HUTECH
Đại học Công nghệ Tp.HCM

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

ĐỒ ÁN CƠ SỞ

PHÂN CỤM CÁC CỬA HÀNG TRỰC TUYẾN 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 CƠ SỞ

PHÂN CỤM CÁC CỬA HÀNG TRỰC TUYẾN 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

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

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

TPHCM, Ngày. . . . . tháng. . . . . năm 2024

Giáo viên hướng dẫn

(Ký tên, đóng dấu)

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 "Phân cụm các cửa hàng trên sàn thương mại điện tử Tiki"và sẽ đóng góp một phần nhỏ vào lĩnh vực nghiên cứu này.

TPHCM, ngày 09 tháng 06 năm 2024

Sinh viên

Nguyễn Văn Đạt

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT VÀ TỪ KHÓA

K-Means	Thuật toán phân cụm dữ liệu phổ biến trong học không giám sát.
Elbow	Phương pháp lựa chọn số lượng cụm tối ưu trong K-Means dựa trên sự biến động của điểm.
Silhouette	Đánh giá chất lượng của việc phân cụm dữ liệu dựa trên độ tương đồng và độ khác biệt giữa các cụm.
Machine Learning	Lĩnh vực của trí tuệ nhân tạo mà máy tính có khả năng học hỏi từ dữ liệu mà không cần phải được lập trình một cách cụ thể.
Unsupervised Learning	Loại học máy mà mô hình học từ dữ liệu không có nhãn.
Clustering	Quá trình phân loại các mẫu dữ liệu thành các nhóm có tính chất tương đồng.
Underthesea	Thư viện xử lý ngôn ngữ tự nhiên cho tiếng Việt trong Python.
Log Transformation	Phương pháp biến đổi dữ liệu bằng cách lấy logarithm của chúng.
IQR(Interquartile Range)	Phạm vi tương phản giữa phân vị thứ 75 và phân vị thứ 25 của một tập dữ liệu, được sử dụng để phát hiện các giá trị ngoại lai.
StandardScaler	Một phương pháp chuẩn hóa dữ liệu bằng cách loại bỏ trung bình và chia tỷ lệ theo độ lệch chuẩn.
API Scraping	Kỹ thuật thu thập dữ liệu từ API.

Mục lục

Danh sách bảng

Danh sách hình vẽ

CHƯƠNG 1: TỔNG QUAN

1.1 Giới thiệu đề tài

Trong bối cảnh thương mại điện tử ngày càng phát triển, việc tối ưu hóa hiệu quả kinh doanh của các cửa hàng trực tuyến trở nên vô cùng quan trọng. Các doanh nghiệp không chỉ cần hiểu rõ về khách hàng mà còn phải nhận diện được các yếu tố ảnh hưởng đến thành công hoặc thất bại của các cửa hàng. Việc phân tích và phân cụm dữ liệu cửa hàng trực tuyến giúp xác định các nhóm cửa hàng có đặc điểm và hiệu suất kinh doanh tương đồng, từ đó đề xuất các chiến lược phát triển phù hợp. Đề tài "Phân Cụm Các Cửa Hàng Trực Tuyến Trên Sàn Thương Mại Điện Tử Tiki"được chọn nhằm mục tiêu cung cấp một cách tiếp cận khoa học và có hệ thống để hỗ trợ doanh nghiệp trong việc cải thiện hiệu quả hoạt động và tăng cường cạnh tranh trên thị trường.

1.2 Nhiệm vụ của đề tài

Nhiệm vụ của đề tài "Phân cụm các cửa hàng trực tuyến trên sàn thương mại điện tử"là quá trình áp dụng các kỹ thuật xử lý dữ liệu và các thuật toán Machine Learning nhằm tìm ra các nhóm chứa các cửa hàng tương đồng nhau. Từ đó, phân tích và tìm ra các đặc điểm chung của từng nhóm. Điều này giúp các doanh nghiệp hiểu biết sâu sắc về thị trường, qua đó có thể tối ưu hóa chiến lược marketing và bán hàng, cải thiện dịch vụ khách hàng và trải nghiệm người dùng, và tối ưu hóa hoạt động kinh doanh.

1.2.1 Tính cấp thiết của đề tài

Trong thời đại số hóa hiện nay, thương mại điện tử đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày và trong các hoạt động kinh doanh. Sự bùng nổ của các sàn thương mại điện tử như Tiki đã tạo ra một môi trường cạnh tranh khốc liệt, nơi hàng ngàn cửa hàng trực tuyến cùng tồn tại và phát triển. Điều này đặt ra nhu cầu cấp bách cho các doanh nghiệp phải liên tục đổi mới và tối ưu hóa các hoạt động kinh doanh của mình để tồn tại và phát triển.

Việc phân cụm các cửa hàng trực tuyến trên sàn thương mại điện tử là một bước đi quan

trọng trong việc tối ưu hóa hiệu quả kinh doanh. Phân cụm giúp các doanh nghiệp nhận diện và phân tích các nhóm cửa hàng có đặc điểm và hiệu suất kinh doanh tương đồng. Điều này không chỉ giúp tiết kiệm thời gian và nguồn lực trong việc xây dựng và triển khai các chiến lược kinh doanh, mà còn mang lại những hiểu biết sâu sắc và có giá trị về thị trường. Cụ thể hơn, việc phân cụm các cửa hàng trực tuyến có thể giúp các doanh nghiệp:

Tối ưu hóa chiến lược marketing và bán hàng: Bằng cách hiểu rõ đặc điểm của từng nhóm cửa hàng, doanh nghiệp có thể thiết kế các chiến lược marketing và bán hàng phù hợp hơn, nhằm tối ưu hóa hiệu quả và chi phí.

Cải thiện dịch vụ khách hàng và trải nghiệm người dùng: Phân tích các nhóm cửa hàng giúp doanh nghiệp nhận diện được nhu cầu và mong muốn của khách hàng, từ đó cải thiện dịch vụ và nâng cao trải nghiệm người dùng.

Tối ưu hóa hoạt động kinh doanh: Bằng cách tập trung vào các nhóm cửa hàng có hiệu suất cao, doanh nghiệp có thể tối ưu hóa các hoạt động kinh doanh, từ quản lý kho hàng đến phân phối sản phẩm.

Nâng cao khả năng cạnh tranh: Trong môi trường cạnh tranh khốc liệt, việc hiểu rõ và đáp ứng nhanh chóng các thay đổi của thị trường là yếu tố then chốt để duy trì và nâng cao vị thế cạnh tranh của doanh nghiệp.

Đưa ra các quyết định chiến lược dựa trên dữ liệu: Việc phân tích dữ liệu và phân cụm cửa hàng dựa trên các thuật toán Machine Learning giúp doanh nghiệp đưa ra các quyết định chiến lược dựa trên dữ liệu thay vì cảm tính, từ đó tăng cường độ chính xác và hiệu quả của các quyết định.

Từ những lý do trên, có thể thấy rằng việc phân cụm các cửa hàng trực tuyến trên sàn thương mại điện tử không chỉ mang lại lợi ích thiết thực cho doanh nghiệp mà còn đóng góp quan trọng vào việc phát triển thị trường thương mại điện tử bền vững và hiệu quả hơn.

1.2.2 Ý nghĩa khoa học và thực tiễn của đề tài

Ý nghĩa khoa học: đề tài đóng góp vào lĩnh vực phân tích dữ liệu bằng cách áp dụng thuật toán phân cụm tiên tiến như K-means và các phương phức, thuật toán xử lý dữ liệu để giải quyết vấn đề thực tiễn trong thương mại điện tử. Qua đó, đề tài không chỉ làm phong phú thêm lý thuyết về phân tích dữ liệu mà còn phát triển các phương pháp mới, tạo ra các mô hình phân tích cụ thể cho các cửa hàng trực tuyến. Những kết quả này giúp các nhà nghiên cứu và chuyên gia có thêm công cụ để nghiên cứu sâu hơn về hành vi người tiêu dùng và hiệu suất kinh doanh.

Ý nghĩa thực tiễn: đề tài hỗ trợ các doanh nghiệp trong việc ra quyết định kinh doanh chính xác và hiệu quả hơn. Kết quả phân cụm giúp doanh nghiệp hiểu rõ hơn về đặc điểm và hiệu suất của từng nhóm cửa hàng, từ đó đề ra các chiến lược quản lý và marketing phù hợp.

Điều này không chỉ nâng cao hiệu quả hoạt động mà còn tăng cường khả năng cạnh tranh của các cửa hàng trực tuyến trên thị trường. Bằng cách nhận diện các nhóm cửa hàng có hiệu suất tương đồng, doanh nghiệp có thể tối ưu hóa hoạt động và giảm thiểu rủi ro trong kinh doanh. Phương pháp và kết quả nghiên cứu từ đề tài còn có thể được ứng dụng rộng rãi trên các nền tảng thương mại điện tử khác, góp phần nâng cao chất lượng và hiệu quả của ngành. Hơn nữa, các nhà quản lý và hoạch định chính sách có thể sử dụng kết quả nghiên cứu này để thiết kế các chính sách hỗ trợ và phát triển ngành, đảm bảo một môi trường kinh doanh lành mạnh và bền vững. Bằng việc kết hợp giữa lý thuyết phân tích dữ liệu và ứng dụng thực tiễn, đề tài không chỉ mang lại giá trị khoa học mà còn đóng góp tích cực vào sự phát triển của ngành thương mại điện tử, tạo ra những cơ hội mới cho các doanh nghiệp và hỗ trợ sự phát triển bền vững của thị trường.

1.3 Mục tiêu

1.3.1 Mục tiêu tổng quát

Đề tài nhằm cung cấp một phương pháp phân tích và phân cụm các cửa hàng trực tuyến trên sàn thương mại điện tử Tiki, giúp nhận diện các nhóm cửa hàng có đặc điểm và hiệu suất kinh doanh tương đồng. Từ đó, đề xuất các chiến lược phát triển phù hợp để tối ưu hóa hiệu quả kinh doanh và nâng cao năng lực cạnh tranh cho các doanh nghiệp.

1.3.2 Mục tiêu cụ thể

Với bài nghiên cứu này, tôi sẽ sử dụng tập dữ liệu về các cửa hàng trực tuyến trên sàn thương mại điện tử bao gồm thông tin về sản phẩm, đánh giá của khách hàng, doanh thu, và thông tin tổng quát về cửa hàng. Quá trình thực hiện sẽ bao gồm các bước chính sau: xác định mục tiêu cụ thể, chuẩn bị dữ liệu, chọn phương pháp khai thác dữ liệu, áp dụng các mô hình, đánh giá hiệu quả của từng mô hình, triển khai mô hình. Cụ thể hơn, tôi sẽ áp dụng mô hình K-Means để phân cụm các cửa hàng tương đồng. Sau đó, các mô hình này sẽ được đánh giá bằng các kỹ thuật như Elbow, Silhouette. Cuối cùng, số cụm tốt nhất sẽ được tiến hành thực nghiệm. Kết quả mong đợi là phân cụm đúng đặc điểm của từng nhóm cửa hàng để đưa ra các đề xuất chiến lược kinh doanh phù hợp.

1.4 Đối tượng và phạm vi

1.4.1 Đối tượng

Đối tượng nghiên cứu của đề tài là các cửa hàng trực tuyến hoạt động trên sàn thương mại điện tử Tiki. Các cửa hàng này đa dạng về ngành hàng, kích thước và hiệu suất kinh doanh. Bằng cách thu thập và phân tích dữ liệu từ các cửa hàng này, đề tài nhằm hiểu rõ hơn về các yếu tố ảnh hưởng đến hiệu suất kinh doanh, từ đó đưa ra các đề xuất cải thiện và tối ưu hóa.

