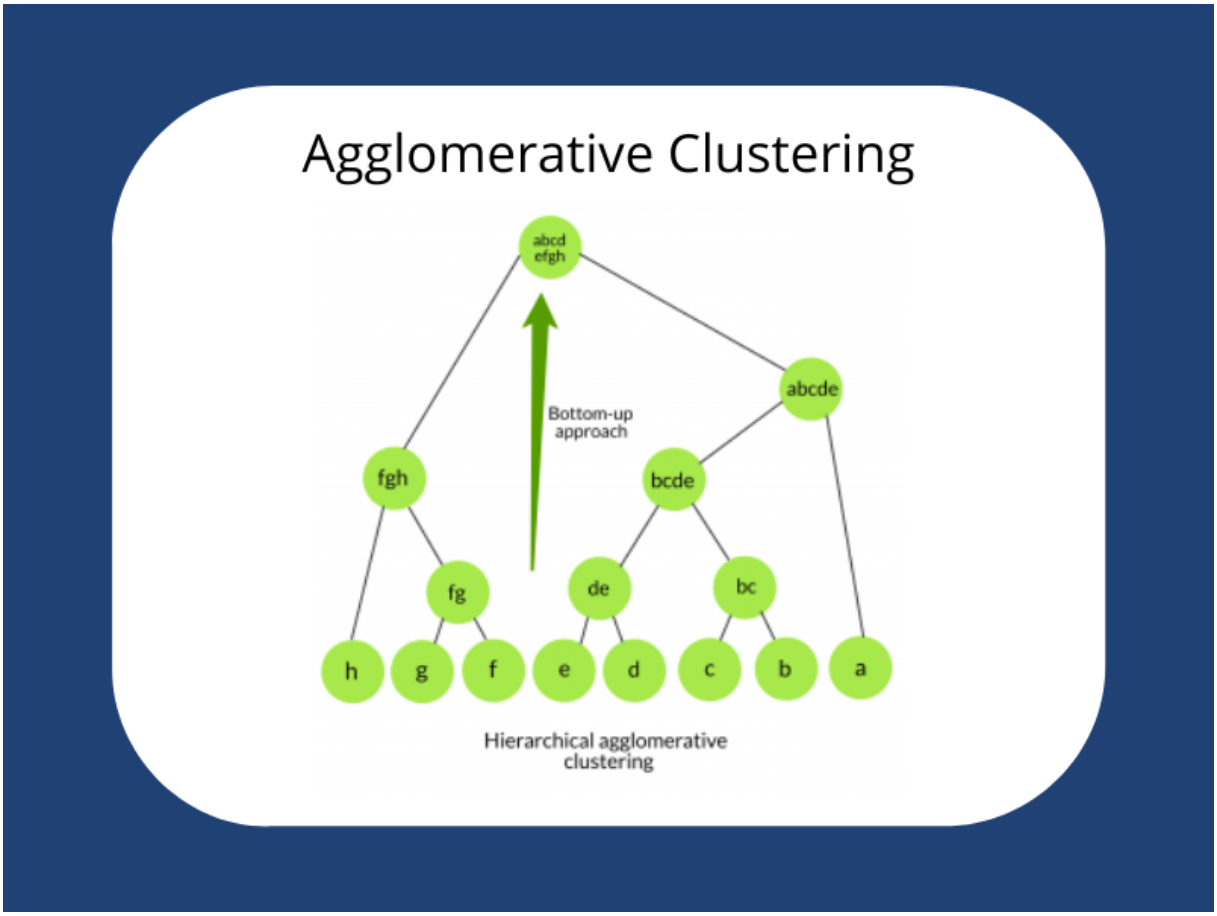
2.6.5. Ứng dụng

- Trong marketing, K-Means được ứng dụng để phân đoạn khách hàng dựa trên hành vi mua sắm và đặc điểm nhân khẩu học.
- Trong lĩnh vực xử lý hình ảnh, thuật toán này được sử dụng để giảm số lượng màu trong hình ảnh, giúp giảm kích thước tệp mà vẫn giữ được chất lượng hình ảnh ở mức chấp nhận được.
- Trong y tế, K-Means hỗ trợ phân loại các loại bệnh hoặc tình trạng sức khỏe dựa trên các chỉ số y tế.
- Trong xử lý ngôn ngữ tự nhiên, K-Means có thể phân loại các tài liệu thành các chủ đề khác nhau, giúp tổ chức và khai thác thông tin hiệu quả hơn.

2.7. Phân cụm Agglomerative

2.7.1. Giới thiệu về thuật toán Agglomerative

Phân cụm Agglomerative [10] là một thuật toán phân cụm phân cấp (hierarchical clustering) thuộc nhóm phương pháp phân cụm kết hợp từ dưới lên. Thuật toán bắt đầu với mỗi điểm dữ liệu là một cụm riêng lẻ, sau đó dần dần kết hợp các cụm gần nhau nhất dựa trên tiêu chí đo khoảng cách, cho đến khi tất cả các điểm được gộp lại thành một cụm duy nhất hoặc đạt số lượng cụm mong muốn. **Hình 2.5** minh họa rõ ràng cách thức hoạt động của Agglomerative.



Hình 2.5: Tổng quan về Agglomerative. Nguồn: [11]

2.7.2. Nền tảng toán học

Giả sử có n điểm dữ liệu $X = \{x_1, x_2, \ldots, x_n\}$, thuật toán Agglomerative thực hiện các bước gộp cụm dựa trên tiêu chí đo khoảng cách giữa các cụm:

1. **Khoảng cách giữa các điểm dữ liệu**:

$$d(x_i, x_j) = \|x_i - x_j\|$$

   hoặc sử dụng các khoảng cách khác như Manhattan, Cosine, hoặc Jaccard. Trong đó: x_i, x_j là hai điểm trong không gian n chiều.

2. **Khoảng cách giữa các cụm**:

   - **Liên kết đơn (Single Linkage)**: Khoảng cách nhỏ nhất giữa các điểm của hai cụm.

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

     Trong đó: $\min_{x \in C_i, y \in C_j} d(x, y)$ là khoảng cách nhỏ nhất giữa hai điểm x và y.

- **Liên kết hoàn chỉnh (Complete Linkage)**: Khoảng cách lớn nhất giữa các điểm của hai cụm.

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

Trong đó: $\max_{x \in C_i, y \in C_j} d(x, y)$ là khoảng cách lớn nhất giữa hai điểm x và y.

- **Liên kết trung bình (Average Linkage)**: Khoảng cách trung bình giữa tất cả các điểm của hai cụm.

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$$

Trong đó: $\sum_{x \in C_i, y \in C_j} d(x, y)$ là tổng khoảng cách giữa tất cả các cặp điểm x trong cụm C_i và y trong cụm C_j.

3. **Ma trận khoảng cách**: Lưu trữ khoảng cách giữa các cụm để tìm cặp cụm gần nhất trong mỗi bước.

2.7.3. Diễn giải thuật toán

Thuật toán Agglomerative bắt đầu bằng cách coi mỗi điểm dữ liệu là một cụm riêng biệt. Sau đó, tính toán ma trận khoảng cách giữa tất cả các cụm hiện tại và tìm hai cụm gần nhất dựa trên tiêu chí khoảng cách (single linkage, complete linkage, hoặc average linkage). Hai cụm gần nhất này được gộp thành một cụm mới, và ma trận khoảng cách được cập nhật. Quá trình này lặp lại liên tục cho đến khi tất cả các điểm được gộp thành một cụm duy nhất hoặc đạt đến số lượng cụm mong muốn.

2.7.4. Ưu điểm và hạn chế

Ưu điểm: Phân cụm Agglomerative không cần xác định trước số cụm, cho phép linh hoạt lựa chọn số cụm sau khi thực hiện. Thuật toán xây dựng cấu trúc phân cấp, giúp hiểu rõ mối quan hệ giữa các điểm và trực quan hóa qua dendrogram. Ngoài ra, Agglomerative linh hoạt với nhiều tiêu chí liên kết và khoảng cách khác nhau, phù hợp với từng bài toán.

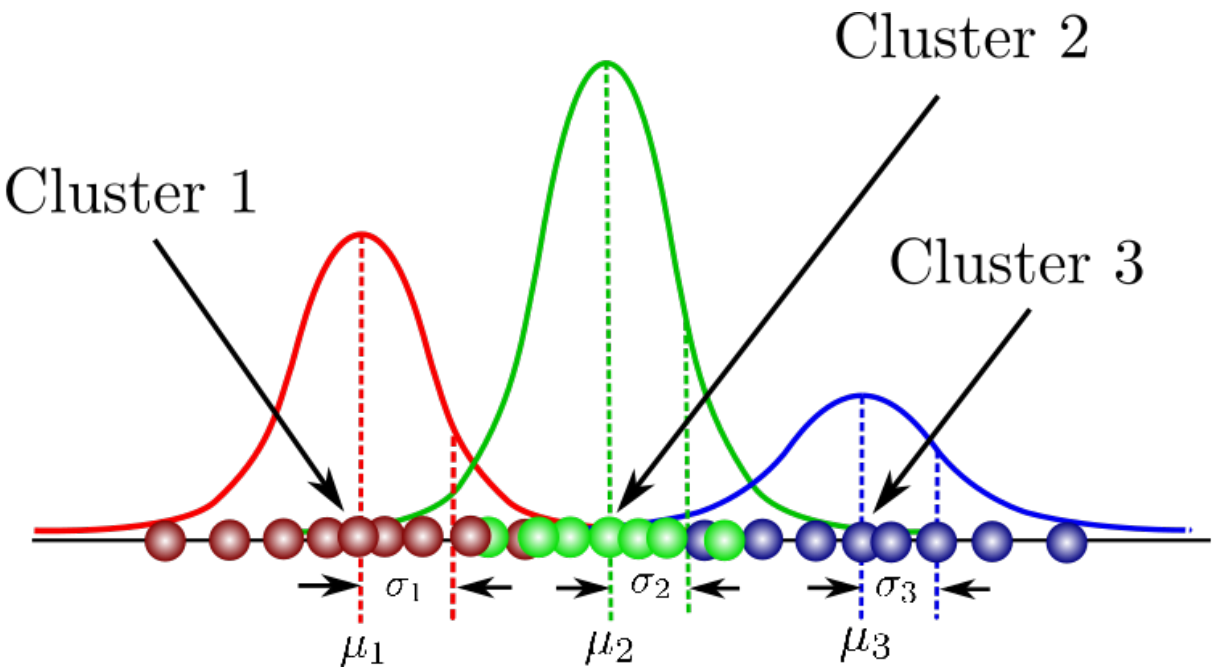
Hạn chế: Thuật toán có độ phức tạp cao $O(n^3)$, gây tốn kém tài nguyên khi xử lý dữ liệu lớn. Agglomerative cũng nhạy cảm với ngoại lai và nhiễu, dễ làm sai lệch kết quả. Với tập dữ liệu lớn, yêu cầu tính toán và lưu trữ trở nên không hiệu quả.

2.7.5. Ứng dụng

- Phân đoạn khách hàng trong marketing: Dựa trên hành vi mua sắm và nhân khẩu học để xác định các nhóm khách hàng tiềm năng.
- Xử lý văn bản: Nhóm tài liệu theo chủ đề để quản lý và tìm kiếm thông tin hiệu quả.
- Phân tích hình ảnh: Nhận diện mẫu và phát hiện các đặc điểm quan trọng trong hình ảnh.

2.8. Phân cụm Gaussian Mixture Model

2.8.1. Giới thiều về thuật toán Gaussian Mixture Model



Hình 2.6: Tổng quan về GMM. Nguồn: [13]

Gaussian Mixture Model (GMM) [12] là một phương pháp phân cụm mềm (soft clustering) trong học máy. Khác với phương pháp phân cụm cứng (hard clustering) như K-Means, GMM cho phép mỗi điểm dữ liệu có thể thuộc về nhiều cụm với các xác suất khác nhau. GMM hoạt động dựa trên giả định rằng dữ liệu được sinh ra từ sự kết hợp của nhiều phân phối Gaussian (normal distributions) như **Hình 2.6**.

2.8.2. Nền tảng toán học

GMM dựa trên mô hình hỗn hợp Gaussian (Gaussian Mixture Model), một mô hình xác suất bao gồm nhiều phân phối Gaussian. Công thức toán học của

GMM.

$$[p(x|\Theta) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k)] \tag{4}$$

- K: Số cụm (thành phần Gaussian).
- π_k: Trọng số của thành phần Gaussian thứ k, thỏa mãn $\sum_{k=1}^{K} \pi_k = 1$
- $\mathcal{N}(x|\mu_k, \Sigma_k)$: Hàm mật độ xác suất Gaussian với trung bình μ_k và ma trận hiệp phương sai Σ_k
- $\Theta = \{\pi_k, \mu_k, \Sigma_k\}$: Tập hợp tham số cần ước lượng.

Hàm mật độ của phân phối Gaussian được định nghĩa như sau:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

2.8.3. Diễn giải thuật toán

GMM sử dụng một quy trình lặp gọi là Expectation-Maximization (EM) để tối ưu hóa các tham số của mô hình. Trong bước đầu tiên (Expectation), GMM ước tính xác suất rằng mỗi điểm dữ liệu thuộc về một cụm nào đó dựa trên các tham số hiện tại. Trong bước thứ hai (Maximization), các tham số của từng cụm như trung bình, hình dạng và trọng số được cập nhật để tối ưu hóa sự khớp với dữ liệu. Quá trình này được lặp đi lặp lại cho đến khi mô hình hội tụ, tức là các tham số không thay đổi đáng kể nữa.

2.8.4. Ưu điểm và hạn chế

Ưu điểm: Gaussian Mixture Model (GMM) cho phép phân cụm mềm, phù hợp với dữ liệu không tách biệt rõ ràng. Thuật toán linh hoạt trong mô hình hóa các cụm có hình dạng khác nhau nhờ sử dụng ma trận hiệp phương sai và dựa trên nền tảng xác suất chặt chẽ, giúp giải thích kết quả rõ ràng.

Hạn chế: GMM yêu cầu xác định trước số cụm, nhạy cảm với giá trị khởi tạo và tiêu tốn nhiều tài nguyên, đặc biệt với dữ liệu lớn hoặc có số chiều cao.

2.8.5. Ứng dụng

- Xử lý tín hiệu: Nhận diện giọng nói, phân đoạn hình ảnh.
- Tài chính: Phân loại khách hàng, ước lượng rủi ro tín dụng.
- Hệ thống khuyến nghị: Nhóm sản phẩm, nhóm khách hàng.

2.9. Thư viện Underthesea

Underthesea [14] là một thư viện xử lý ngôn ngữ tự nhiên (NLP) dành cho tiếng Việt trên nền tảng Python, cung cấp các công cụ và chức năng hỗ trợ thực hiện nhiều tác vụ khác nhau. Các chức năng chính của thư viện bao gồm tách từ (tokenization), tách câu (sentence segmentation), phân loại từ loại (part-of-speech tagging), phân tích cú pháp (parsing), và phân loại cảm xúc (sentiment analysis). Underthesea giúp các nhà phát triển và nghiên cứu dễ dàng thực hiện các nhiệm vụ liên quan đến ngôn ngữ trong môi trường Python.

2.10. Chuẩn hóa dữ liệu Standard Scaler

2.10.1. Giới thiệu về StardardScaler

StandardScaler [15] là phương pháp chuẩn hóa dữ liệu phổ biến trong tiền xử lý học máy. Phương pháp này chuyển đổi các biến về dạng phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp cải thiện hiệu quả và độ chính xác của các thuật toán.

2.10.2. Nền tảng toán học

StandardScaler biến đổi dữ liệu theo phương trình:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{5}$$

Trong đó:

- X là giá trị gốc của dữ liệu
- X_{scaled} là giá trị sau khi được chuẩn hóa
- μ là giá trị trung bình của dữ liệu
- σ là độ lệch chuẩn của dữ liệu.

2.10.3. Lý do sử dụng

StandardScaler thường được sử dụng trong tiền xử lý dữ liệu để cải thiện hiệu suất và ổn định quá trình huấn luyện mô hình học máy. Trong phân tích thống kê, chuẩn hóa giúp so sánh các biến có đơn vị và phân phối khác nhau, làm rõ mối quan hệ và ảnh hưởng của chúng đến kết quả.

2.10.4. Ưu điểm và hạn chế

Ưu điểm: Chuẩn hóa dữ liệu cải thiện hiệu suất, ổn định mô hình học máy và giảm ảnh hưởng của các biến không cần thiết, đặc biệt với các thuật toán nhạy cảm với thang đo dữ liệu.

Hạn chế: Chuẩn hóa dữ liệu có thể làm mất mát thông tin, đặc biệt là trong các trường hợp mà phân phối của dữ liệu không phải là phân phối chuẩn. Có thể làm cho dữ liệu trở nên khó hiểu và diễn giải.

2.11. Xử lý nhiễu Winsorization

2.11.1. Giới thiệu về Winsorization

Winsorization [16] là một kỹ thuật xử lý dữ liệu nhằm giảm thiểu ảnh hưởng của các giá trị ngoại lai (outliers) trong tập dữ liệu. Thay vì loại bỏ hoàn toàn các giá trị ngoại lai, phương pháp này giới hạn chúng bằng cách thay thế các giá trị cực trị bằng một ngưỡng cụ thể, thường dựa trên các phân vị (percentile) của dữ liệu.

2.11.2. Lý do sử dụng

Winsorization giúp giảm tác động của các giá trị ngoại lai, vốn có thể làm sai lệch trung bình và độ lệch chuẩn, mà không cần loại bỏ dữ liệu. Phương pháp này cải thiện tính ổn định của các mô hình học máy và thống kê, giúp kết quả chính xác hơn. Đồng thời, Winsorization bảo toàn dữ liệu bằng cách thay thế các giá trị cực trị, giữ nguyên số lượng mẫu trong phân tích.

2.11.3. Ưu điểm và hạn chế

Ưu điểm: Giảm ảnh hưởng của các giá trị ngoại lai mà không loại bỏ dữ liệu. Giữ nguyên cấu trúc tổng thể của dữ liệu, bao gồm cả các đặc điểm phân phối. Đơn giản và dễ thực hiện trong các bước tiền xử lý dữ liệu.

Hạn chế: Có thể làm mất thông tin quan trọng nếu ngoại lai chứa ý nghĩa đặc biệt trong ngữ cảnh dữ liệu. Phương pháp này phụ thuộc vào ngưỡng phân vị được chọn, có thể dẫn đến việc thay đổi quá mức dữ liệu. Không phù hợp với các phân tích yêu cầu tính toàn vẹn của dữ liệu thô.

2.12. Xử lý nhiễu Isolation Forest

2.12.1. Giới thiệu về Isolation Forest

Isolation Forest [17] là một thuật toán phát hiện ngoại lai (outlier detection) được thiết kế đặc biệt cho các tập dữ liệu lớn và đa chiều. Thuật toán dựa trên ý

tưởng cô lập (isolate) các điểm dữ liệu ngoại lai thông qua cây phân chia dữ liệu (isolation trees), giúp phát hiện các giá trị bất thường một cách hiệu quả và nhanh chóng.

2.12.2. Nền tảng toán học

2.12.3. Lý do sử dụng

Isolation Forest được sử dụng vì khả năng phát hiện ngoại lai hiệu quả, đặc biệt với các tập dữ liệu lớn và đa chiều. Thuật toán có thời gian tính toán tuyến tính theo số lượng mẫu, giúp nó phù hợp cho các ứng dụng thời gian thực. Hơn nữa, Isolation Forest không yêu cầu dữ liệu tuân theo bất kỳ phân phối cụ thể nào, tạo sự linh hoạt vượt trội so với nhiều phương pháp phát hiện ngoại lai khác.

2.12.4. Ưu điểm và hạn chế

Ưu điểm: Isolation Forest có hiệu quả cao với thời gian tính toán tuyến tính, phù hợp cho cả tập dữ liệu lớn và đa chiều. Phương pháp này hoạt động tự động, không yêu cầu giả định về phân phối dữ liệu, giúp tăng tính linh hoạt. Ngoài ra, thuật toán đạt độ chính xác cao trong việc phát hiện các giá trị ngoại lai ngay cả trên dữ liệu phức tạp và không cân đối.

Hạn chế: Việc phân chia ngẫu nhiên trong Isolation Forest đôi khi dẫn đến kết quả không ổn định, đặc biệt khi số lượng cây được sử dụng ít. Phương pháp này cũng không tối ưu khi áp dụng cho các tập dữ liệu rất nhỏ, khiến hiệu quả giảm đi đáng kể. Hơn nữa, kết quả từ Isolation Forest thường khó giải thích trực quan so với một số phương pháp khác, gây khó khăn trong việc diễn giải và phân tích.

2.13. Phân cụm DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

2.13.1. Giới thiệu về thuật toán DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [18] là một thuật toán phân cụm dựa trên mật độ, được giới thiệu bởi Martin Ester và cộng sự vào năm 1996. Thuật toán này không yêu cầu biết trước số lượng cụm và có khả năng xác định các điểm nhiễu (noise) trong dữ liệu. DBSCAN hoạt động dựa trên việc xác định các cụm dựa trên mật độ của các điểm dữ liệu trong không gian.

2.13.2. Nền tảng toán học

DBSCAN dựa trên hai khái niệm chính:

1. **Mật độ lân cận:** Thuật toán sử dụng hai tham số:
   - ε: Bán kính để xác định vùng lân cận của một điểm.
   - MinPts: Số lượng điểm tối thiểu cần thiết để một vùng được xem là dày đặc.
2. **Phân loại điểm dữ liệu:**
   - **Core points**: Là các điểm có ít nhất MinPts điểm lân cận trong bán kính ε.
   - **Border points**: Là các điểm không phải core point nhưng nằm trong vùng lân cận của một core point.
   - **Noise points**: Là các điểm không thuộc cụm nào (ngoài bán kính ε của mọi core point).

2.13.3. Công thức toán học

1. **Vùng lân cận của một điểm p:**

$$N(p) = \{q \in D \mid \text{distance}(p, q) \leq \varepsilon\}$$

   Trong đó:
   - $N(p)$: Tập các điểm trong vùng lân cận ε của p.
   - distance(p, q): Khoảng cách giữa p và q, có thể là khoảng cách Euclid hoặc các loại khoảng cách khác.
2. **Điều kiện để p là core point:**

$$|N(p)| \geq \text{MinPts}$$

   Trong đó:
   - $|N(p)|$: Số lượng điểm trong vùng lân cận ε của p.

2.13.4. Diễn giải thuật toán

Thuật toán DBSCAN bắt đầu bằng cách duyệt qua từng điểm dữ liệu chưa được thăm. Với mỗi điểm, thuật toán kiểm tra xem nó có đủ số lượng điểm lân cận tối thiểu $MinPts$ trong bán kính hay không. Nếu đủ, điểm đó trở thành một điểm lõi (core point) và khởi tạo một cụm mới. Thuật toán sau đó mở rộng cụm bằng cách thêm các điểm lân cận, tiếp tục kiểm tra các điểm lõi mới và lặp lại quá trình cho đến khi không thể mở rộng thêm. Điểm không phải lõi nhưng nằm gần cụm

sẽ được gán là điểm biên (border point), còn điểm không thuộc cụm nào sẽ được đánh dấu là nhiễu (noise). Quá trình lặp lại cho đến khi toàn bộ các điểm được xử lý, kết quả là các cụm dựa trên mật độ cùng các điểm nhiễu được xác định.

2.13.5. Ưu điểm và hạn chế

Ưu điểm: DBSCAN không cần xác định trước số cụm, có thể phát hiện điểm nhiễu và xử lý tốt các cụm có hình dạng phức tạp. Thuật toán hoạt động hiệu quả với dữ liệu lớn và không phụ thuộc nhiều vào tỷ lệ đặc trưng nếu sử dụng khoảng cách phù hợp.