1.4.2 Phạm vi

Đề tài tập trung vào thu thập và phân tích dữ liệu từ các cửa hàng trực tuyến trên sàn thương mại điện tử Tiki. Chúng tôi sẽ áp dụng các phương pháp phân tích dữ liệu và thuật toán phân cụm để nhận diện các nhóm cửa hàng có đặc điểm và hiệu suất kinh doanh tương đồng. Dựa trên kết quả này, chúng tôi sẽ đề xuất các chiến lược kinh doanh phù hợp và đánh giá hiệu quả của chúng. Điều này giúp mang lại giá trị lý thuyết và thực tiễn cho các doanh nghiệp trên thị trường thương mại điện tử.

1.5 Phương pháp nghiên cứu

1.5.1 Phương pháp nghiên cứu sơ bộ

Trước khi tiến hành thu thập dữ liệu, chúng tôi sẽ tiến hành một nghiên cứu sơ bộ để hiểu rõ hơn về lĩnh vực nghiên cứu và các yếu tố quan trọng liên quan. Nghiên cứu này bao gồm việc tìm hiểu về 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. Thông qua việc nghiên cứu sơ bộ, chúng tôi sẽ xác định các vấn đề cụ thể cần giải quyết và đề xuất các phương pháp nghiên cứu phù hợp.

1.5.2 Phương pháp nghiên cứu tài liệu

Chúng tôi sẽ tiến hành nghiên cứu tài liệu để thu thập thông tin về các phương pháp và công cụ phân tích dữ liệu trong lĩnh vực thương mại điện tử và học máy. Qua việc đánh giá các nghiên cứu trước đây và các công trình khoa học liên quan, chúng tôi sẽ xác định các phương pháp phân cụm phù hợp nhất cho nghiên cứu của mình và áp dụng chúng vào việc phân tích dữ liệu.

1.5.3 Phương pháp nghiên cứu thống kê

Trong quá trình phân tích dữ liệu, chúng tôi sẽ sử dụng các phương pháp thống kê để mô tả và phân tích các biến số quan trọng. Các phương pháp thống kê bao gồm phân tích đơn biến, đa biến, phân tích phương sai, và kiểm tra độ tương quan của các biến. Thông qua việc áp dụng các phương pháp thống kê này, chúng tôi sẽ đánh giá mối quan hệ giữa các biến số và xác định các yếu tố ảnh hưởng đến hiệu suất kinh doanh của các cửa hàng trực tuyến.

1.5.4 Phương pháp thực nghiệm

Chúng tôi sẽ tiến hành thực nghiệm trên dữ liệu thu thập được từ sàn thương mại điện tử Tiki. Quá trình này bao gồm việc tiền xử lý dữ liệu, áp dụng các phương pháp phân cụm để nhận diện các nhóm cửa hàng, và đánh giá hiệu quả của các chiến lược kinh doanh đề xuất. Thông qua việc thực nghiệm trên thực tế, chúng tôi sẽ kiểm tra và đảm bảo tính khả thi và hiệu quả của phương pháp nghiên cứu.

1.5.5 Phương pháp đánh giá

Cuối cùng, chúng tôi sẽ thực hiện phương pháp đánh giá để đo lường hiệu quả của các phương pháp phân tích dữ liệu. Quá trình này bao gồm việc so sánh các chỉ số và thước đo hiệu quả kinh doanh giữa các nhóm cửa hàng.

1.6 Những đóng góp nghiên cứu của đề tài

1.6.1 Trong lĩnh vực học thuật

Đề tài đóng góp bằng cách áp dụng và phát triển các phương pháp phân tích dữ liệu tiên tiến trong ngữ cảnh thương mại điện tử. Việc áp dụng thuật toán phân cụm như K-means vào việc phân tích các cửa hàng trực tuyến trên Tiki mở ra một lĩnh vực nghiên cứu mới, đồng thời cung cấp các mô hình phân cụm cụ thể cho việc nghiên cứu tiếp theo. Ngoài ra, đề tài cũng đóng góp vào việc tạo ra một cơ sở dữ liệu lớn và đa dạng về các cửa hàng trực tuyến, từ đó cung cấp nguồn tài nguyên quý giá cho các nghiên cứu liên quan về thương mại điện tử và học máy.

1.6.2 Trong thực tiễn kinh doanh

Đề tài cung cấp các chiến lược cụ thể và có tính ứng dụng cao cho các doanh nghiệp hoạt động trong lĩnh vực thương mại điện tử. Việc nhận diện các nhóm cửa hàng có đặc điểm và hiệu suất tương đồng giúp doanh nghiệp hiểu rõ hơn về thị trường và khách hàng

của mình. Đồng thời, việc đề xuất các chiến lược kinh doanh phù hợp cho từng nhóm cửa hàng giúp tối ưu hóa hoạt động kinh doanh và tăng cường cạnh tranh trên thị trường. Điều này có thể giúp các doanh nghiệp cải thiện hiệu suất kinh doanh và tăng trưởng doanh thu trong một thị trường thương mại điện tử ngày càng cạnh tranh.

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 [8] 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 [20]. 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 tốt: API cung cấp dữ liệu theo định dạng có cấu trúc, thường là JSON hoặc XML, giúp dễ dàng phân tích và xử lý.
- Độ chính xác cao: API được cung cấp bởi chính trang web hoặc dịch vụ, đảm bảo dữ liệu chính xác và cập nhật.
- Nhanh chóng: Việc truy vấn API thường nhanh hơn so với việc phải tải và phân tích HTML của trang web.
- Tối ưu hóa hiệu suất: API thường được thiết kế để xử lý truy vấn một cách hiệu quả, giảm thiểu tải mạng và thời gian xử lý.

– Thay đổi dễ dàng: Khi có thay đổi trên trang web, các API thường được cập nhật và duy trì ổn định, giúp giảm thiểu các lỗi phát sinh do thay đổi cấu trúc HTML.

Hạn chế:

– Hạn chế truy cập: Nhiều API có giới hạn về số lượng truy vấn mà bạn có thể thực hiện trong một khoảng thời gian nhất định, điều này có thể ảnh hưởng đến việc thu thập dữ liệu lớn.

– Yêu cầu xác thực: Một số API yêu cầu xác thực nghiêm ngặt, có thể phức tạp đối với người dùng mới hoặc người dùng không có quyền truy cập hợp lệ.

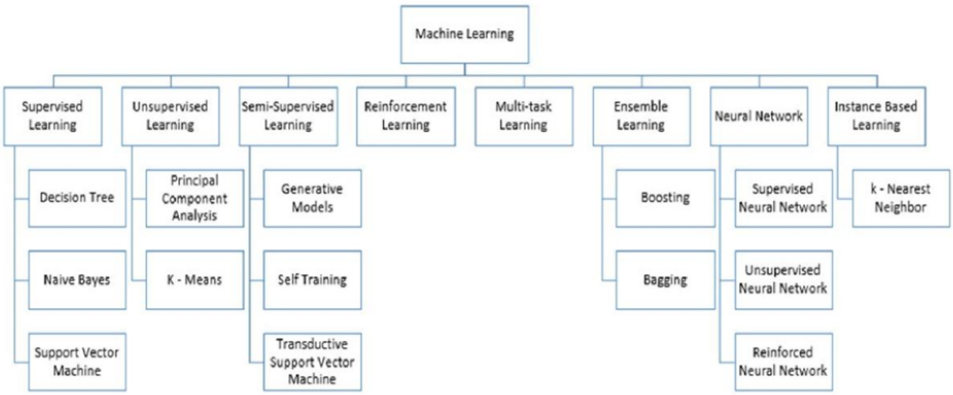
– Phụ thuộc vào nhà cung cấp: Nhà cung cấp API có thể thay đổi hoặc ngừng cung cấp dịch vụ mà không báo trước, gây ảnh hưởng đến ứng dụng sử dụng API.

– Chi phí: Một số API yêu cầu phí sử dụng, đặc biệt khi cần truy cập số lượng lớn dữ liệu hoặc các tính năng cao cấp.

– Giới hạn loại dữ liệu: API thường chỉ cung cấp một phần dữ liệu mà trang web hiển thị, không phải tất cả các thông tin có trên trang web.

2.2 Machine Learning

Máy học, còn được gọi là học máy (machine learning) [10], 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 [10]. Thông qua các thuật toán máy học, hệ thống xây dựng mô hình dựa trên dữ liệu mẫu hay còn gọi là dữ liệu huấn luyện, từ đó đưa ra các dự đoán hoặc quyết định mà không cần phải lập trình chi tiết về cách thức đưa ra những dự đoán hoặc quyết định này. Ở đề tài này, tôi tập chung vào sử dụng các thuật toán ở nhánh Unsupervised Learning trong Hình 2.1 để xử lý bài toán phân cụm.

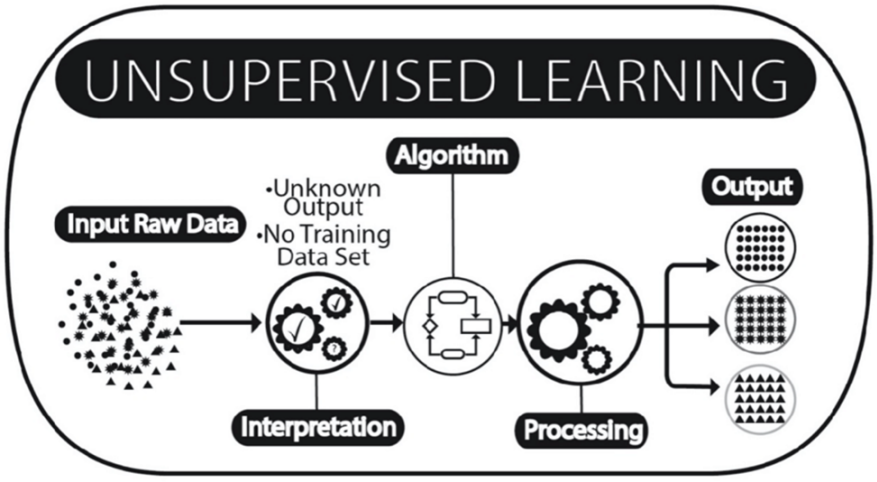


Hình 2.1: Tổng quan về Machine Learning. Nguồn: [4].

2.2.1 Unsupervised Learning

Học không có giám sát (tiếng Anh: unsupervised learning) là một phương pháp của ngành học máy nhằm tìm ra một mô hình mà phù hợp với các quan sát [9]. Nó khác biệt với học có giám sát ở chỗ là đầu ra đúng tương ứng cho mỗi đầu vào là không biết trước.