Hạn chế: Hạn chế: DBSCAN nhạy cảm với tham số và $MinPts$, hoạt động kém với dữ liệu chiều cao hoặc các cụm có mật độ thay đổi. Thuật toán cũng gặp khó khăn khi dữ liệu có nhiều nhiễu hoặc cụm gần nhau.

2.14. Ứng dụng

- Phát hiện bất thường: Được sử dụng để phát hiện gian lận, lỗi hệ thống hoặc điểm bất thường trong dữ liệu.
- Xử lý hình ảnh và sinh học: Phân cụm dữ liệu hình ảnh, dữ liệu gen.
- Khai phá văn bản: Phân cụm tài liệu hoặc thông tin tương tự.
- Robotics và IoT: Lập bản đồ, định vị robot, phân tích dữ liệu cảm biến.

2.15. Khai thác tập phổ biến tối đại FP-Max

2.15.1. Giới thiệu về FP-Max

FP-Max (Frequent Pattern Maximum) [19] là một thuật toán trong khai phá dữ liệu, được sử dụng để khai thác tập phổ biến tối đại (maximal frequent itemsets) từ một tập dữ liệu giao dịch. Tập phổ biến tối đại là tập con lớn nhất trong các tập phổ biến mà không có tập phổ biến nào khác là tập con của nó. FP-Max là một biến thể của thuật toán FP-Growth, tối ưu hóa việc giảm số lượng tập phổ biến được sinh ra, giúp tiết kiệm không gian lưu trữ và giảm thời gian xử lý.

2.15.2. Nền tảng toán học

Một tập phổ biến I được gọi là **tối đại** nếu:

$$I' \supset I \text{ mà } \text{Support}(I') \geq \text{min\_support}. \tag{6}$$

Trong đó: Support(I): Số lượng giao dịch chứa I, biểu diễn độ phổ biến của tập I. min_support: Ngưỡng hỗ trợ tối thiểu được chỉ định trước.

FP-Max sử dụng cây *FP-tree* (Frequent Pattern Tree) để biểu diễn dữ liệu giao dịch theo cách nén gọn, giúp khai thác các tập phổ biến trực tiếp mà không cần quét lại toàn bộ dữ liệu.

2.15.3. Diễn giải thuật toán

Thuật toán FP-Max gồm ba bước chính. Đầu tiên, xây dựng FP-tree để nén dữ liệu giao dịch thành một cấu trúc cây, lưu thông tin về tần suất và các mục. Tiếp theo, khai thác tập phổ biến từ FP-tree bằng chiến lược đệ quy. Cuối cùng, lọc ra các tập phổ biến tối đại bằng cách loại bỏ các tập con bị bao phủ bởi tập lớn hơn. FP-Max khác FP-Growth ở chỗ kiểm tra điều kiện tối đại ngay trong quá trình khai thác, giúp giảm số lượng tập phổ biến cần xử lý.

2.15.4. Lý do sử dụng

Tập trung vào các tập phổ biến lớn nhất, giảm số lượng tập phổ biến không cần thiết. Xử lý các tập dữ liệu giao dịch lớn mà các thuật toán truyền thống như Apriori không hiệu quả. Tăng hiệu suất trong các bài toán như phân tích hành vi người dùng, quản lý giỏ hàng hoặc phát hiện mẫu trong dữ liệu lớn.

2.15.5. Ưu điểm và hạn chế

Ưu điểm: Giảm số lượng tập phổ biến được sinh ra, tiết kiệm không gian lưu trữ và thời gian xử lý. Không yêu cầu quét lại toàn bộ dữ liệu nhiều lần như thuật toán Apriori. Sử dụng FP-tree để nén dữ liệu, tăng hiệu quả trong việc xử lý tập dữ liệu lớn.

Hạn chế: Tốn kém bộ nhớ nếu tập dữ liệu có nhiều mục hoặc giao dịch tương tự nhau, dẫn đến FP-tree lớn. Cần kiểm tra điều kiện tối đại, có thể phức tạp khi số lượng tập phổ biến lớn. Đòi hỏi hiểu biết sâu hơn về thuật toán và cấu trúc dữ liệu FP-tree để triển khai hiệu quả.

2.16. Khoảng cách Gower

2.16.1. Giới thiệu về Gower

Khoảng cách Gower [20] được thiết kế đặc biệt để xử lý các tập dữ liệu chứa nhiều loại biến khác nhau như biến số liên tục, biến phân loại, biến thứ tự, và biến nhị phân. Phương pháp này tính toán khoảng cách giữa hai đối tượng bằng cách kết hợp khoảng cách riêng lẻ trên từng thuộc tính, sau đó chuẩn hóa và tổng hợp chúng. Nhờ tính linh hoạt và khả năng xử lý nhiều loại biến, khoảng cách Gower

thường được sử dụng trong các bài toán phân cụm và so sánh sự tương đồng trong tập dữ liệu hỗn hợp.

2.16.2. Nền tảng toán học

Khoảng cách Gower được tính toán như một trung bình có trọng số của các độ tương đồng theo từng biến. Công thức tổng quát để tính độ tương đồng giữa hai đối tượng i và j là:

$$d_{ij} = 1 - s_{ij}$$

Trong đó, s_{ij} là hệ số tương đồng giữa hai đối tượng i và j, được xác định bởi:

$$s_{ij} = \frac{\sum_{k=1}^{p} w_k \cdot s_{ij}^{(k)}}{\sum_{k=1}^{p} w_k}$$

Các thành phần trong công thức:

- p: Số lượng biến trong tập dữ liệu.
- w_k: Trọng số của biến thứ k, thường bằng 1 nếu không có giá trị bị thiếu.
- $s_{ij}^{(k)}$: Độ tương đồng giữa hai đối tượng theo biến k, được tính theo từng loại biến:
  - **Biến định lượng (continuous):**

$$s_{ij}^{(k)} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{R_k}$$

    Trong đó, $R_k = \max(x^{(k)}) - \min(x^{(k)})$ là khoảng giá trị của biến k.
  - **Biến định tính (categorical):**

$$s_{ij}^{(k)} = \begin{cases} 1, & \text{nếu } x_i^{(k)} = x_j^{(k)} \\ 0, & \text{nếu } x_i^{(k)} \neq x_j^{(k)} \end{cases}$$

  - **Biến nhị phân (binary):** Áp dụng tương tự biến định tính.

Khoảng cách Gower luôn nằm trong khoảng $[0, 1]$, trong đó giá trị càng nhỏ thì hai đối tượng càng tương đồng.

2.16.3. Lý do sử dụng

Khoảng cách Gower [20] được sử dụng vì khả năng xử lý dữ liệu hỗn hợp, cho phép đo lường khoảng cách trên các tập dữ liệu chứa cả biến liên tục và biến phân loại. Phương pháp này có khả năng chuẩn hóa từng loại biến, đảm bảo rằng không loại biến nào chiếm ưu thế trong quá trình tính toán. Ngoài ra, Gower rất linh hoạt, dễ dàng mở rộng và điều chỉnh để phù hợp với các loại biến mới hoặc áp dụng trọng số cụ thể theo yêu cầu của từng bài toán.

2.16.4. Ưu điểm và hạn chế

Ưu điểm: Khoảng cách Gower hỗ trợ nhiều loại biến trong cùng một tập dữ liệu, từ biến liên tục đến biến phân loại, với công thức đơn giản, dễ áp dụng. Phương pháp này tự động chuẩn hóa các biến, đảm bảo tính nhất quán và giảm ảnh hưởng của các đơn vị đo lường khác nhau.

Việc tính toán khoảng cách Gower có thể tốn thời gian khi dữ liệu lớn hoặc nhiều chiều. Kết quả phụ thuộc vào trọng số, và việc xác định trọng số phù hợp đôi khi phức tạp. Ngoài ra, phương pháp này không hiệu quả với các quan hệ phi tuyến giữa các biến.

Chương 3. PHƯƠNG PHÁP THỰC HIỆN

3.1. Phương pháp thu thập dữ liệu

Tôi đã thực hiện quá trình thu thập dữ liệu từ trang web thương mại điện tử Tiki.vn thông qua việc sử dụng API request và trích xuất file JSON từ API của trang web. Quá trình này được thực hiện theo các bước sau:

3.1.1. Truy Xuất Thông Tin Cửa Hàng

- Sử dụng API request để truy xuất thông tin về các cửa hàng trên Tiki.
- Bằng cách lấy curl (chứa headers và params) của một sản phẩm ngẫu nhiên trên nền tảng Tiki, tôi xác định đường dẫn chung của sản phẩm và trích xuất file JSON để thu thập danh sách các ID sản phẩm từ đường dẫn đã xác định.
- Tiếp theo, thu thập thông tin từ API của cửa hàng, bao gồm "id", "name", và "link".

3.1.2. Thu Thập Thông Tin Sản Phẩm

- Đối với mỗi cửa hàng, tiếp tục lấy curl từ API của cửa hàng, sau đó chuyển mã curl đã thu thập về mã Python, sử dụng vòng lặp để lấy ra ID của từng sản phẩm trong cửa hàng đó.
- Chọn một sản phẩm bất kỳ từ danh sách sản phẩm của cửa hàng để lấy headers và params từ API của sản phẩm đó.
- Đối với mỗi cửa hàng, tiếp tục lấy curl từ API của cửa hàng, sau đó chuyển mã curl đã thu thập về mã Python, sử dụng vòng lặp để lấy ra ID của từng sản phẩm trong cửa hàng đó.
- Chọn một sản phẩm bất kỳ từ danh sách sản phẩm của cửa hàng để lấy headers và params từ API của sản phẩm đó.
- Trích xuất dữ liệu và thu thập thông tin như "id", "name", "price", "rating_average", "review_count", "quantity_sold", "quantity_sold_2weeks" của mỗi sản phẩm trong cửa hàng. Quá trình này lặp lại cho tối đa 95 sản phẩm trong mỗi cửa hàng.

3.1.3. Thu Thập Đánh Giá Khách Hàng và Thông Tin Khác

- Tiếp tục lấy curl (chứa headers và params) từ API chứa thông tin về đánh giá của sản phẩm.

- Áp dụng phương pháp tương tự để thu thập các đánh giá từ khách hàng.
- Thu thập thông tin về "Name_Shop", “Shop_Rating”, “Year_Joined”, “Follower”, và “Chat_Response” cũng được thực hiện tương tự như trên.
- Qua phương pháp này, tôi có thể thu thập dữ liệu đa dạng và chi tiết về các cửa hàng, sản phẩm và đánh giá từ người dùng trên nền tảng Tiki.vn.

3.2. Mô tả dữ liệu ban đầu

Tên biến	Mô tả	Kiểu dữ liệu
Id	Id của từng sản phẩm	int
QuantitySold	Số lượng bán của sản phẩm	int
QuantitySoldMonth	Số lượng bán của sản phẩm sau 1 tháng	int
Price	Giá từng sản phẩm	float
YearJoined	Năm tham gia bán hàng của cửa hàng	int
ShopRating	Số sao trung bình cửa hàng	float
CounterRating	Số lượng chấm điểm để ra số sao trung bình	int
Followers	Số người theo dõi	int
ChatResponse	Tỷ lệ phản hồi chat	float
Feedbacks	Các bình luận của khác hàng về sản phẩm	string

Bảng 3.1: Bảng thông tin về các biến

Dữ liệu được thu thập ban đầu bao gồm 73300 mẫu về thông tin của cửa hàng, thông tin sản phẩm và các bình luận của khách hàng về chất lượng sản phẩm. Dữ liệu được thu thập gồm 10 đặc chưng, bao gồm các thông tin về:

3.3. Tiền xử lý dữ liệu

3.3.1. Tiền xử lý dữ liệu "Feedbacks"

Chuẩn hóa ký tự đặc biệt và emoji: Quá trình tiền xử lý dữ liệu được thực hiện bằng cách thay thế các dấu như ".", ",", "\n" (xuống dòng) bằng dấu "." và loại bỏ tất cả các ký tự đặc biệt như: `!, @, #, %, ▪ &, ....` Tiếp theo, các biểu tượng cảm xúc như hình 3.1 sẽ được loại bỏ khỏi mỗi đoạn văn. Mục đích của quy trình này là làm sạch các đoạn văn và chuẩn bị dữ liệu sẵn sàng để sử dụng trong thư viện `underthesea`.

Hình 3.1: Hình ảnh minh họa Emoji.

Chuẩn hóa dữ liệu Tiếng Việt: Sử dụng thư viện `underthesea` để chuẩn hóa dữ liệu tiếng Việt. Quá trình chuẩn hóa này giúp loại bỏ các yếu tố không mong muốn từ văn bản như dấu câu, ký tự đặc biệt và biểu tượng cảm xúc. Ví dụ, việc sử dụng có thể sửa chữa các lỗi chính tả như "Đảm bao chất lựa chọn phòng thí nghiệm hóa học" thành "Đảm bảo chất lượng phòng thí nghiệm hóa học". Điều này giúp tạo ra một tập dữ liệu thuần khiết hơn, giúp tăng hiệu suất ở bước phân loại cảm xúc văn bản ở bước sau.

Tách câu: Vì phân loại cảm xúc nhờ thư viện `underthesea` đạt hiệu suất cao với các câu riêng lẻ nên việc tách các câu bình luận dài từ khách hàng là điều cần thiết. Ở bài báo cáo này, tôi sử dụng thư viện `underthesea` để tách các bình luận dài của khách hàng thành các đoạn ngắn hơn. Điều này tăng cơ hội hiểu đúng ngữ cảnh và nội dung của bình luận. Ví dụ, đối với một đánh giá của khách hàng "Rất tuyệt vời....giá cả hợp lý,.thành phần hữu ít vk em rất thích...cám ơn shop bán và ứng dụng tiki", chúng ta sẽ phân tách thành các câu như sau: "Rất tuyệt vời", "giá cả hợp lý", "thành phần hữu ít vk em rất thích", và "cám ơn shop bán và ứng dụng tiki".

Phân loại cảm xúc văn bản: Phân loại cảm xúc trong bộ dữ liệu này xác định số lượng bình luận tích cực và tiêu cực của từng sản phẩm, hỗ trợ phân cụm và ra quyết định chiến lược. Tôi sử dụng hàm `sentiment()` từ thư viện `underthesea` để phân loại cảm xúc từng câu, sau đó áp dụng phương pháp voting để xác định cảm xúc (tích cực hoặc tiêu cực). Kết quả thu được ba cột mới: số lượng đánh giá tích cực, tiêu cực và tổng số đánh giá của từng sản phẩm. Cuối cùng, tôi tính tổng số đánh giá tích cực, tiêu cực và tổng số đánh giá cho từng cửa hàng.

3.3.2. Ước tính doanh thu cửa hàng

Phân đoạn này mô tả quá trình tính toán giữa các cột để ước tính doanh thu cửa hàng sau 1 tháng giúp ta có thể đánh giá hiệu suất kinh doanh của cửa hàng trong khoảng thời gian cụ thể này.

$$\text{Revenue}_{\text{store}} = \sum \left((\text{QuantitySold}_{1 \text{ month}} - \text{QuantitySold}_{\text{initial}}) \cdot \text{price}_{\text{product}} \right)$$

Trong đó: Revenue_{store} là doanh thu ước tính của từng cửa hàng trong 1 tháng, $\text{QuantitySold}_{initial}$ số lượng bán sau thu thập lần đầu, $\text{QuantitySold}_{1month}$ là số lượng bán sau thu thập lần đầu sau 1 tháng, $\text{price}_{product}$ là giá sản phẩm cửa hàng bán.

Cuối cùng, ta sinh ra 1 đặc trưng dữ liệu mới là *Revenue* hay "Doanh thu cửa hàng".

3.3.3. Đo lường chất lượng đặc trưng bằng Wilson score Interval

RatingQuality: là sự kết hợp giữa hai đặc trưng `ShopRating` và `CounterRating`. Phương pháp `Wilson score Interval` được sử dụng để đánh giá độ tin cậy của các cửa hàng, nhằm tạo ra thứ hạng khách quan hơn. Cụ thể, các cửa hàng có số lượng đánh giá lớn hơn và đáng tin cậy hơn sẽ được ưu tiên. Ví dụ, một cửa hàng có `ShopRating` là 5.0 nhưng chỉ có 10 `CounterRating` sẽ có độ tin cậy thấp. Trong khi đó, một cửa hàng khác có `ShopRating` là 4.8 nhưng có 100 `CounterRating` sẽ đáng tin cậy hơn và được ưu tiên hơn trong quá trình đánh giá.

PositiveQuality: được thực hiện bằng cách kết hợp hai biến `Positive` và `TotalFeedback`. Quá trình này ưu tiên các cửa hàng có số lượng bình luận tích cực gần với tổng số lượng bình luận. Ví dụ, một cửa hàng có `Positive` là 30 trên tổng số 50 TotalFeedback sẽ có độ tin cậy cao hơn so với một cửa hàng khác cũng có `Positive` là 30 trên 200 `TotalFeedback`. Ta không áp dụng với cột `Negative` vì sẽ xảy ra tương quan cao giữa 2 biến, làm tăng số lượng biến không cần thiết cho quá trình phân cụm.

Cuối cùng, ta có thêm 2 đặc trưng mới: *RatingQuality* và *PositiveQuality*.

3.4. Bộ dữ liệu sau khi tiền xử lý

Sau khi hoàn thành quá trình tiền xử lý dữ liệu ban đầu, chúng ta thu được bộ dữ liệu mới, là cơ sở chính để thực hiện phân tích và áp dụng các mô hình học máy.

Dữ liệu lúc này bao gồm 933 mẫu và 6 đặc trưng chính về thông tin của các cửa hàng trên sàn thương mại điện tử như **Bảng 3.2**.

Tên biến	Mô tả	Kiểu dữ liệu
Revenue	Doanh thu ước tính của cửa hàng	int
PositiveQuality	Tỷ lệ chất lượng đánh giá là tích cực của cửa hàng	float
YearJoined	Năm tham gia bán hàng của cửa hàng	int
RatingQuality	Chất lượng số sao trung bình của cửa hàng	float
Followers	Số lượng người theo dõi	int
ChatResponse	Tỷ lệ phản hồi chat	float

Bảng 3.2: Các đặc trưng phân cụm

3.5. Chuẩn hóa dữ liệu

Revenue	YearJoined	Followers	ChatResponse	RatingQuality	PositiveQuality
-0.326840	-0.279443	-0.143533	-0.819518	0.517466	0.731572
-0.143226	-0.972848	0.031235	-0.819518	0.539159	0.189270
-0.260906	0.413962	1.358861	1.525984	0.663007	1.056228
-0.313024	1.107368	-0.313240	-0.819518	-0.090066	0.500853
-0.287069	0.413962	-0.284224	1.525984	0.379822	0.768611

Bảng 3.3: Bảng dữ liệu sau khi áp dụng StandardScaler

Trong báo cáo này, tôi sử dụng phương pháp `StandardScaler` để chuẩn hóa dữ liệu, nhằm tối ưu hóa hiệu suất và độ chính xác của các thuật toán phân cụm. Quá trình chuẩn hóa giúp loại bỏ sự chênh lệch về quy mô và đơn vị đo lường giữa các đặc trưng, tạo điều kiện thuận lợi cho các thuật toán phân cụm, đặc biết là các thuật toán dựa trên khoảng cách hoạt động hiệu quả hơn.

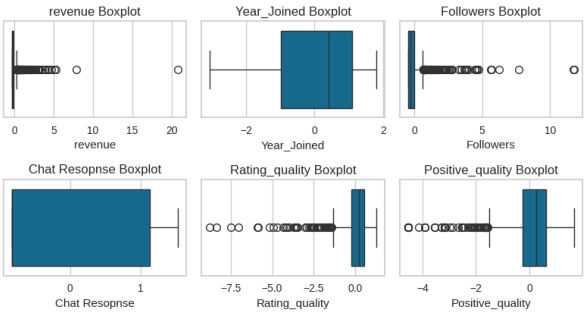
Ta thấy, ở **Bảng 3.3** dữ liệu đã được đưa về cùng thang đo.

3.6. Xử lý nhiễu (Outliers)

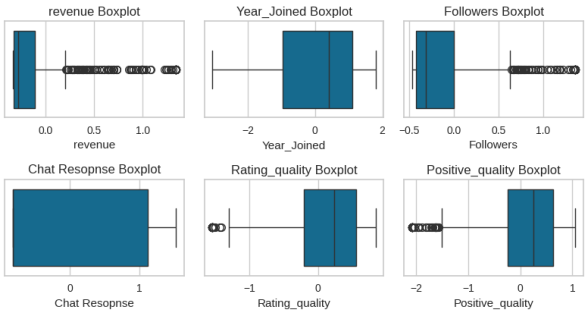
3.6.1. Xử lý nhiễu bằng Winsorization

Phương pháp Winsorization là kỹ thuật xử lý nhiễu bằng cách thay thế các giá trị ngoại lệ (outliers) nằm ngoài phạm vi chấp nhận được bằng các giá trị trong phạm vi này. Điều này giúp giảm ảnh hưởng của các giá trị cực đoan mà không làm mất quá nhiều thông tin, từ đó cải thiện tính chính xác trong phân tích thống kê và mô hình học máy.

Ở **Hình 3.2**, ta thấy các giá trị ngoại lệ nằm xa các nhóm dữ liệu chính, điều này có thể làm sai lệch kết quả đặc biệt là với các thuật toán phân cụ dựa trên



Hình 3.2: Dữ liệu ban đầu.

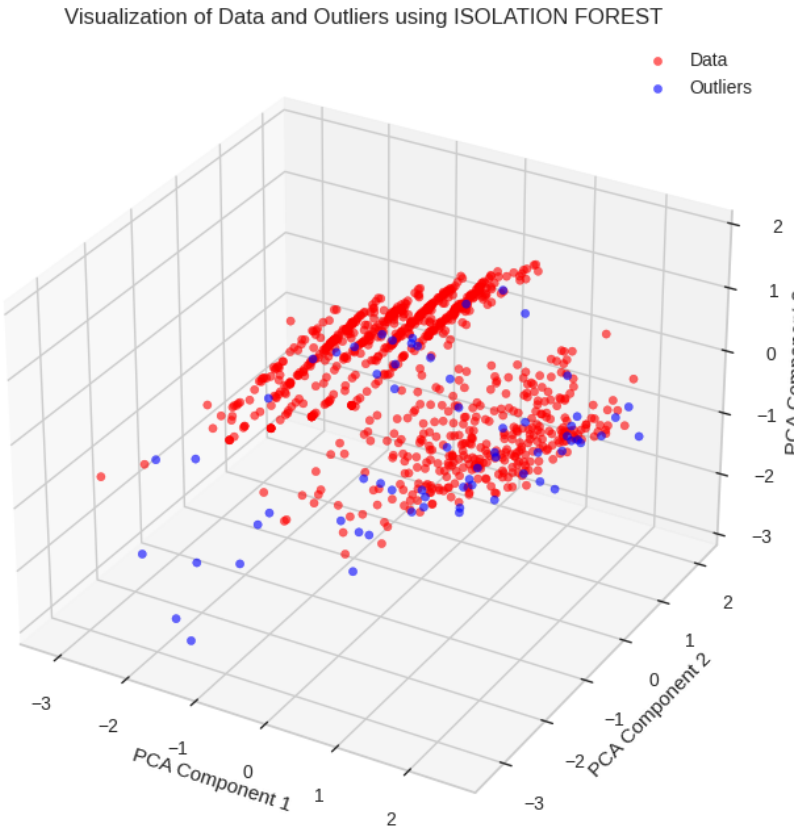


Hình 3.3: Dữ liệu sau khi xử lý outliers.

khoảng cách.