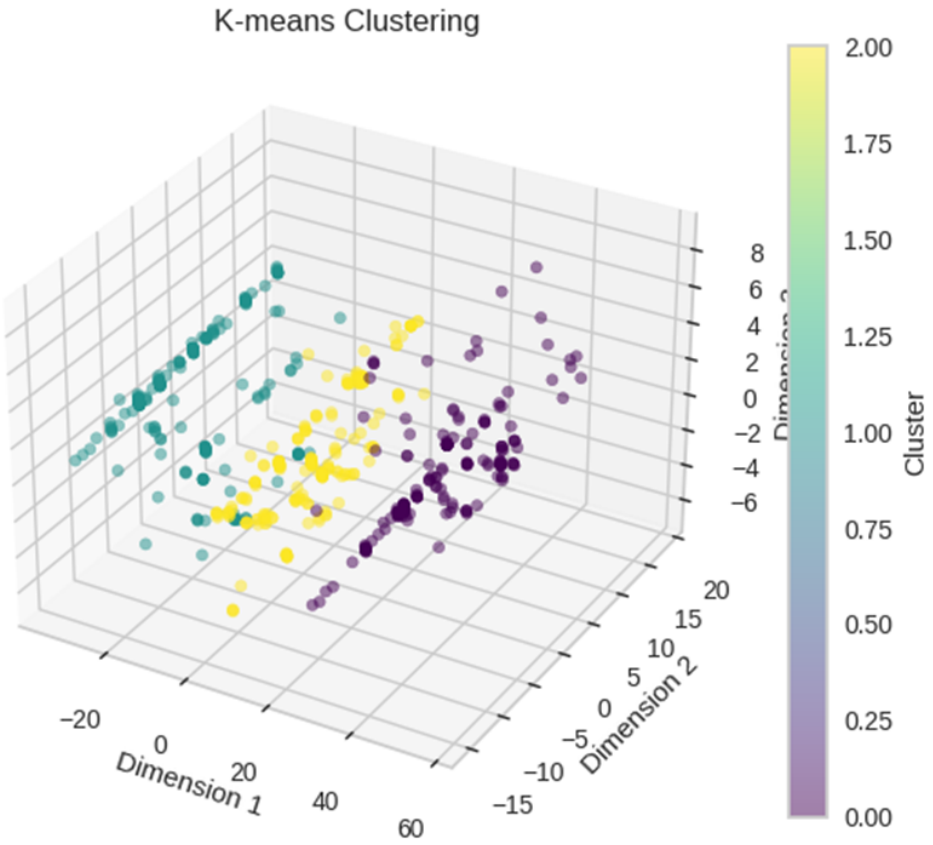
Các thuật toán học không giám sát học một số tính năng từ dữ liệu. Khi dữ liệu mới được đưa vào, nó sẽ sử dụng các tính năng đã học trước đó để nhận dạng lớp dữ liệu [10]. Nó chủ yếu được sử dụng để phân cụm (Clustering), phát hiện bất thường (Anomaly Detection), giảm chiều dữ liệu (Dimensionality Reduction) và xây dựng mô hình generative (Generative Modeling). Hình 2.2 dưới đây mô tả tổng quan về học không giám sát.



Hình 2.2: Minh họa về học không giám sát. Nguồn: [12]

2.2.2 *Clustering*

Phân cụm (Clustering) là một kỹ thuật phân tích dữ liệu không giám sát 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 [14].

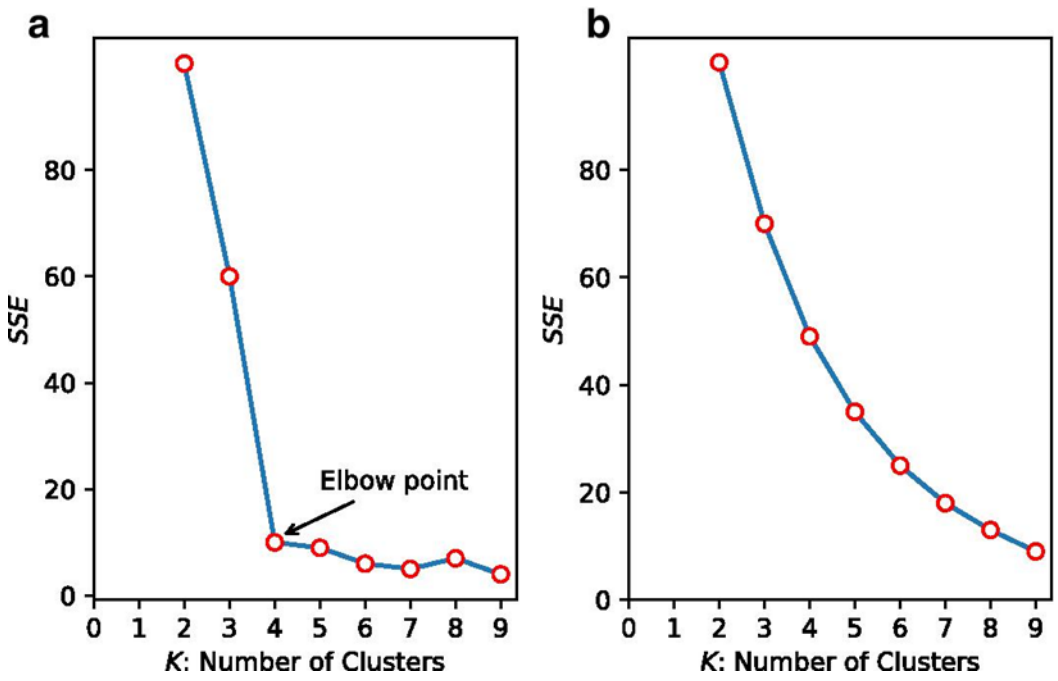


Hình 2.3: Minh họa thuật toán phân cụm

2.3 Elbow

2.3.1 *Giới thiệu về phương pháp Elbow*

Phương pháp Elbow [15] là một kỹ thuật được sử dụng để chọn số lượng nhóm cụm tối ưu trong thuật toán phân cụm như K-Means. Mục tiêu của phương pháp này là tìm ra một giá trị của số lượng nhóm cụm sao cho việc thêm một nhóm cụm mới không cải thiện đáng kể việc giảm tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm của nhóm cụm [19]. Chỉ số Elbow point như Hình 2.4 được xem là chỉ số tối ưu để chọn cụm.



Hình 2.4: Minh họa phương pháp Elbow. Nguồn:[17]]

2.3.2 Nền tảng toán học

Phương pháp Elbow dựa trên việc tính toán giá trị của một thước đo hoặc một hàm mất mát cho mỗi số lượng cụm khác nhau. Thường là Sum of Squared Errors (SSE) được sử dụng, là tổng của bình phương khoảng cách giữa mỗi điểm dữ liệu và trọng tâm của cụm gần nhất. Elbow method giả định rằng khi số lượng cụm tăng lên, SSE sẽ giảm dần và sẽ có một điểm khiến cho việc tăng thêm cụm không còn mang lại lợi ích lớn nữa.

Cho một tập dữ liệu $X = \{x_1, x_2, \ldots, x_n\}$ với n điểm dữ liệu và một số lượng cụm k, công thức toán học để tính SSE cho mỗi số lượng cụm là:

$$SSE(k) = \sum_{i=1}^{n} \min_{i=1}^{k} ||x_i - \mu_j||^2 \tag{2.1}$$

Trong đó:

- SSE(k): là Sum of Squared Errors cho số lượng cụm k.
- x_i là điểm dữ liệu thứ i.
- μ_j là trọng tâm của cụm thứ j
- $|||x_i - \mu_j||^2$ là bình phương khoảng cách của điểm dữ liệu x_i và trọng tâm của cụm μ_j

2.3.3 Diễn giải thuật toán

Giai đoạn 1:

- Đầu tiên, chúng ta xây dựng một chuỗi các mô hình phân cụm với số lượng cụm từ 1 đến một giới hạn nào đó.
- Tiếp theo, chúng ta tính toán SSE cho mỗi mô hình.

Giai đoạn 2:

- Đối với mỗi tập dữ liệu minh họa, chúng ta biểu diễn SSE theo số lượng cụm trên biểu đồ.
- Chúng ta quan sát biểu đồ để xác định điểm "elbow", nơi mà sự giảm của SSE giảm dần đi đáng kể.Phương pháp Elbow cung cấp một phương tiện đơn giản và trực quan để chọn số lượng nhóm cụm tối ưu mà không cần kiến thức trước về dữ liệu.

2.3.4 Phân tích độ phức tạp

Phương pháp Elbow có độ phức tạp tính toán thấp, vì chỉ yêu cầu tính toán SSE cho mỗi số lượng cụm. Tuy nhiên, độ phức tạp về thời gian có thể tăng lên đối với các tập dữ liệu lớn.

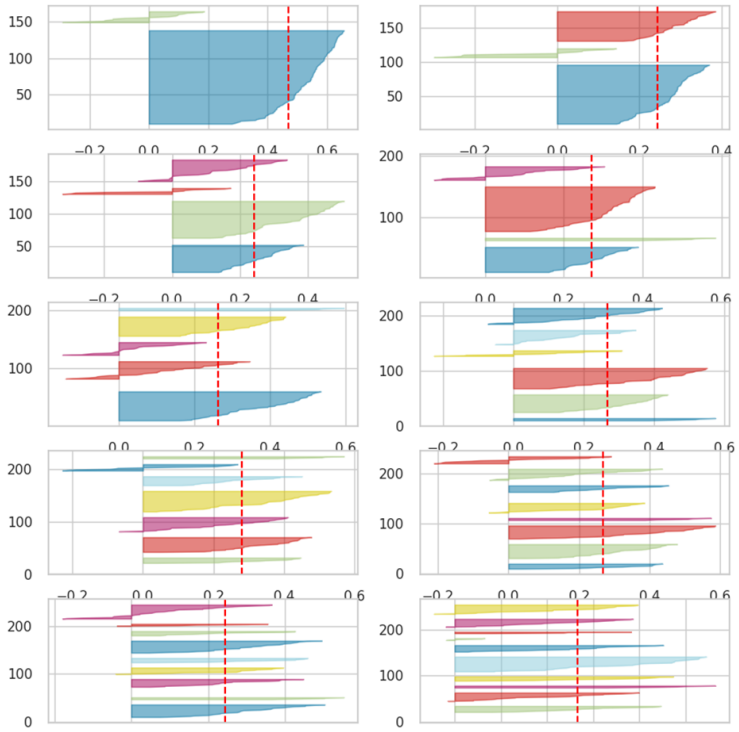
2.3.5 Ưu điểm và hạn chế

- Ưu điểm: dễ dàng triển khai và hiểu, cung cấp một cách trực quan để chọn số lượng cụm tối ưu.
- Hạn chế: Không phải lúc nào cũng cho ra kết quả chính xác, đặc biệt là đối với dữ liệu có cấu trúc phức tạp.

2.4 Silhouette

2.4.1 Giới thiệu về phương pháp Silhouette

Chỉ số Silhouette [18] là một phương pháp không giám sát để đánh giá hiệu suất của phương pháp 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.



Hình 2.5: Minh họa phương pháp Silhouette

2.4.2 Nền tảng toán học

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2.2}$$

Trong đó:

- $S(i)$ là Silhouette score của điểm dữ liệu
- $a(i)$ là trung bình khoảng cách giữa điểm dữ liệu i và các điểm khác trong cùng một cụm.
- $b(i)$ trung bình khoảng cách giữa điểm dữ liệu i và các điểm trong cụm gần nhất khác.

Silhouette score cho mỗi điểm dữ liệu nằm trong khoảng [-1, 1], trong đó:

- 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.

2.4.3 Diễn giải thuật toán

- Giai Đoạn 1: Đầu tiên, chúng ta phân cụm dữ liệu bằng một thuật toán phân cụm nào đó, chẳng hạn như K-Means. Tiếp theo, chúng ta tính toán Silhouette score cho mỗi điểm dữ liệu trong từng cụm.
- Giai Đoạn 2: Chúng ta tính toán Silhouette score trung bình cho tất cả các điểm dữ liệu trong tập dữ liệu.

2.4.4 Phân tích độ phức tạp

Phương pháp Silhouette yêu cầu tính toán khoảng cách giữa mỗi cặp điểm dữ liệu, điều này có thể tốn nhiều thời gian và tài nguyên tính toán đối với các tập dữ liệu lớn. Tuy nhiên, độ phức tạp thời gian và không gian của phương pháp Silhouette là tuyến tính, vì vậy nó thích hợp cho các tập dữ liệu có kích thước lớn.