Ở **Hình 3.3** sau khi áp dụng Winsorization, các ngoại lệ được thay thế bằng giá trị trong phạm vi chấp nhận (5-95%), giúp giảm ảnh hưởng của outliers mà không làm mất thông tin .

3.6.2. Xử lý nhiễu bằng Isolation Forest



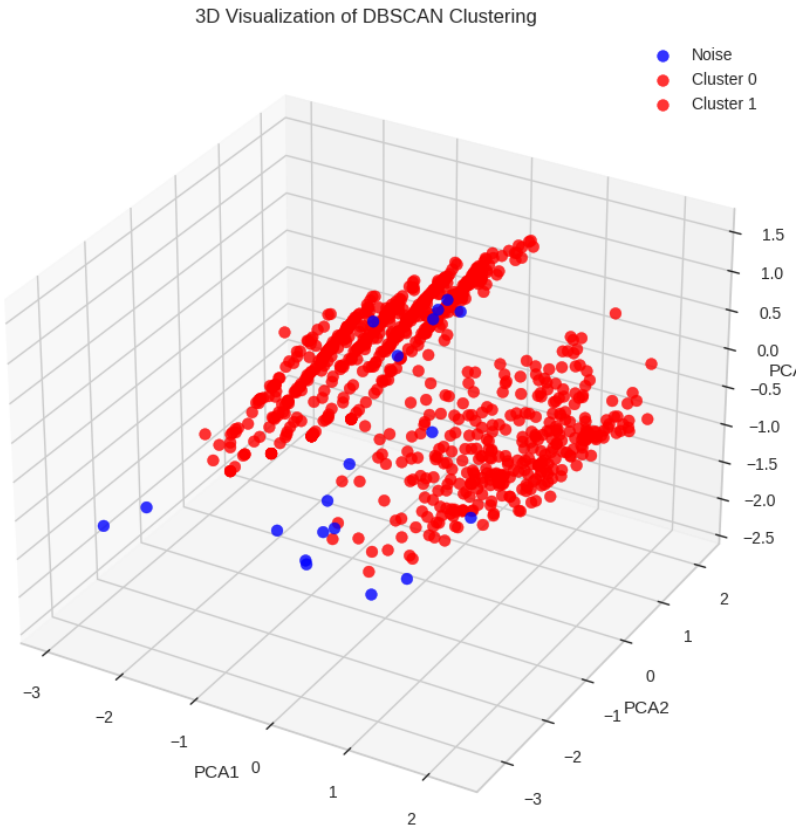
Hình 3.4: Dữ liệu sau khi xử lý outliers bằng Isolation Forest.

Phương pháp Isolation Forest là một kỹ thuật hiệu quả để phát hiện và loại bỏ các giá trị ngoại lệ (outliers) trong dữ liệu. Sau khi sử dụng Winsorization để giảm thiểu ảnh hưởng của các giá trị cực trị, Isolation Forest sẽ tiếp tục tìm và loại

bỏ những điểm ngoại lệ còn lại. Thuật toán này hoạt động bằng cách xây dựng các cây quyết định, tách biệt các điểm dữ liệu bất thường, vì các ngoại lệ sẽ dễ dàng bị cô lập khỏi phần lớn dữ liệu còn lại.

Hình 3.4 xác định các điểm ngoại lệ trong dữ liệu bằng các chấm màu xanh (contamination = 0.7). Các điểm này sẽ bị loại bỏ khỏi bộ dữ liệu sau khi áp dụng Isolation Forest. Việc này giúp làm sạch dữ liệu, giảm thiểu ảnh hưởng của các điểm bất thường.

3.6.3. Xử lý nhiễu bằng DBSCAN

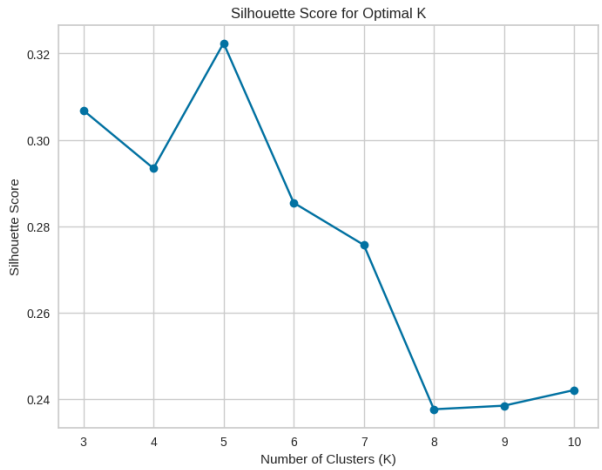


Hình 3.5: Dữ liệu sau khi xử lý outliers bằng DBSCAN.

Trong báo cáo này, sau khi áp dụng các phương pháp xử lý nhiễu ban đầu như Winsorization và Isolation Forest, DBSCAN được sử dụng để kiểm tra lại các điểm dữ liệu ngoại lệ còn sót lại. Cuối cùng ta tiến hành loại bỏ các ngoại lệ cuối cùng đó. Tôi áp dụng 2 tham số sau: minpts = 12 (2 * số chiều dữ liệu (6)) và eps = 1.05 (nơi mà đường cong đồ thị k-distance graph không thay đổi đáng kể).

3.7. Chọn số lượng cụm tối ưu

Phương pháp thực nghiệm này nhằm chọn số lượng cụm tối ưu từ 3 đến 10 cho ba thuật toán phân cụm khác nhau, bao gồm *K-Means* với `init = 'k-means++'`, *Agglomerative Clustering* với `linkage = 'ward'`, và *Gaussian Mixture Model (GMM)* với `covariance_type = 'spherical'`. Để xác định số lượng cụm tối ưu, ta sử dụng chỉ số *Silhouette Score*, một tiêu chí đánh giá sự phân tách của các cụm. Silhouette Score đo lường mức độ tương đồng của các điểm trong mỗi cụm so với các điểm trong các cụm khác, giúp lựa chọn số cụm cho kết quả phân cụm tốt nhất. Số cụm có giá trị Silhouette cao nhất sẽ được chọn là tối ưu cho mỗi thuật toán.



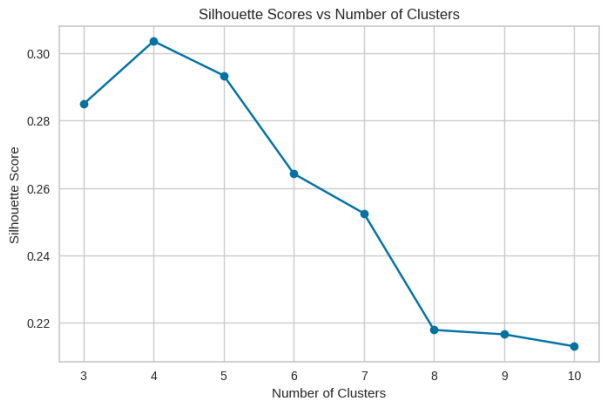
Hình 3.6: Điểm Silhouette từng cụm.

Number of Clusters	Silhouette Score
3	0.273251
4	0.300402
5	0.329759
6	0.298026
7	0.281516
8	0.240029
9	0.251132
10	0.233334

Hình 3.7: Bảng chỉ số Silhouette

3.7.1. Chọn số lượng cụm tối ưu cho K-Means

- Chỉ số Silhouette trung bình ở **Hình 3.6** và **Hình 3.7** cho thấy số cụm tối ưu khi sử dụng K-Means là 5.



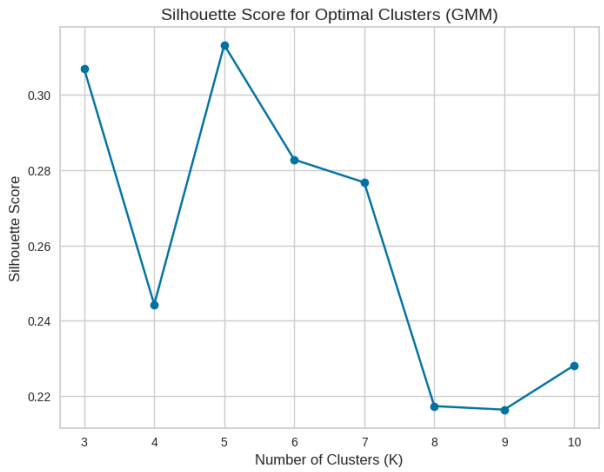
Hình 3.8: Điểm Silhouette từng cụm.

Number of Clusters	Silhouette Score
3	0.273251
4	0.300402
5	0.329759
6	0.298026
7	0.281516
8	0.240029
9	0.251132
10	0.233334

Hình 3.9: Bảng chỉ số Silhouette

3.7.2. Chọn số lượng cụm tối ưu cho Agglomerative Đánh giá:

- Chỉ số Silhouette trung bình ở **Hình 3.8** và **Hình 3.9** cho thấy số cụm tối ưu khi sử dụng Agglomerative là 4.



Hình 3.10: Điểm Silhouette từng cụm.

Number of Clusters	Silhouette Score
3	0.273251
4	0.300402
5	0.329759
6	0.298026
7	0.281516
8	0.240029
9	0.251132
10	0.233334

Hình 3.11: Bảng chỉ số Silhouette

3.7.3. Chọn số lượng cụm tối ưu cho Gaussian Mixture Model Đánh giá:

- Chỉ số Silhouette trung bình ở **Hình 3.10** và **Hình 3.11** cho thấy số cụm tối ưu khi sử dụng Gaussian Mixture Model là 3.

3.8. Áp dụng khai thác tập mục thường xuyên (FP-Max)

3.8.1. Rời rạc hóa dữ liệu

Sử dụng K-Bin Discretizer với tham số K-Means để phân chia các giá trị liên tục thành 3 đoạn: Low, Medium, High. Sau đó, chuyển đổi dữ liệu rời rạc thành dạng nhị phân (binary), tạo ra các cột tương ứng với các đoạn Low, Medium, và High.

3.8.2. Khai thác tập mục thường xuyên

Áp dụng thuật toán FP-Max để tìm các tập mục phổ biến tối đại trong dữ liệu, kết quả ra được 14 tập mục, nhỏ hơn so với việc sử dụng Apriori hay Fp-Growth (63 tập mục).

3.8.3. Tạo đặc trưng nhị phân mới

Ta tiến thành chuyển đổi các itemset đã tạo ra ở bước trên về dạng các đặc trưng nhị phân. Cụ thể, nếu các itemset thỏa mãn điều kiện, đặc trưng nhận giá trị 1. Ngược lại 0.

3.8.4. Chọn lựa các đặc trưng nhị phân vào phân cụm

Tôi sử dụng phương pháp Forward Selection để lựa chọn xem biến nhị phân nào có tác động tích cực tới quá trình phân cụm. Bằng cách lần lượt lựa chọn các biến nhị phân thêm vào bộ dữ liệu ban đầu, nếu biến nào làm tăng chỉ số Silhouette thì thêm vào, ngược lại tiến hành loại bỏ. Khoảng cách Gower được sử dụng để đánh giá bộ dữ liệu hỗn hợp sau khi được thêm các biến nhị phân vào.

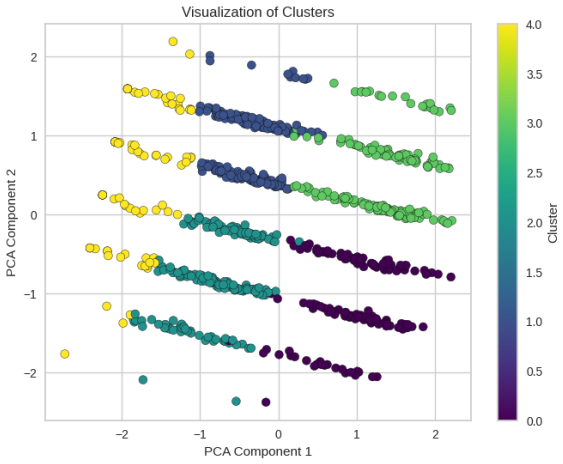
Số biến nhị phân được thêm vào 3 thuật toán lần lượt là 2 biến với K-Prototype, 3 biến với Aglomerative và 3 biến với GMM.