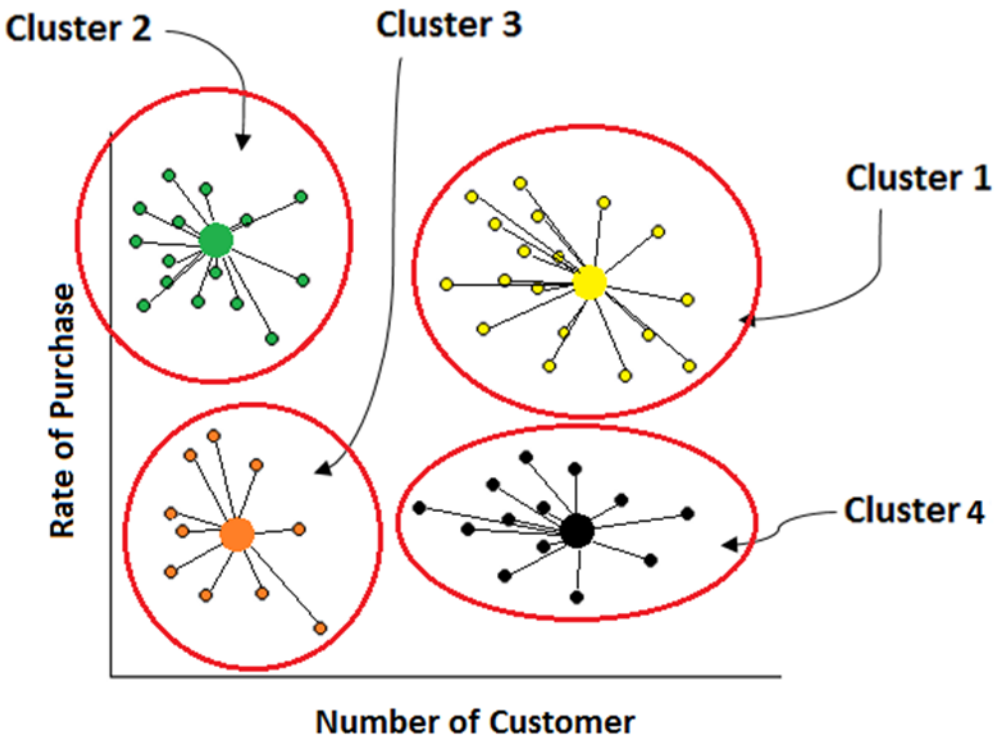
2.4.5 Ưu điểm và hạn chế

- 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.

2.5 K-Means

2.5.1 Giới thiều về thuật toán K-Means

K-means [16] là một trong những thuật toán học không giám sát đơn giản nhất giúp giải quyết vấn đề phân cụm phổ biến [10]. Thuật toán k-means phù hợp nhất cho việc khai thác dữ liệu vì tính hiệu quả của nó trong việc xử lý các tập dữ liệu lớn. Phân cụm là một trong những kỹ thuật khai thác dữ liệu nổi tiếng để tìm mẫu hữu ích từ dữ liệu trong cơ sở dữ liệu lớn [2].



Hình 2.6: Minh họa về thuật toán K-Means. Nguồn: [5]

2.5.2 Nền tảng toán học

Bước 1: Tạo các trung tâm ngẫu nhiên

$$c^{(0)} = (m_1^{(0)}, m_2^{(0)}, \ldots, m_k^{(0)}) \tag{2.3}$$

Bước 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.

$$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2\}, \forall j, 1 \leq j \leq k \tag{2.4}$$

Trong đó:

- $S_i^{(t)}$: Tập hợp các điểm dữ liệu được gán vào cụm i tại bước thứ t.
- x_p: Một điểm dữ liệu trong tập dữ liệu
- $m_i^{(t)}$: Trung tâm của cụm i 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.

Bước 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.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x \in S_i^{(t)}} x_j \tag{2.5}$$

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 số lượng 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 thuận 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.

Thuật toán sẽ lặp lại các bước trên cho tới khi đạt được kết quả chấp nhận được.

2.5.3 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.

```
Giả mã (pseudo-code)

initialize centroids randomly

while not converged:

  for each point in dataset:

    assign point to the nearest centroid

  for each centroid:

    update centroid to be the mean of points assigned to it
```

Hình 2.7: Hình minh họa về mã giả thuật toán K-Means.

2.5.4 Phân tích độ phức tạp

Độ phức tạp của thuật toán K-Means phụ thuộc vào số lượng điểm dữ liệu n, số lượng cụm k và số lần lặp lại của quá trình gán và cập nhật trọng tâm. Độ phức tạp thời gian trung bình của K-Means là $O(n \cdot k \cdot I \cdot d)$ trong đó I là số lần lặp lại và d là số chiều của không gian dữ liệu.

2.5.5 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.

2.5.6 Ứng dụng:

- Trong marketing, K-Means được sử dụng để phân đoạn khách hàng dựa trên hành vi mua sắm và đặc điểm nhân khẩu học.
- K-Means có thể được sử dụng để giảm số lượng màu trong hình ảnh, làm giảm kích thước tệp mà không làm giảm chất lượng hình ảnh quá nhiều.
- K-Means giúp 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.

2.6 Underthesea

Underthesea là một thư viện xử lý ngôn ngữ tự nhiên (NLP) dành cho tiếng Việt trong Python[3]. Thư viện này cung cấp các công cụ và chức năng để thực hiện các tác vụ như tách từ (tokenization), tách câu (sentence segmentation), phân loại từ loại (part-of-speech tagging), và phân tích cú pháp (parsing) và phân loại cảm xúc (sentiment), ...

Hình 2.8: Hình minh họa về thư viện underthesea. Nguồn:[21].

Dưới đây là một số chức năng chính của thư viện Underthesea:

– Tách từ (Tokenization): Chia văn bản thành các đơn vị từ riêng lẻ như từ, số hoặc dấu câu.

– Tách câu (Sentence Segmentation): Phân chia văn bản thành các câu riêng lẻ.

– Phân loại từ loại (Part-of-Speech Tagging): Xác định loại từ (động từ, danh từ, tính từ,...) của mỗi từ trong văn bản.

– Phân tích cú pháp (Parsing): Phân tích cấu trúc câu để hiểu ý nghĩa của câu.

– Phân loại cảm xúc (Sentiment): Phân tích cảm xúc của câu thuộc về tích cực hay tiêu cực.

Underthesea được phát triển để hỗ trợ xử lý ngôn ngữ tự nhiên tiếng Việt, giúp cho các nhà phát triển và nghiên cứu dễ dàng thực hiện các nhiệm vụ liên quan đến ngôn ngữ trong môi trường Python.

2.7 Log Transformation

2.7.1 Giới thiệu về phương pháp Log transformation

Log Transformation là một phương pháp được sử dụng rộng rãi để giải quyết dữ liệu sai lệch [7]. Phương pháp này được sử dụng để biến đổi các dữ liệu không đối xứng hoặc không

tuân theo phân phối chuẩn thành dữ liệu có phân phối gần như chuẩn hóa. Phương pháp này đặc biệt hữu ích khi làm việc với dữ liệu có phân phối đuôi dài hoặc biến đổi không tuyến tính.

2.7.2 Nền tảng toán học

Log Transformation biến đổi dữ liệu theo công thức logarithmic. Đối với dữ liệu dương và không bằng 0, logarit tự nhiên (cơ số e) thường được sử dụng, được biểu diễn như sau:

$$y = \log(x) \tag{2.6}$$

Trong đó:

- y là giá trị sau khi biến đổi.
- x là giá trị gốc của dữ liệu.

Nó cũng có thể được biến đổi với các cơ số khác như logarit cơ số 10 hoặc logarit cơ số 2 tùy thuộc vào nhu cầu của bài toán.

2.7.3 Lý do sử dụng

- Log Transformation thường được sử dụng để chuẩn hóa phân phối của dữ liệu, đặc biệt là trong các trường hợp mà dữ liệu không tuân theo phân phối chuẩn.
- Đối với dữ liệu có độ biến thiên không đồng nhất hoặc có đuôi dài, Log Transformation có thể giúp giảm độ biến thiên và làm cho dữ liệu dễ dàng hơn trong việc xử lý và phân tích.
- Log Transformation có thể tăng sự tương quan giữa các biến, đặc biệt là trong các tình huống mà các biến có sự tương quan đồng biến hoặc phân phối lệch.

2.7.4 Ưu điểm và hạn chế

Ưu điểm:

- Chuẩn hóa phân phối dữ liệu.
- Giảm độ biến thiên của dữ liệu.
- Tăng sự tương quan giữa các biến.

Hạn chế:

– Không thể áp dụng cho các giá trị bằng 0 hoặc âm.

– Có thể làm mất mát thông tin nếu không được sử dụng đúng cách.

– Cần lưu ý về các biến chứa giá trị gần 0, vì Log Transformation có thể tạo ra giá trị vô cùng nhỏ.

2.7.5 Ứng dụng

– Log Transformation thường được sử dụng trong các nghiên cứu y tế để chuẩn hóa phân phối của các biến như huyết áp, cholesterol, và các chỉ số sinh hóa.

– Trong lĩnh vực tài chính, Log Transformation có thể được áp dụng để chuẩn hóa các biến như lợi nhuận, tỷ lệ sinh lợi suất và biến động giá.

– Trong lĩnh vực xử lý ảnh, Log Transformation có thể được sử dụng để cải thiện độ tương phản và giảm độ sáng của hình ảnh.

2.8 IQR

2.8.1 Giới thiệu về phương pháp IQR

Trong thống kê mô tả , phạm vi liên tứ phân vị (IQR) [6] là thước đo độ phân tán thống kê , là mức độ lan truyền của dữ liệu. IQR còn có thể được gọi là mức chênh lệch giữa , mức chênh lệch giữa 50% , mức chênh lệch thứ tư hoặc mức chênh lệch H. Nó được định nghĩa là sự khác biệt giữa phần trăm thứ 75% và 25% của dữ liệu [11].

2.8.2 Nền tảng toán học

IQR được tính toán bằng cách lấy hiệu của phần tư thứ ba (Q3) và phần tư thứ nhất (Q1) của tập dữ liệu:

$$IQR = Q3 - Q1 \tag{2.7}$$

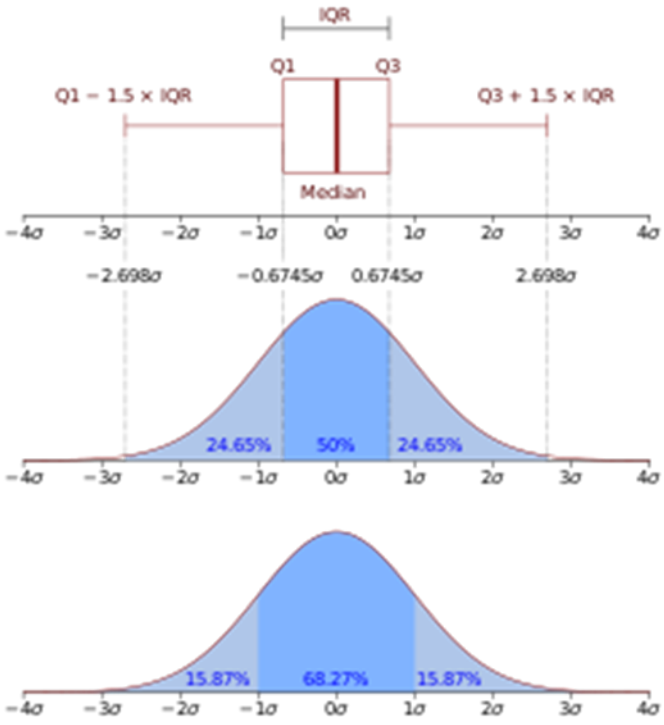
Trong đó:

– IQR: Interquartile Range.

– $Q3$: phần tư thứ ba (75th percentile) của dữ liệu.

– $Q1$: phần tư thứ nhất (25th percentile) của dữ liệu.

IQR cho biết phạm vi giữa các giá trị dữ liệu mà 50% số lượng quan sát nằm trong đó. Nó thường được sử dụng để xác định các giá trị ngoại lai (outliers) trong tập dữ liệu.



Hình 2.9: Minh họa về phương pháp IQR. Nguồn: [13].

2.8.3 Lý do sử dụng

– IQR thường được sử dụng để phát hiện và loại bỏ các giá trị ngoại lai từ tập dữ liệu. Các giá trị ngoại lai thường là các quan sát có giá trị rất cao hoặc thấp so với phân phối chung của dữ liệu.

– IQR cung cấp thông tin về biến động của dữ liệu trong một phạm vi cụ thể. Khi IQR lớn, nghĩa là dữ liệu có sự biến động lớn, và ngược lại.

– IQR cũng có thể được sử dụng để xác định phân phối của dữ liệu, đặc biệt là khi kết hợp với các biểu đồ như biểu đồ hộp (box plot) để hiểu rõ hơn về phân phối của tập dữ liệu.

2.8.4 Ưu điểm và hạn chế

Ưu điểm:

– Dễ dàng hiểu và tính toán.

– Không bị ảnh hưởng bởi giá trị cực đại hoặc cực tiểu trong tập dữ liệu.

– Cung cấp một phương tiện đơn giản để phát hiện giá trị ngoại lai.

Hạn chế:

- Không cung cấp thông tin chi tiết về phân phối dữ liệu như mean và standard deviation.
- Không phản ánh được sự biến động của dữ liệu ở phần đuôi của phân phối.

2.8.5 Ứng dụng

- IQR thường được sử dụng để phát hiện và loại bỏ các giá trị ngoại lai trong dữ liệu y tế như huyết áp, cân nặng, và chỉ số sinh học.
- Trong lĩnh vực tài chính và kinh doanh, IQR có thể được sử dụng để phân tích biến động của giá cổ phiếu, thu nhập, và các chỉ số tài chính khác.
- Trong xử lý ngôn ngữ tự nhiên, IQR có thể được sử dụng để phân tích độ biến động của độ dài của các văn bản, từ đó giúp hiểu rõ hơn về cấu trúc của văn bản.

2.9 Standard Scaler

2.9.1 Giới thiệu về StardardScaler

StandardScaler [1] là một phương pháp chuẩn hóa dữ liệu phổ biến được sử dụng trong quá trình tiền xử lý dữ liệu trong học máy và khai phá dữ liệu. Phương pháp này nhằm mục đích biến đổi các biến số của dữ liệu sao cho chúng có phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1.

2.9.2 Nền tảng toán học

StandardScaler biến đổi dữ liệu theo phương trình:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{2.8}$$

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.

2.9.3 Lý do sử dụng

- StandardScaler thường được sử dụng trong quá trình tiền xử lý dữ liệu trước khi huấn luyện mô hình học máy. Việc chuẩn hóa dữ liệu giúp cải thiện hiệu suất của các mô hình và làm cho quá trình huấn luyện mô hình ổn định hơn.
- Trong phân tích thống kê, việc chuẩn hóa dữ liệu giúp so sánh các biến số có đơn vị đo khác nhau và có phân phối khác nhau. Điều này giúp dễ dàng hơn trong việc so sánh các mối quan hệ và hiệu ứng của các biến số trên các kết quả.

2.9.4 Ưu điểm và hạn chế

Ưu điểm:

- Chuẩn hóa dữ liệu giúp cải thiện hiệu suất của các mô hình học máy.
- Giúp loại bỏ ảnh hưởng của các biến số không cần thiết trong quá trình huấn luyện mô hình.
- Giúp cải thiện ổn định của các thuật toán học máy.

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.
- Chuẩn hóa dữ liệu có thể làm cho dữ liệu trở nên khó hiểu và diễn giải.

2.9.5 Ứng dụng

- Trong các ứng dụng học máy như phân loại, dự đoán và gom cụm, StandardScaler là một phương pháp tiền xử lý dữ liệu phổ biến.
- Trong lĩnh vực tài chính, việc chuẩn hóa dữ liệu giúp phân tích mối quan hệ giữa các biến số tài chính một cách dễ dàng và chính xác hơn.
- Trong nghiên cứu y học và y sinh, việc chuẩn hóa dữ liệu giúp cải thiện khả năng dự đoán và phát hiện bất thường từ dữ liệu y sinh.

CHƯƠNG 3: PHƯƠNG PHÁP THỰC NGHIỆM

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" và "categories"của mỗi sản phẩm trong cửa hàng. Quá trình này lặp lại cho tối đa 80 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

Dữ liệu được thu thập ban đầu bao gồm 9500 mẫu về các thông tin của cửa hàng, chi tiết của sản phẩm và các lời bình luận của khách hàng về chất lượng sản phẩm. Dữ liệu được thu thập gồm 16 đặc chưng, bao gồm các thông tin về:

Tên biến	**Mô tả**	**Kiểu dữ liệu**
id	Chỉ số id của từng sản phẩm trong 1 cửa hàng	Interger
name	Tên sản phẩm	String
price	Giá tiền của các sản phẩm	Interger
rating_average	Điểm đánh giá sản phẩm	Float
review_count	Số lượng người đánh giá sản phẩm	Interger
quantity_sold	Số lượng sản phẩm đã được bán	Interger
quantity_sold_2weeks	Số lượng sản phẩm đã được bán sau 2 tuần	Interger
product_categories	Danh mục từng sản phẩm	String
shop_categories	Danh mục cửa hàng	String
Name_Shop	Tên cửa hàng	String
Shop_Rating	Điểm đánh giá của cửa hàng	Float
Year_Joined	Năm tham gia bán hàng trên Tiki	Interger
Followers	Số người theo dõi	Interger
Chat_Response	Tỷ lệ phản hồi chat	Interger
Reviews	Đánh giá của khách hàng về sản phẩm	String

Bảng 3.1: Bảng mô tả các biến và kiểu dữ liệu của chúng

3.3 Tiền xử lý dữ liệu

Quá trình này để xử lý các dữ liệu thô để đưa dữ liệu về bảng dữ liệu được sử dụng để phục vụ quá trình phân cụm.

3.3.1 Chuẩn hóa ký tự đặc biệt và emoji

Sau khi thu thập dữ liệu, các đánh giá về sản phẩm được lưu dưới dạng từ điển, trong đó mỗi đánh giá có cấu trúc như hình dưới đây:

"content": "Khăn giấy ướt giao nhanh đóng gói kỹ chất lượng giấy thật đúng như mô tả lau rất thích"

"content": "Sản phẩm dùng ok"

"content": "🍉Cuốn sách thay đổi cuộc đời.\n\n🍀Ai đã lớn, đã biết suy nghĩ thì phải nên đọc cuốn sách này 1 lần trong đời, ít nhất là 1 lần, nhiều nhất là "n" lần.😀"

Hình 3.1: Hình minh họa trước khi chuẩn hóa ký tự đặc biệt và emoji

Ta tiến hành chuyển đổi kiểu dữ liệu từ dạng chuỗi sang từ điển và trích xuất giá trị của mỗi từ điển. Điều này tạo ra tập hợp các đánh giá từ khách hàng cho từng sản phẩm, với mỗi đánh giá được phân bổ vào một cột riêng trong tập dữ liệu. Số lượng đánh giá thu thập được cho mỗi sản phẩm không đồng đều. Sau khi phân tích dữ liệu từ định dạng từ điển, thực hiện quy trình tiền xử lý bằng cách thay thế các dấu như ".", ",", "/n (xuống dòng)"thành dấu "."và loại bỏ tất cả các ký tự đặc biệt như: !, @, #, $, %, ^, &,...

Tiếp theo, loại bỏ các biểu tượng cảm xúc như Hình 3.2 khỏi mỗi đoạn văn và kết hợp tất cả các đánh giá vào một danh sách. Mỗi đánh giá được coi là một phần tử trong danh sách này. Mục đích của quy trình trên là làm sạch đoạn văn và chuẩn bị dữ liệu để có thể sử dụng trong thư viện underthesea.

😔, 👍, 😂, 🍉,...

Hình 3.2: Hình minh họa emoji

3.3.2 Chuẩn hóa dữ liệu Tiếng Việt

Sử dụng hàm `text_normalize()` trong thư viện underthesea, dữ liệu đánh giá từ khách hàng được chuẩn hóa để đảm bảo tính nhất quán và đồng nhất. 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 `text_normalize()` có thể sửa chữa các lỗi chính tả như "Đảm baỏ 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 của các mô hình và phương pháp xử lý dữ liệu.

Chuẩn hóa dữ liệu cũng đảm bảo tính nhất quán trong quá trình phân tích và đánh giá, đặc biệt là khi xử lý văn bản từ nhiều nguồn khác nhau. Kết quả là, quá trình phân tích và đánh giá dữ liệu trở nên mạch lạc hơn và dễ dàng hơn cho các mô hình và công cụ sau này.

3.3.3 Tách câu

Hàm `sentiment()` trong thư viện underthesea chỉ đạt hiệu suất cao khi phân loại cảm xúc với các câu ngắn, do đó việc tách một đoạn đánh giá dài của khách hàng thành các đoạn ngắn là cần thiết. Phân tách văn bản thành các câu riêng biệt giúp tập trung phân tích và hiểu rõ ý nghĩa của từng câu một. Điều này tăng cơ hội hiểu đúng ngữ cảnh và nội dung của văn bản.

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”. Các đoạn văn bản được tách được lưu vào như các phần tử của một danh sách và được lồng trong một danh sách lớn hơn chứa các đánh giá khác.

3.3.4 Phân loại cảm xúc văn bản

Phân loại cảm xúc đóng vai trò quan trọng trong việc giúp các tổ chức và doanh nghiệp hiểu rõ hơn về cảm xúc và ý kiến của người dùng đối với sản phẩm. Nó cung cấp thông tin chi tiết về nhu cầu, mong muốn và phản hồi của khách hàng, đồng thời đo lường mức độ hài lòng hoặc không hài lòng đối với sản phẩm hoặc dịch vụ cụ thể. Thông tin này quan trọng để cải thiện chất lượng và đáp ứng nhu cầu của khách hàng.

Trong bộ dữ liệu này, phân loại cảm xúc giúp xác định số lượng đánh giá tích cực và tiêu cực của từng sản phẩm, từ đó làm cho quá trình phân cụm dữ liệu trở nên trực quan hơn. Việc này đóng vai trò quan trọng trong việc đưa ra quyết định về sản phẩm và chiến lược kinh doanh.

Để thực hiện phân loại cảm xúc, tôi sử dụng hàm `sentiment()` của thư viện `underthesea` trên từng câu nhỏ trong một danh sách đã được tách ra từ bước trước. Sau đó, tôi sử dụng phương pháp voting để chọn ra cảm xúc của câu đó là tích hay tiêu cực. Đối với các dữ liệu đánh giá còn lại, tôi thực hiện tương tự. Kết quả là, tôi thu được 2 cột mới là số lượng đánh

giá tích cực và số lượng đánh giá tiêu cực của từng sản phẩm.

Cuối cùng, tôi tiến hành cộng tất cả các đánh giá tích cực của từng sản phẩm và tất cả các đánh giá tiêu cực của từng sản phẩm của mỗi cửa hàng và sinh ra được 2 cột mới là ‘‘positive_y’’ và ‘‘negative_y’’ biểu thị cho tổng số đánh giá tích cực và tiêu cực của từng cửa hàng.

3.4 Ước tính doanh thu từng cửa hàng

Phân đoạn này mô tả quá trình tiếp tục thu thập dữ liệu cửa từng cửa hàng để ước tính số lượng bán từng sản phẩm và doanh thu của các cửa hàng trong bộ dữ liệu. Việc này nhằm ước tính doanh thu sau 2 tuần giúp ta có thể đánh giá hiệu suất kinh doanh của cửa hàng trong khoảng thời gian cụ thể này.