1. Các biến nhị phân được thêm cho K-Prototype:
    - ChatLow_YearLow_FollowersLow_revenueLow.
    - YearLow_PositiveHigh.
2. Các biến nhị phân được thêm cho Agglomerative.
    - RatingMedium_FollowersLow_ChatLow.
    - RatingHigh_revenueLow_FollowersLow_ChatLow.
    - revenueLow_RatingMedium_ChatLow.
3. Các biến nhị phân được thêm cho Gaussian Mixture Model:
    - ChatLow_revenueLow_PositiveHigh.
    - ChatLow_FollowersLow_PositiveHigh.
    - ChatLow_FollowersLow_PositiveMedium_revenueLow.

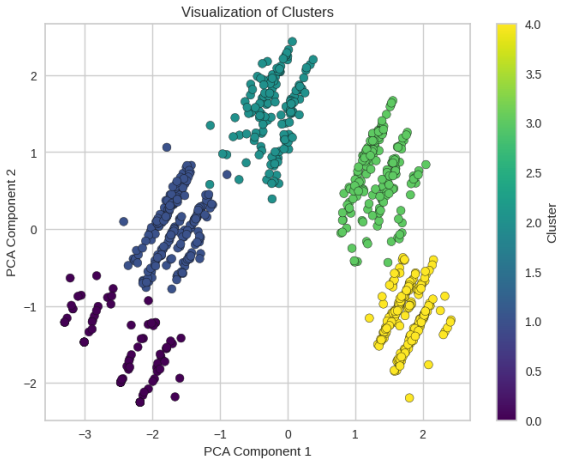
Chương 4. KẾT QUẢ THỰC NGHIỆM

4.1. Phân cụm trước và sau khi kết hợp khai thác tập mục thường xuyên

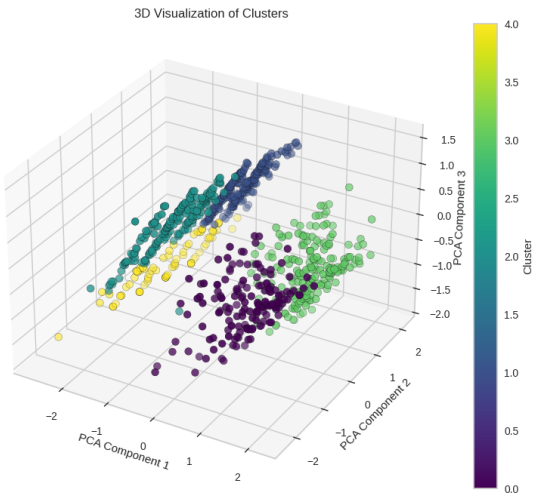
Sau khi tiến hành phân cụm, tôi so sánh trước và sau phân cụm kết hợp các đặc trưng nhị phân mới từ quá trình khai thác tập phổ biến tối đại Fp-Max, kết quả được thể hiện như bên dưới:



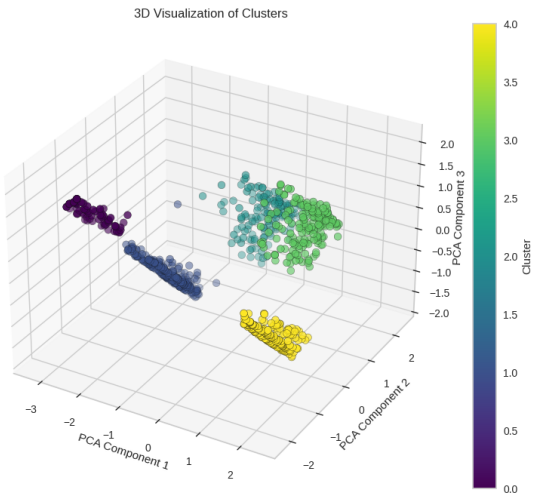
Hình 4.1: K-Means trước khi sử dụng FIM.



Hình 4.2: K-Prototypes sau khi sử dụng FIM.



Hình 4.3: K-Means trước khi sử dụng FIM.



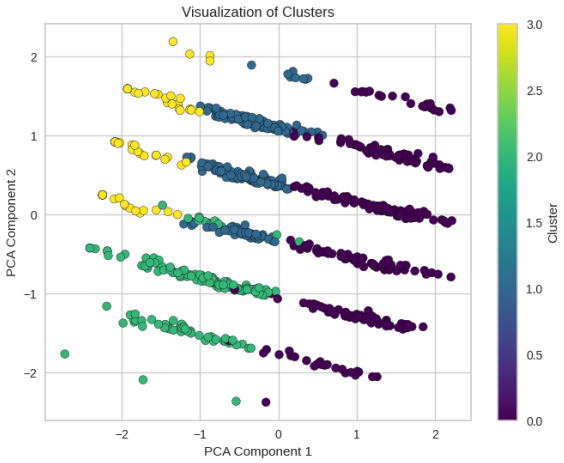
Hình 4.4: K-Prototypes sau khi sử dụng FIM.

4.1.1. Phân cụm K-Means và K-Prototypes

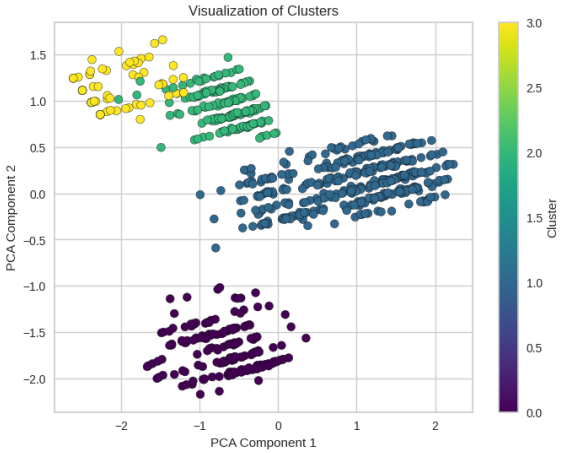
Bảng 4.1: Chỉ số đánh giá phân cụm K-Means và K-Prototype trước và sau kết hợp đặc trưng nhị phân

	Algorithms	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Before	K-Means	0.325	1.088	415.935
After	K-Means	0.459	0.963	495.393

Các chỉ số đánh giá ở **Bảng 4.1** cho thấy việc kết hợp đặc trưng nhị phân đã cải thiện rõ rệt chất lượng phân cụm của thuật toán K-Means. Silhouette và Calinski-Harabasz tăng đáng kể, khẳng định cấu trúc cụm rõ ràng và khoảng cách giữa các cụm được mở rộng, trong khi Davies-Bouldin giảm, phản ánh tính chặt chẽ và khả năng phân biệt cụm tốt hơn. Tóm lại, việc kết hợp đặc trưng nhị phân đã nâng cao chất lượng phân cụm, tạo kết quả rõ ràng và chính xác hơn.



Hình 4.5: Agglomerative trước khi sử dụng FIM.



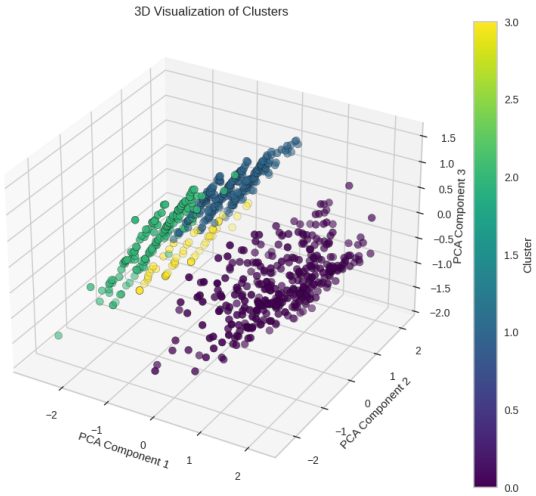
Hình 4.6: Agglomerative sau khi sử dụng FIM.

Bảng 4.2: Chỉ số đánh giá phân cụm Agglomerative trước và sau kết hợp đặc trưng nhị phân

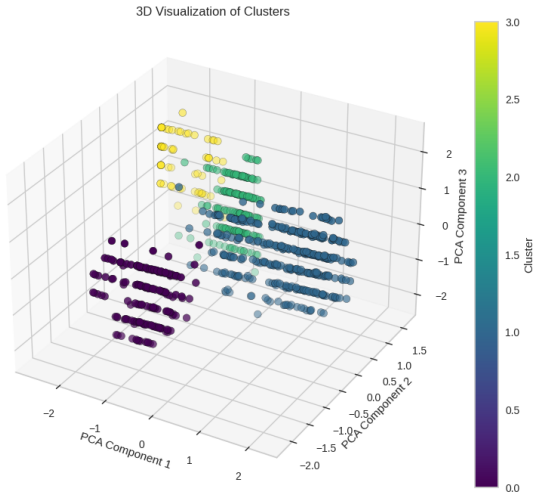
	Algorithms	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Before	Agglomerative	0.308	1.036	361.240
After	Agglomerative	0.553	0.628	1451.059

4.1.2. Phân cụm Agglomerative

Việc kết hợp đặc trưng nhị phân đã cải thiện đáng kể chất lượng phân cụm của thuật toán Agglomerative Clustering (**Bảng 4.2**). Silhouette tăng mạnh, thể hiện các cụm rõ ràng và tách biệt hơn; Davies-Bouldin giảm đáng kể, cho thấy tính

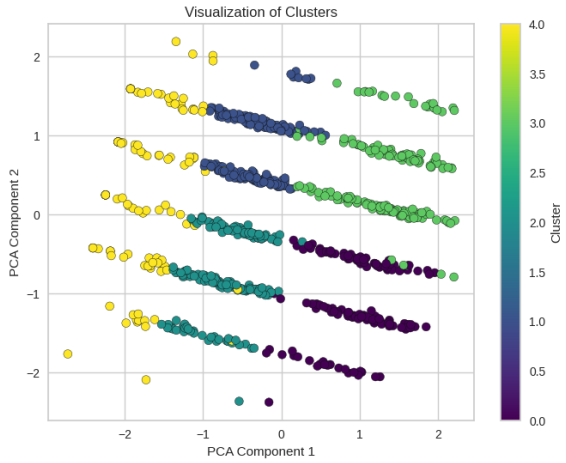


Hình 4.7: Agglomerative trước khi sử dụng FIM.

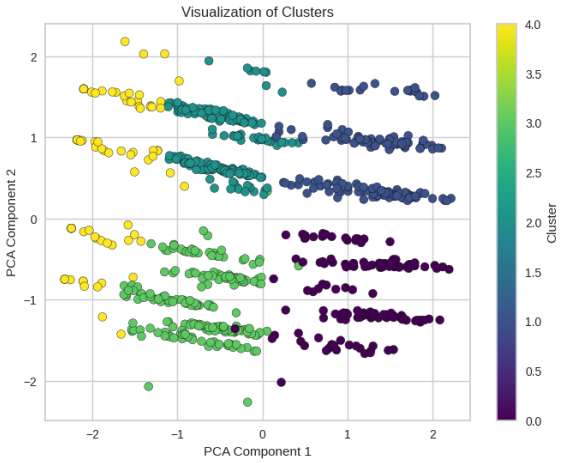


Hình 4.8: Agglomerative sau khi sử dụng FIM.

chặt chẽ giữa các cụm; Calinski-Harabasz tăng đột phá, phản ánh độ phân tách và cấu trúc cụm tối ưu hơn. Kết quả cho thấy các cụm được hình thành rõ nét, phân tách tốt và chất lượng cao hơn so với ban đầu.



Hình 4.9: Gaussian Mixture Model trước khi sử dụng FIM.