Quá trình này diễn ra bằng cách thu thập dữ liệu số lượng bán của từng sản phẩm trong cửa hàng thông qua biến “id” của sản phẩm. Sau khi có được số lượng bán sau 2 tuần, ta tiến hành ước lượng doanh thu của từng cửa hàng bằng cách lấy số lượng bán sau 2 tuần trừ số lượng bán ban đầu, sau đó nhân với giá của sản phẩm. Sau đó, cộng tất cả các doanh thu của từng sản phẩm của mỗi cửa hàng ta được doanh thu của cửa hàng. Cuối cùng ta sinh ra được 1 đặc chưng mới là ‘‘revenue_y’’ tương ứng với tổng doanh thu từng cửa hàng.

3.5 Hợp nhất dữ liệu

Sau khi thực hiện các bước tiền xử lý ở trên ta đã có thể ra được 1 bảng dữ liệu mới để có thể đem đi xử lý trước khi đưa vào mô hình học máy. Ta tiến hành trích chọn các đặc trưng quan trọng bao gồm: `"shop_categories"`, `"Name_Shop"`, `"Shop_Rating"`: , `"Year_Joined"`, `"Followers"`, `"Chat_Response"`: , `"revenue_y"`, `"positive_y"`, `"negative_y"`. Bước tiếp theo là tách các đặc trưng quan trọng trên ra khỏi bộ dữ liệu ban đầu và tiến hành xóa các dữ liệu trùng lặp cho chúng.

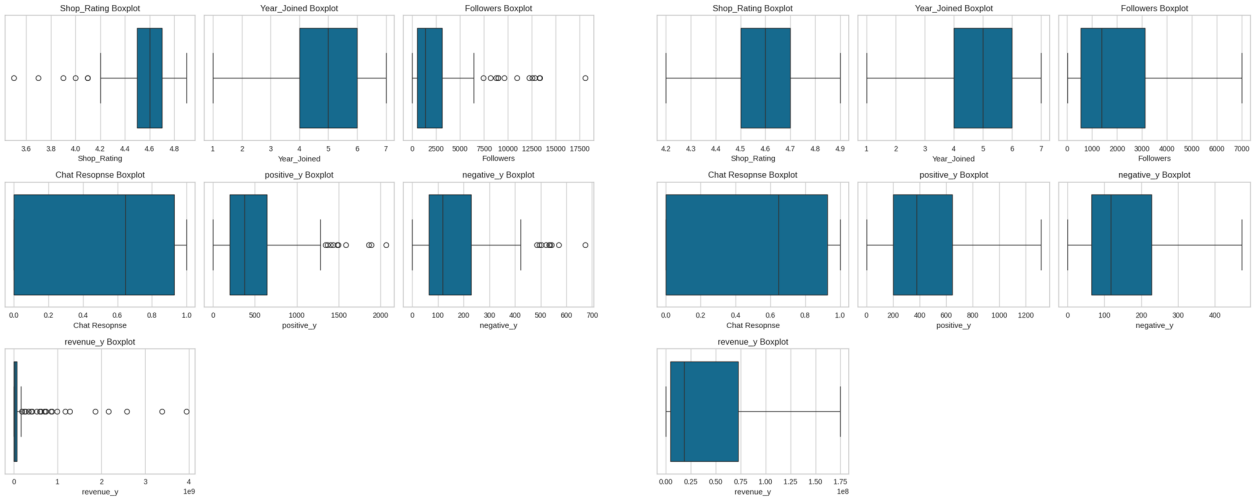
Bảng dữ liệu hoàn thiện của tôi gồm 9 đặc chưng như trên và 146 mẫu tương ứng với 146 cửa hàng. Tuy nhiên chỉ 7 trong số chúng được đem đi phân cụm.

3.6 Xử lý ngoại lai

Trong bài báo cáo này, tôi sử dụng kỹ thuật IQR (interquartile range) để xác định các giá trị ngoại lai. Kỹ thuật này dựa trên sự phân bố của dữ liệu thành các phần bằng nhau. Cụ thể, ta xác định giá trị q1 (quantile thứ 25), q3 (quantile thứ 75) và iqr (IQR) bằng cách lấy q3 trừ q1. Giá trị lower fence và upper fence được tính bằng cách lấy `(q1 - 1.5) * iqr` và `(q3 + 1.5) * iqr`. Các giá trị nằm ngoài upper_bound hoặc lower_bound được coi là

giá trị ngoại lai.

Trong bài báo cáo, ta nhận thấy có 5 cột trong bộ dữ liệu xuất hiện ngoại lai là: `"Shop_Rating"`: , `"Followers"`, `"revenue_y"`, `"positive_y"`, `"negative_y"`. Để xử lý chúng, ta duyệt qua từng cột dữ liệu và thực hiện các bước như đã nêu ở trên. Các giá trị ngoại lai sẽ được thay thế thành giá trị max nếu nằm phía trên của boxplot hoặc giá trị min nếu nằm phía dưới của boxplot.



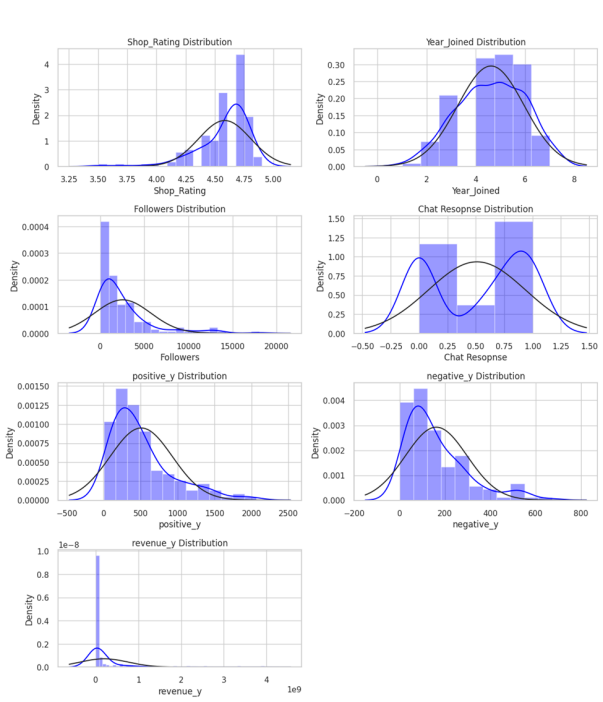
(a) Biểu đồ Boxplot trước khi xử lý ngoại lai. (b) Biểu đồ Boxplot sau khi xử lý ngoại lai.

Hình 3.3: So sánh biểu đồ Boxplot trước và sau khi xử lý ngoại lai.

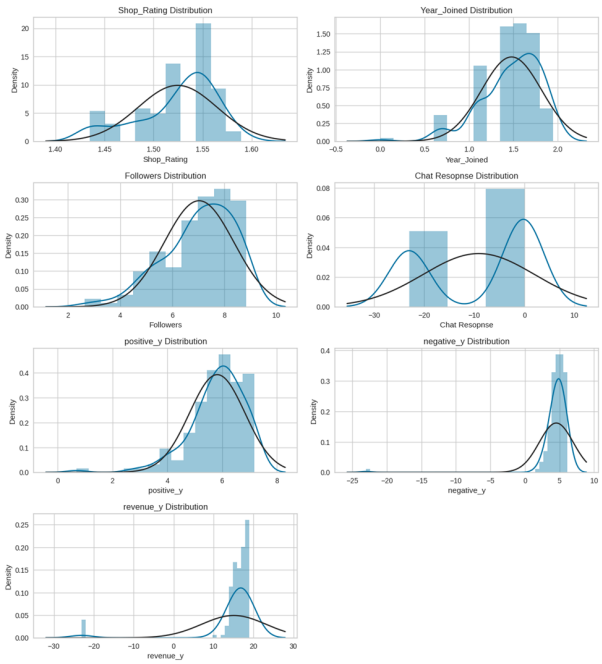
Sau khi xử lý, các giá trị ngoại lai đã được thay thế bằng các giá trị hợp lý hơn như Hình3.3(b), giúp giảm thiểu tác động của các yếu tố bất thường đến kết quả phân tích.

3.7 Chuyển đổi Log (Log Transformation)

Nhằm giảm thiểu ảnh hưởng của các giá trị ngoại lai và làm cho dữ liệu tuân theo phân phối gần chuẩn hơn, tôi đã sử dụng phương pháp phép biến đổi log. Phép biến đổi log giúp thu hẹp khoảng cách giữa các giá trị lớn và nhỏ, làm giảm độ lệch chuẩn và làm cho dữ liệu phân phối đều hơn. Điều này hữu ích cho dữ liệu gốc có sự chênh lệch lớn giữa các giá trị, chẳng hạn như cột **‘‘`revenue_y`’’** so với các cột còn lại. Bằng cách biến đổi dữ liệu, chúng ta có thể cải thiện độ chính xác và hiệu quả của thuật toán học máy, đặc biệt là các thuật toán nhạy cảm với sự phân phối của dữ liệu như K-Means. Hơn nữa, phép biến đổi log có thể giúp làm rõ các mối quan hệ tuyến tính tiềm ẩn trong dữ liệu và làm cho các mẫu trở nên rõ ràng hơn, từ đó hỗ trợ trong việc phát hiện các xu hướng và mẫu hình quan trọng.



(a) Biểu đồ phân phối dữ liệu ban đầu



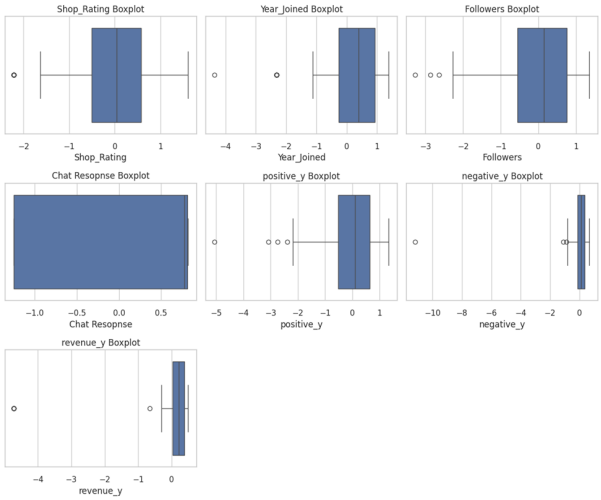
(b) Biểu đồ phân phối dữ liệu sau khi xử lý oulier và chuyển đổi log.

Hình 3.4: So sánh biểu đồ phân phối dữ liệu trước và sau khi xử lý outlier và chuyển đổi log.

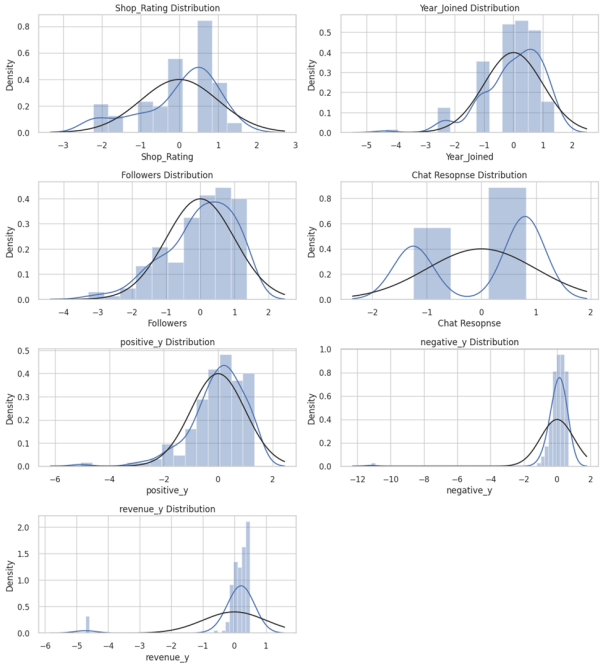
3.8 Chuẩn hóa dữ liệu

Mục đích của việc áp dụng phương pháp StandardScaler trong báo cáo này là để chuẩn hóa dữ liệu, giúp cải thiện hiệu suất và độ chính xác của các thuật toán học máy. StandardScaler chuyển đổi các đặc trưng (features) của dữ liệu sao cho mỗi đặc trưng có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Quá trình này giúp loại bỏ các khác biệt về quy mô và đơn vị đo lường giữa các đặc trưng, tạo điều kiện cho các thuật toán học máy hoạt động hiệu quả hơn.