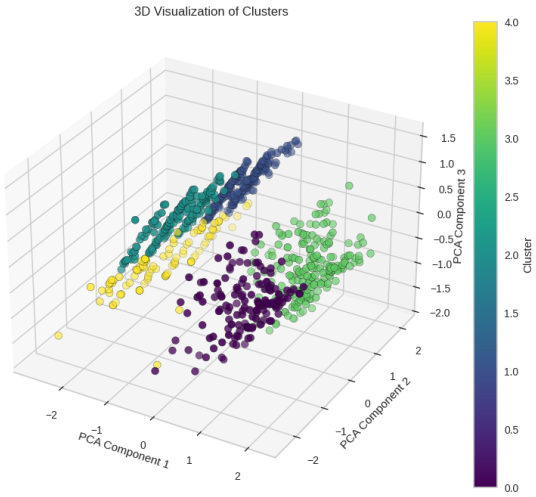


Hình 4.10: Gaussian Mixture Model sau khi sử dụng FIM.

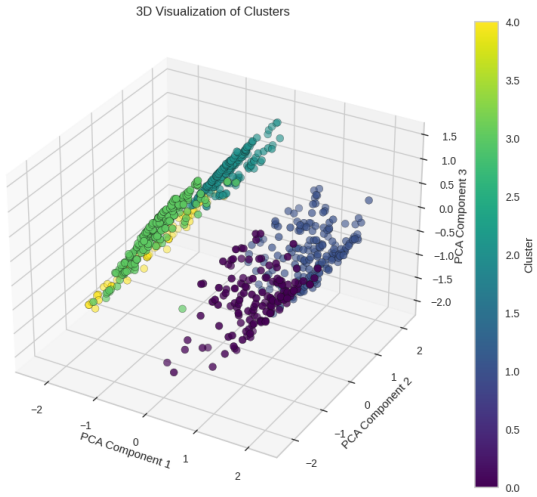
Bảng 4.3: Chỉ số đánh giá phân cụm Gaussian Mixture Model trước và sau kết hợp đặc trưng nhị phân

	Algorithms	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Before	GMM	0.325	1.249	378.246
After	GMM	0.341	1.234	373.190

4.1.3. Phân cụm Gaussian Mixture Model



Hình 4.11: Gaussian Mixture Model trước khi sử dụng FIM.



Hình 4.12: Gaussian Mixture Model sau khi sử dụng FIM.

Việc kết hợp đặc trưng nhị phân vào Gaussian Mixture Model (GMM) chỉ mang lại cải thiện nhỏ (**Bảng 4.3**): Silhouette tăng nhẹ, cho thấy sự tách biệt giữa các cụm cải thiện không đáng kể; Davies-Bouldin giảm nhẹ, phản ánh độ chặt chẽ giữa các cụm ít thay đổi; trong khi Calinski-Harabasz giảm, cho thấy độ phân tách giữa các cụm giảm nhẹ. Tóm lại, hiệu quả của việc bổ sung đặc trưng nhị phân là không đáng kể.

4.2. Phân tích từng cụm

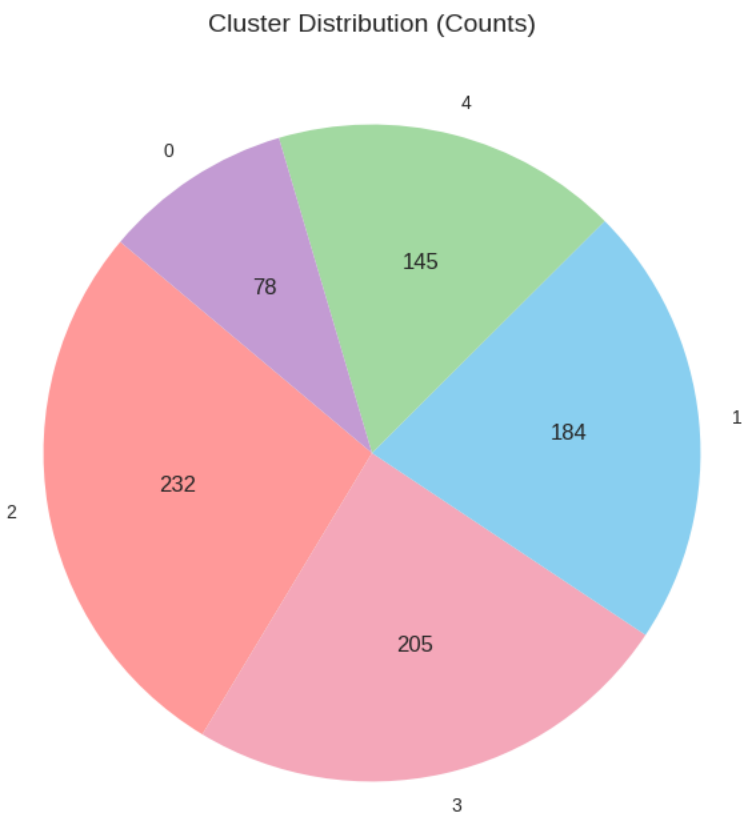
4.2.1. Phân phối từng cụm

Sau khi thực hiện phân cụm với 3 thuật toán, tôi lựa chọn K-Prototypes với 5 cụm để phân tích từng cụm. Thuật toán này tạo ra số lượng cụm lớn hơn so với hai thuật toán còn lại, giúp việc phân tích sự tương đồng giữa các cửa hàng trở nên chi tiết và sâu sắc hơn.

Dựa vào biểu đồ **Hình 4.13** cho thấy biểu đồ này có tổng cộng 5 cụm được đánh số từ 0 đến 4, với tỷ lệ phân bố như sau: Cụm 0 (màu tím nhạt) chiếm 78 cửa hàng. Cụm 1 (màu xanh nước) chiếm 184 cửa hàng. Cụm 2 (màu đỏ nhạt) chiếm 232 cửa hàng. Cụm 3 (màu hồng nhạt) chiếm tổng số cửa hàng. Cụm 4 (màu xanh lá) chiếm 145 cửa hàng.

4.2.2. Đặc điểm từng cụm

Cụm 1: Thời gian hoạt động khá lâu nhưng hiệu quả kinh doanh thấp. Doanh thu, lượng người theo dõi, và phản hồi tích cực đều ở mức rất thấp. Chất lượng dịch vụ cũng chưa được đánh giá cao.



Hình 4.13: Pie Chart thể hiện số lượng cửa hàng trong từng cụm

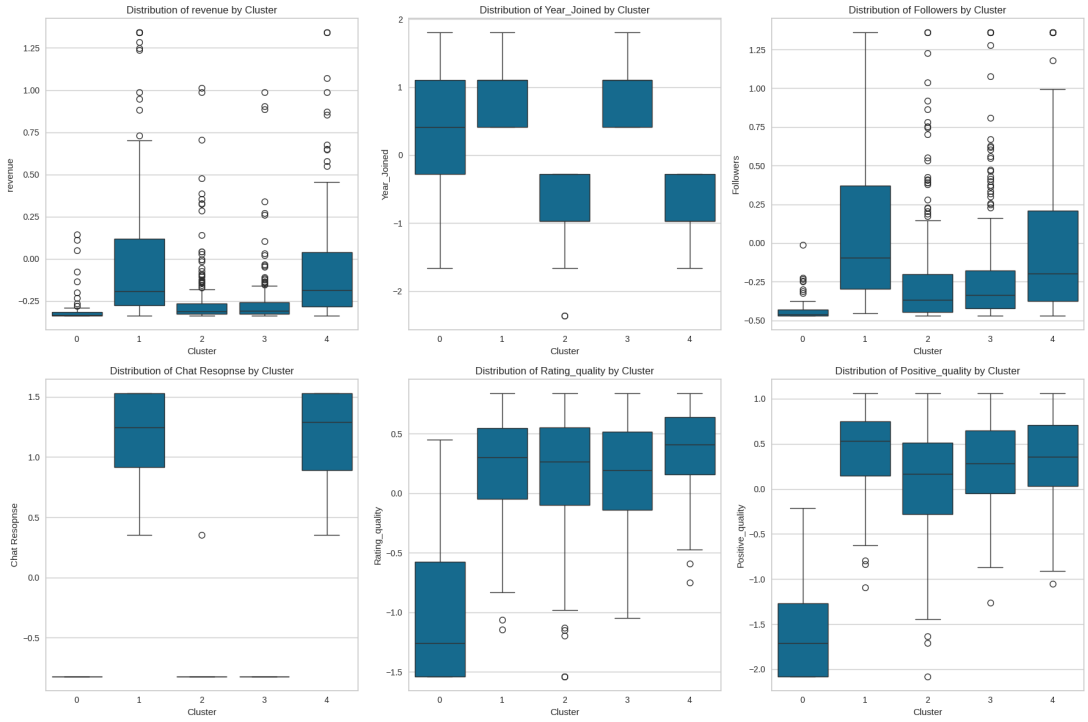
Cụm 2: Đây là các cửa hàng mới gia nhập thị trường, với hiệu quả kinh doanh và chất lượng dịch vụ tương đối tốt. Doanh thu và các chỉ số khác dao động lớn giữa các cửa hàng, cho thấy sự không đồng đều về hiệu quả hoạt động.

Cụm 3: Nhóm này chủ yếu là các cửa hàng lâu năm nhưng hoạt động không hiệu quả. Doanh thu và các chỉ số liên quan đều thấp, tuy nhiên một số cửa hàng vẫn giữ được điểm tích cực. Đây là nhóm cần cải thiện hiệu quả kinh doanh nhưng vẫn có tiềm năng để phát triển.

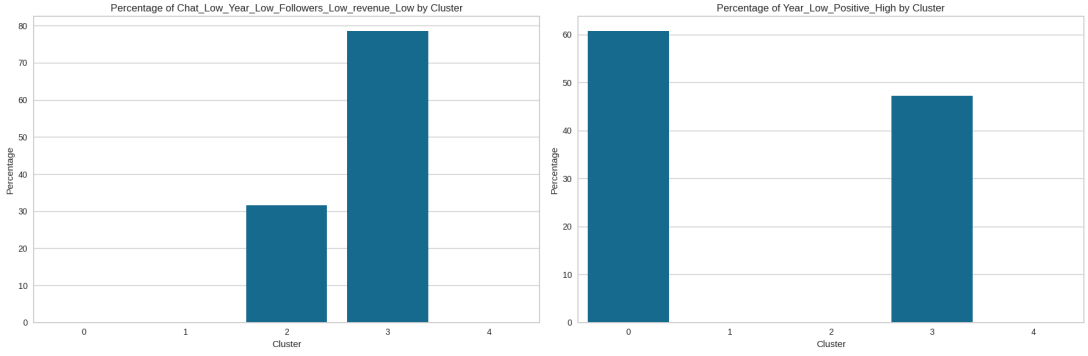
Cụm 4: Các cửa hàng trong cụm này tương đối mới và có hiệu quả trung bình. Điểm phản hồi tích cực và chất lượng dịch vụ không cao nhưng ổn định. Nhóm này có tiềm năng phát triển nếu cải thiện thêm chất lượng và trải nghiệm khách hàng.

Cụm 5: Đây là nhóm cửa hàng lâu năm hoạt động hiệu quả nhất. Doanh thu dao động lớn nhưng điểm chất lượng và phản hồi tích cực cao hơn các cụm khác. Nhóm này có thể được coi là các cửa hàng tiêu biểu về hiệu quả kinh doanh và dịch vụ.

4.2.3. Đề xuất chiến lược cho từng cụm



Hình 4.14: BoxPlot phân tích cụm các biến liên tục



Hình 4.15: BoxPlot phân tích cụm các biến nhị phân

Cụm 1: Cải thiện toàn diện về mọi khía cạnh

1. Nâng cao chất lượng sản phẩm và dịch vụ:
    - Thực hiện khảo sát để xác định các vấn đề cụ thể khiến khách hàng không hài lòng.
    - Đào tạo nhân viên để cải thiện kỹ năng phục vụ và chăm sóc khách hàng.
2. Chiến lược marketing và xây dựng thương hiệu:
    - Tăng cường quảng bá trên các nền tảng online và offline để thu hút khách hàng.
    - Thực hiện các chương trình khuyến mãi để kích thích doanh thu trong ngắn hạn.
3. Tối ưu vận hành và kiểm soát chi phí:
    - Rà soát lại các quy trình hoạt động để giảm chi phí và tăng hiệu quả.
    - Xây dựng các gói dịch vụ hoặc sản phẩm phù hợp với nhu cầu thị trường.