Đối với thuật toán dựa trên khoảng cách K-Means thường bị ảnh hưởng bởi các đặc trưng có phạm vi giá trị lớn. Khi các đặc trưng được chuẩn hóa, mô hình có thể học từ dữ liệu một cách nhất quán và chính xác hơn.



(a) Biểu đồ Boxplot dữ liệu sau khi chuẩn hóa StandardScaler.



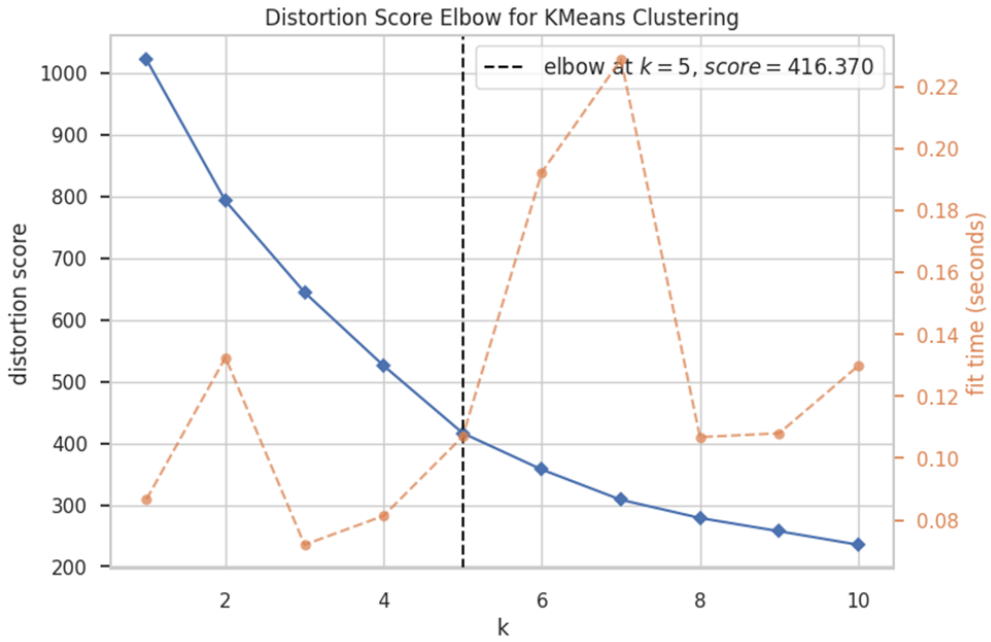
(b) Biểu đồ phân phối dữ liệu sau khi chuẩn hóa StandardScaler.

Hình 3.5: Biểu đồ so sánh dữ liệu sau khi chuẩn hóa StandardScaler.

3.9 Chọn số cụm tối ưu

Trước khi áp dụng thuật toán gom cụm (clustering) vào dữ liệu của tôi, một trong những quyết định quan trọng là chọn số cụm tối ưu. Đối với quá trình này, tôi sử dụng phương pháp Elbow (khuỷu tay) và Silhouette để xác định số lượng cụm phù hợp nhất cho dữ liệu của tôi.

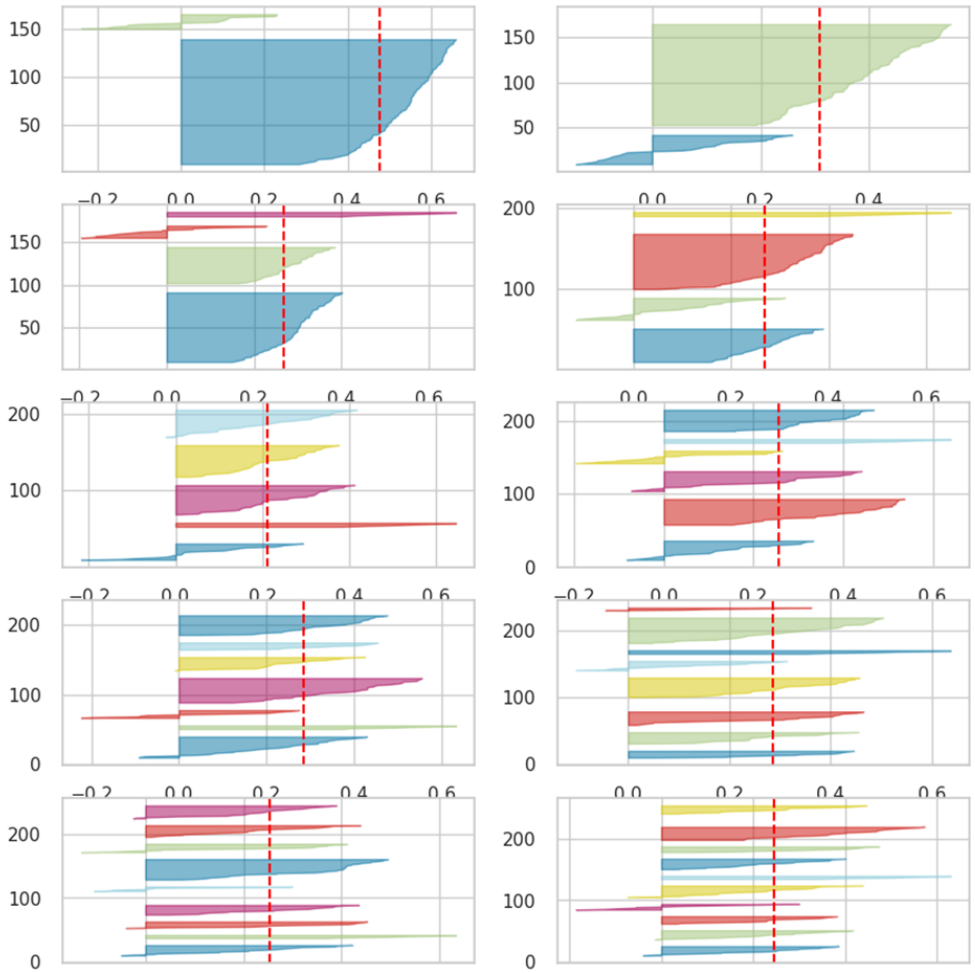
3.9.1 Elbow



Hình 3.6: Hình ảnh sau khi thực hiện phương pháp Elbow.

Kết quả của phương pháp Elbow cho ra số cụm tối ưu nhất có thể chọn là 5. Tôi quyết định đánh giá lại 1 lần nữa để chọn ra số cụm tối ưu bằng thuật toán phân tích hình bóng Silhouette.

3.9.2 Silhouette

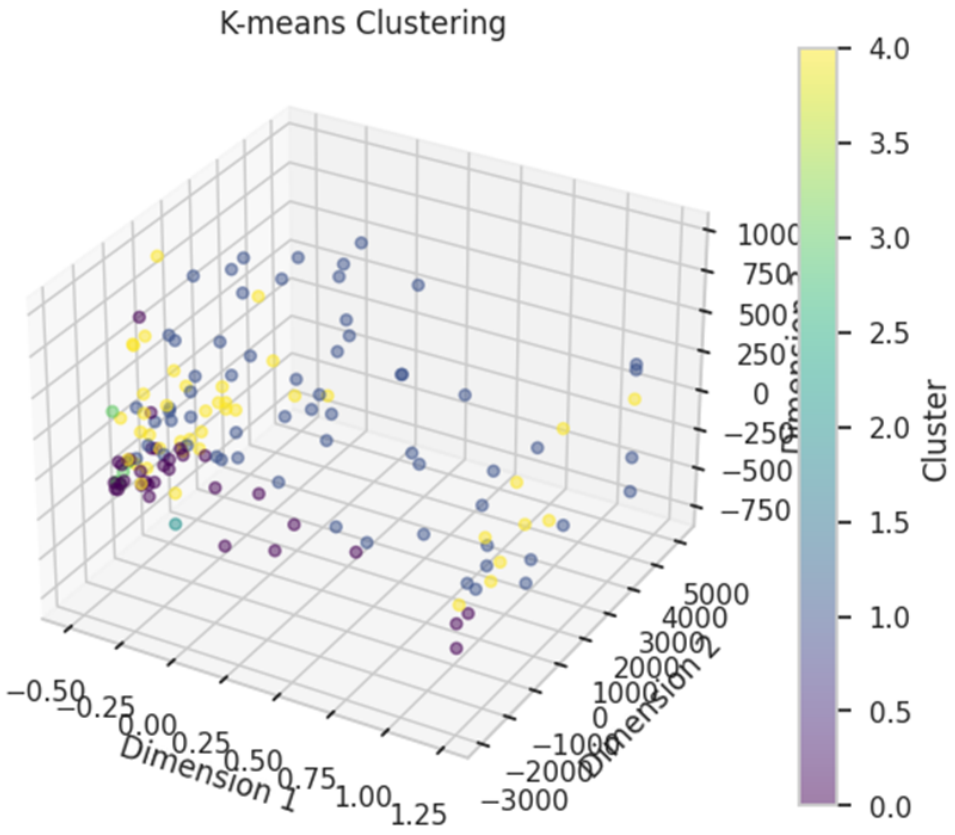


Hình 3.7: Hình ảnh sau khi thực hiện phương pháp Silhouette.

Kết quả của phương pháp cho ra số cụm tối ưu nhất có thể chọn là 5. Dựa trên kết quả của 2 phương pháp trên, chúng tôi quyết định chọn số cụm là 5 để thực hiện tiếp các phân tích gom cụm tiếp theo.

3.10 Phân cụm bằng thuật toán K-Means

Sau khi đã xác định số cụm tối ưu bằng phương pháp Elbow, chúng tôi tiến hành thực hiện phân cụm bằng thuật toán K-Means trên 7 đặc trưng quan trọng nhất đã được chọn ở phía trên.

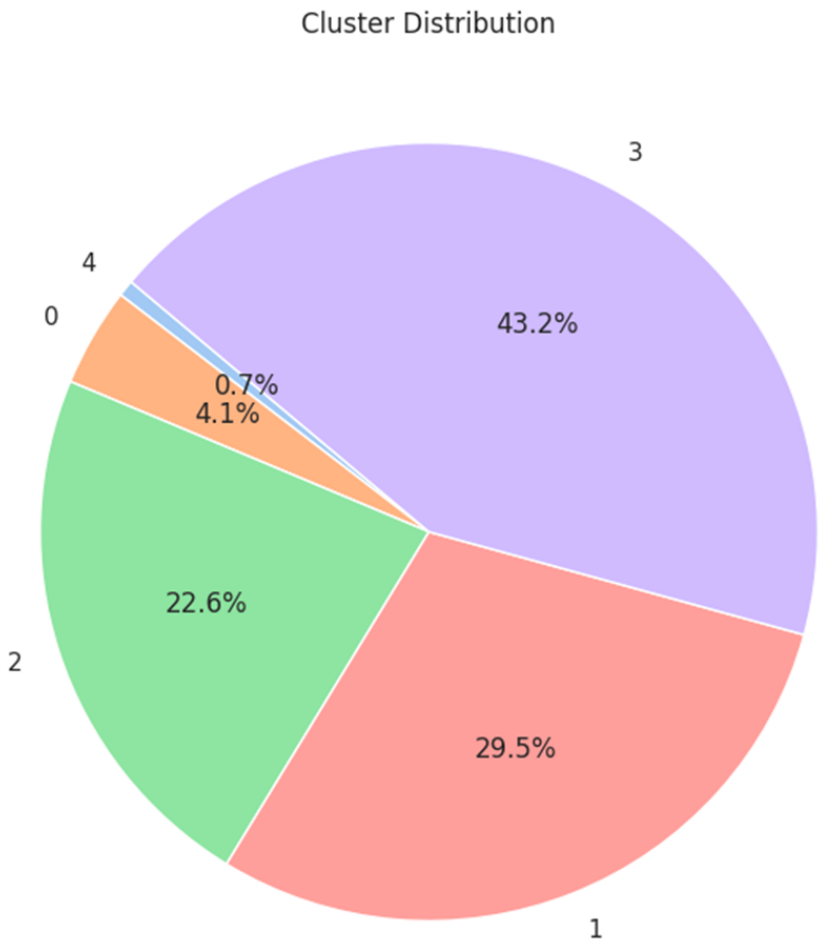


Hình 3.8: Hình ảnh sau khi thực hiện thuật toán K-Means.