Cụm 2: Hỗ trợ phát triển bền vững và đồng đều

1. Hỗ trợ cửa hàng có hiệu quả thấp hơn:
    - Tổ chức các buổi chia sẻ kinh nghiệm từ các cửa hàng hoạt động tốt trong cụm.
    - Phân tích nguyên nhân sự chênh lệch và đưa ra giải pháp cải thiện phù hợp.
2. Tăng cường hiệu quả marketing:
    - Đẩy mạnh quảng bá các sản phẩm và dịch vụ nổi bật.
    - Sử dụng mạng xã hội để tăng khả năng tiếp cận và thu hút khách hàng mới.
3. Tăng tính cạnh tranh và chuyên nghiệp:
    - Thực hiện các chương trình khuyến mãi và ưu đãi để tăng khả năng cạnh tranh.
    - Xây dựng tiêu chuẩn dịch vụ đồng đều giữa các cửa hàng trong cụm.

Cụm 3: Khai thác tiềm năng và cải thiện hiệu quả kinh doanh

1. Phân tích và tập trung vào điểm mạnh:
    - Xác định các cửa hàng có tiềm năng phát triển và tập trung nguồn lực hỗ trợ.
    - Phát huy các yếu tố đang được khách hàng đánh giá tích cực.
2. Tái cấu trúc sản phẩm và dịch vụ:
    - Loại bỏ hoặc cải tiến các sản phẩm/dịch vụ không hiệu quả.
    - Thêm các sản phẩm/dịch vụ mới phù hợp với xu hướng thị trường.
3. Cải thiện hiệu quả vận hành:
    - Áp dụng công nghệ quản lý để giảm chi phí và tăng năng suất.
    - Thực hiện các chiến dịch marketing tập trung vào khách hàng trung thành.

Cụm 4: Phát triển tiềm năng và cải thiện chất lượng dịch vụ

1. Nâng cao trải nghiệm khách hàng:
    - Tập trung cải thiện chất lượng dịch vụ và trải nghiệm mua hàng.
    - Đào tạo nhân viên để tăng khả năng chăm sóc khách hàng.
2. Phát triển thương hiệu và marketing:
    - Xây dựng hình ảnh thương hiệu thông qua các chiến dịch truyền thông.
    - Thực hiện chương trình ưu đãi để thu hút khách hàng mới.
3. Đa dạng hóa sản phẩm và dịch vụ:
    - Tìm kiếm các sản phẩm/dịch vụ bổ trợ để tăng giá trị cho khách hàng.
    - Phân tích nhu cầu thị trường để tối ưu danh mục sản phẩm.

Cụm 5: Duy trì và mở rộng hiệu quả kinh doanh

1. Tăng cường duy trì và mở rộng thị phần:
    - Đẩy mạnh các chiến dịch quảng bá thương hiệu để thu hút thêm khách hàng mới.
    - Mở rộng phạm vi hoạt động hoặc chi nhánh để tận dụng lợi thế cạnh tranh.

2. Phát triển khách hàng trung thành:
    - Triển khai các chương trình khách hàng thân thiết để giữ chân khách hàng.
    - Tăng cường dịch vụ hậu mãi và chăm sóc khách hàng.
3. Đổi mới và nâng cao chất lượng:
    - Tiếp tục nâng cao chất lượng sản phẩm và dịch vụ để duy trì vị thế dẫn đầu.
    - Đầu tư vào công nghệ và cải tiến quy trình để tối ưu hiệu quả vận hành.

Chương 5. KẾT LUẬN VÀ KIẾN NGHỊ

Nghiên cứu này đã chứng minh rằng việc tích hợp khai thác tập phổ biến lớn nhất (FP-Max) vào quá trình phân cụm các cửa hàng trên nền tảng thương mại điện tử đã mang lại những cải thiện rõ rệt về chất lượng phân cụm. Bằng cách bổ sung các đặc trưng nhị phân vào tập dữ liệu, phương pháp đề xuất không chỉ làm giàu không gian đặc trưng mà còn giúp phát hiện các mối quan hệ ẩn sâu trong dữ liệu, vốn khó nhận thấy khi chỉ sử dụng các biến liên tục. Những cải tiến đáng kể trên các chỉ số đánh giá như Silhouette Score, Davies-Bouldin Index và Calinski-Harabasz Index đã khẳng định vai trò quan trọng của khai thác tập phổ biến trong việc nâng cao hiệu quả phân cụm. Các cụm kết quả đạt mức độ kết nối tốt hơn và giúp xác định rõ ràng các nhóm cửa hàng có đặc điểm kinh doanh và chất lượng dịch vụ tương đồng, từ đó cung cấp cơ sở phát triển các chiến lược kinh doanh phù hợp và hiệu quả.

Các nền tảng thương mại điện tử như Tiki có thể áp dụng phương pháp này để tối ưu hóa phân cụm cửa hàng, từ đó xác định chính xác các nhóm cửa hàng cần hỗ trợ hoặc thúc đẩy nhằm nâng cao chất lượng dịch vụ và tăng sự hài lòng của khách hàng. Đồng thời, phương pháp FP-Max cũng có thể được mở rộng bằng cách kết hợp với các kỹ thuật khai thác tập phổ biến có trọng số cao nhằm phát hiện những mẫu dữ liệu có giá trị lớn hơn, qua đó tối ưu hóa hiệu quả của các chiến lược phân cụm. Kết quả phân cụm này cũng cung cấp cho các cửa hàng cơ hội điều chỉnh chiến lược kinh doanh, cải thiện chất lượng sản phẩm và dịch vụ để gia tăng hiệu quả hoạt động và năng lực cạnh tranh. Ngoài ra, trong tương lai, việc áp dụng các kỹ thuật khai phá mẫu chiếm dụng trọng số cao hứa hẹn giúp các nền tảng thương mại điện tử phát hiện các mối quan hệ tiềm năng và đặc điểm quan trọng hơn, góp phần tối ưu hóa chiến lược kinh doanh và cải thiện chất lượng phân cụm trong bối cảnh dữ liệu lớn và phức tạp. Kết quả từ nghiên cứu không chỉ mang ý nghĩa khoa học mà còn có giá trị ứng dụng cao, đặc biệt trong bối cảnh phát triển nhanh chóng của thương mại điện tử tại Việt Nam và khu vực Đông Nam Á.

TÀI LIỆU THAM KHẢO

[1] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," en, *Briefings in Bioinformatics*, vol. 15, no. 5, pp. 788–797, Sep. 2014, ISSN: 1477-4054, 1467-5463. DOI: 10.1093/bib/bbt026. [Online]. Available: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt026 (visited on 12/16/2024).

[2] B. Mahesh, "Machine Learning Algorithms - A Review," en, *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, ISSN: 23197064. DOI: 10.21275/ART20203995. [Online]. Available: https://www.ijsr.net/archive/v9i1/ART20203995.pdf (visited on 12/16/2024).

[3] M. Usama, J. Qadir, A. Raza, *et al.*, "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," en, *IEEE Access*, vol. 7, pp. 65 579–65 615, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2916648. [Online]. Available: https://ieeexplore.ieee.org/document/8713992/ (visited on 12/16/2024).

[4] T. S. Madhulatha, *An Overview on Clustering Methods*, arXiv:1205.1117 [cs], May 2012. DOI: 10.48550/arXiv.1205.1117. [Online]. Available: http://arxiv.org/abs/1205.1117 (visited on 12/16/2024).

[5] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," en, *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, ISSN: 1099-4300. DOI: 10.3390/e23060759. [Online]. Available: https://www.mdpi.com/1099-4300/23/6/759 (visited on 12/15/2024).

[6] S. Petrovic, "A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters," en,

[7] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," en, *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 5, p. 052 024, Jul. 2019, ISSN: 1757-8981, 1757-899X. DOI: 10.1088/1757-899X/569/5/052024. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/569/5/052024 (visited on 12/15/2024).

[8] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," en, *J*, vol. 2, no. 2, pp. 226–235, Jun. 2019, ISSN: 2571-8800. DOI: 10.3390/j2020016. [Online]. Available: https://www.mdpi.com/2571-8800/2/2/16 (visited on 12/15/2024).

[9] A. Jeffares, *K-means: A Complete Introduction*, en, Nov. 2019. [Online]. Available: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c (visited on 12/16/2024).

[10] E. K. Tokuda, C. H. Comin, and L. D. F. Costa, "Revisiting agglomerative clustering," en, *Physica A: Statistical Mechanics and its Applications*, vol. 585, p. 126 433, Jan. 2022, ISSN: 03784371. DOI: 10.1016/j.physa.2021.126433. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378437121007068 (visited on 12/15/2024).

[11] A. Arora, *The Most Important Things You Need To Know About Agglomerative Clustering*, en-US, Jul. 2022. [Online]. Available: https://analyticsarora.com/the-most-important-things-you-need-to-know-about-agglomerative-clustering/ (visited on 12/16/2024).

[12] Z. Ban, J. Liu, and L. Cao, "Superpixel Segmentation Using Gaussian Mixture Model," en, *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4105–4117, Aug. 2018, ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2018.2836306. [Online]. Available: https://ieeexplore.ieee.org/document/8360143/ (visited on 12/15/2024).

[13] O. C. Carrasco, *Gaussian Mixture Models Explained*, en, Feb. 2020. [Online]. Available: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95 (visited on 12/16/2024).

[14] *Undertheseanlp/underthesea*, original-date: 2017-03-01T10:24:26Z, Dec. 2024. [Online]. Available: https://github.com/undertheseanlp/underthesea (visited on 12/16/2024).

[15] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning," en, *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 5, no. 1, pp. 401–413, Dec. 2023, Number: 1, ISSN: 2715-6079. DOI: 10.37385/jaets.v5i1.3080. [Online]. Available: https://www.journal.yrpipku.com/index.php/jaets/article/view/3080 (visited on 12/16/2024).

[16] T. Nyitrai and M. Virág, "The effects of handling outliers on the performance of bankruptcy prediction models," en, *Socio-Economic Planning Sciences*, vol. 67, pp. 34–42, Sep. 2019, ISSN: 00380121. DOI: 10.1016/j.seps.2018.08.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S003801211730232X (visited on 12/16/2024).

[17] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier factor," en, in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, Chongqing China: ACM, Sep. 2019, pp. 161–168, ISBN: 978-1-4503-6843-8. DOI: 10.1145/3338840.3355641. [Online]. Available: https://dl.acm.org/doi/10.1145/3338840.3355641 (visited on 12/16/2024).

[18] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Hefei, China: IEEE, Sep. 2020, pp. 949–953, ISBN: 978-1-7281-9627-5. DOI: 10.1109/IFEEA51475.2020.00199. [Online]. Available: https://ieeexplore.ieee.org/document/9356727/ (visited on 12/15/2024).

[19] B. Ziani and Y. Ouinten, "Mining maximal frequent itemsets: A java implementation of FPMAX algorithm," en, in *2009 International Conference on Innovations in Information Technology (IIT)*, Al-Ain, United Arab Emirates: IEEE, Dec. 2009, pp. 330–334, ISBN: 978-1-4244-5698-7. DOI: 10.1109/IIT.2009.5413790. [Online]. Available: http://ieeexplore.ieee.org/document/5413790/ (visited on 12/15/2024).

[20] G. Tuerhong and S. B. Kim, "Gower distance-based multivariate control charts for a mixture of continuous and categorical variables," en, *Expert Systems with Applications*, vol. 41, no. 4, pp. 1701–1707, Mar. 2014, ISSN: 09574174. DOI: 10.1016/j.eswa.2013.08.068. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417413006891 (visited on 12/16/2024).