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1 Phân tích cụm và đề xuất chiến lược

4.1.1 Phân tích cụm



Hình 4.1: Biểu đồ tròn thể hiện phần trăm các Cụm.

Dựa vào biểu đồ trên 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 cam) chiếm 4.1% tổng số cửa hàng.
- Cụm 1 (màu hồng) chiếm 29.5% tổng số cửa hàng.
- Cụm 2 (màu xanh lá) chiếm 22.6% tổng số cửa hàng.
- Cụm 3 (màu tím nhạt) chiếm 43.2% tổng số cửa hàng.
- Cụm 4 (màu xanh dương nhạt) chiếm 0.7% tổng số cửa hàng.

Nhận xét: Biểu đồ này giúp chúng ta dễ dàng hình dung và so sánh tỷ lệ phân bố của các nhóm cửa hàng sau khi phân cụm, cho thấy cụm 3 là nhóm có số lượng cửa hàng lớn nhất, sau đó tới cụm 1 và cụm 2. Do cụm 0 và cụm 4 có số lượng cửa hàng quá ít nên không thực hiện phân tích vào 2 cụm này.

Feature	Mean
Shop_Rating	4.59
Year_Joined	4.91
Followers	2308
Chat_Response	0
positive_y	465
negative_y	171
revenue_y	40.99

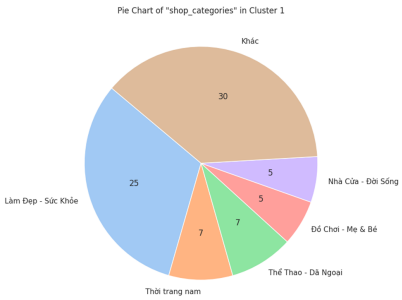
Bảng 4.1: Bảng phân tích cụm 1

Feature	Mean
Shop_Rating	4.60
Year_Joined	3.52
Followers	521
Chat_Response	65%
positive_y	204
negative_y	61
revenue_y	36.92

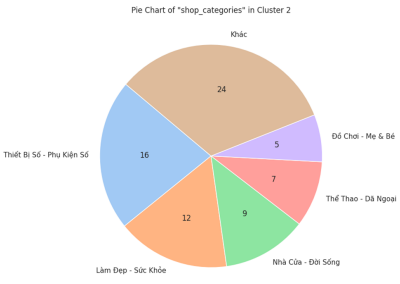
Bảng 4.2: Bảng phân tích cụm 2

Feature	Mean
Shop_Rating	4.60
Year_Joined	5.13
Followers	3149
Chat_Response	83%
positive_y	697
negative_y	213
revenue_y	69.66

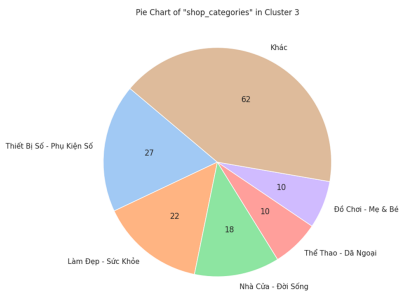
Bảng 4.3: Bảng phân tích cụm 3



(a) Biểu đồ Pie Chart biến "shop_categories" cho Cụm 1.



(b) Biểu đồ Pie Chart biến "shop_categories" cho Cụm 2.



(c) Biểu đồ Pie Chart biến "shop_categories" cho Cụm 3.

Hình 4.2: Biểu đồ Pie Chart biến "shop_categories" của các cụm.

Hiểu rõ phân khúc thị trường

Phân bố đánh giá cửa hàng:

- **Cụm 0:** Tỷ lệ đánh giá (`Shop_Rating`) cao (mean = 4.62) nhưng số người theo dõi `Followers` thấp (mean = 258). Các cửa hàng này có thể mới tham gia hoặc có sản phẩm độc đáo nhưng chưa được nhiều người biết đến.
- **Cụm 1:** Đánh giá cao (mean = 4.59) với số người theo dõi trung bình cao (mean = 2308). Các cửa hàng này có uy tín và lượng khách hàng ổn định.
- **Cụm 2:** Đánh giá trung bình (mean = 4.60) với số người theo dõi thấp (mean = 521). Các cửa hàng này có tiềm năng phát triển nhưng cần cải thiện marketing.
- **Cụm 3:** Đánh giá cao nhất (mean = 4.60) với số người theo dõi rất cao (mean = 3149). Đây là các cửa hàng thành công nhất, có thể là những thương hiệu lớn.
- **Cụm 4:** Số lượng cửa hàng chỉ có 1 nên không thực hiện phân tích.

Số năm tham gia bán hàng:

- **Cụm 0 và 2:** Các cửa hàng trong nhóm này thường mới hơn (mean năm tham gia lần lượt là 3.83 và 3.52).
- **Cụm 1 và 3:** Các cửa hàng có nhiều kinh nghiệm hơn (mean năm tham gia lần lượt là 4.91 và 5.13).

Phát triển chiến lược kinh doanh

Tỷ lệ phản hồi chat:

- **Cụm 0 và 1:** Tỷ lệ phản hồi chat bằng 0. Điều này có thể gây ảnh hưởng tiêu cực đến trải nghiệm khách hàng và cần được cải thiện.
- **Cụm 2:** Tỷ lệ phản hồi chat khá cao (mean = 0.65), cho thấy sự tương tác tốt với khách hàng.
- **Cụm 3:** Tỷ lệ phản hồi chat rất cao (mean = 0.83), đây là chuẩn mực mà các cửa hàng khác nên hướng tới.

Cải thiện chất lượng dịch vụ

Phản hồi tích cực và tiêu cực:

- **Cụm 0:** Số lượng đánh giá tích cực thấp (mean = 150) và tiêu cực cũng thấp (mean = 44).
- **Cụm 1:** Đánh giá tích cực cao (mean = 465) nhưng cũng có nhiều phản hồi tiêu cực (mean = 171).
- **Cụm 2:** Đánh giá tích cực trung bình (mean = 204) và tiêu cực thấp (mean = 61).
- **Cụm 3:** Đánh giá tích cực rất cao (mean = 697) nhưng cũng có nhiều phản hồi tiêu cực (mean = 213).

Quản lý hiệu suất

Doanh thu:

- **Cụm 0:** Không có doanh thu ghi nhận, có thể là các cửa hàng mới hoặc không hiệu quả.
- **Cụm 1:** Doanh thu trung bình (mean = 40.99 triệu đồng).
- **Cụm 2:** Doanh thu tương đối cao (mean = 36.92 triệu đồng).
- **Cụm 3:** Doanh thu rất cao (mean = 69.66 triệu đồng), cho thấy các cửa hàng này rất thành công.

Xu hướng thị trường

Danh mục sản phẩm:

- **Cụm 1:** Chủ yếu kinh doanh các mặt hàng về **Làm Đẹp - Sức Khỏe**.
- **Cụm 2:** Chủ yếu kinh doanh các mặt hàng về **Thiết Bị Số - Phụ Kiện Số**, **Làm Đẹp - Sức Khỏe**, **Nhà Cửa - Đời Sống**.
- **Cụm 3:** Chủ yếu kinh doanh các mặt hàng giống Cụm 2.

Ta có thể thấy được khách hàng có xu hướng mua các mặt hàng liên quan đến danh mục **Làm Đẹp - Sức Khỏe**, **Thiết Bị Số - Phụ Kiện Số**, **Nhà Cửa - Đời Sống** trong thời đại hiện nay.

Đào tạo và phát triển nhân viên

Tập trung vào dịch vụ khách hàng:

Các Cụm với tỷ lệ phản hồi chat thấp (Cụm 0 và 1) cần được chú trọng đào tạo về dịch vụ khách hàng để cải thiện trải nghiệm người mua.

4.2 Đề xuất chiến lược

4.2.1 Cụm 1: Cửa hàng có uy tín và lượng khách hàng ổn định

Chiến lược 1: Tối ưu hóa chất lượng sản phẩm và dịch vụ.

- Mục tiêu: Duy trì và nâng cao chất lượng dịch vụ.

- Hành động:

  - Phân tích các phản hồi tiêu cực để tìm ra nguyên nhân và cải thiện dịch vụ.
  - Đưa ra các chương trình chăm sóc khách hàng thân thiết, ví dụ như tích điểm đổi quà, giảm giá đặc biệt.

Chiến lược 2: Đẩy mạnh chương trình khách hàng thân thiết

- Mục tiêu: Tăng lượng khách hàng trung thành.

- Hành động:

  - Tạo các chương trình khách hàng thân thiết, ưu đãi dành riêng cho khách hàng thường xuyên.
  - Thu thập và phân tích dữ liệu khách hàng để cá nhân hóa các chiến dịch marketing.

4.2.2 Cụm 2: Cửa hàng có tiềm năng nhưng cần cải thiện marketing

Chiến lược 1: Nâng cao hiệu quả marketing.

- Mục tiêu: Tăng số lượng người theo dõi và doanh thu.

- Hành động:

  - Sử dụng SEO và content marketing để thu hút khách hàng.
  - Tạo các video quảng cáo, bài viết blog, và review sản phẩm trên các nền tảng mạng xã hội.

Chiến lược 2: Cải thiện chất lượng dịch vụ và sản phẩm

- Mục tiêu: Tăng số lượng đánh giá tích cực và giảm số lượng đánh giá tiêu cực.

- Hành động:

  - Đảm bảo chất lượng sản phẩm trước khi giao hàng.
  - Tăng cường dịch vụ hỗ trợ sau bán hàng.

4.2.3 Cụm 3: Cửa hàng thành công nhất

Chiến lược 1: Duy trì và phát triển chất lượng.

- Mục tiêu: Duy trì vị thế dẫn đầu và tăng doanh thu..

- Hành động:

  - Tăng cường các chương trình khách hàng VIP.
  - Phát triển thêm các sản phẩm hoặc dịch vụ mới để mở rộng thị trường.

Chiến lược 2: Xây dựng hình ảnh thương hiệu cao cấp.

- Mục tiêu: Nâng cao giá trị thương hiệu và lòng trung thành của khách hàng.

- Hành động:

  - Tổ chức các sự kiện offline hoặc online cho khách hàng VIP.
  - Xây dựng các nội dung quảng bá thương hiệu chất lượng cao, tập trung vào giá trị độc đáo của sản phẩm.

CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ

Qua nghiên cứu và phân tích dữ liệu các cửa hàng trực tuyến trên Tiki, chúng tôi nhận thấy rằng việc sử dụng thuật toán phân cụm K-means đã đem lại nhiều kết quả quan trọng. Thứ nhất, phân loại các cửa hàng thành các nhóm có đặc điểm và hiệu suất kinh doanh tương đồng đã giúp chúng tôi nhận diện rõ ràng các đặc trưng của từng nhóm, từ đó đề xuất chiến lược kinh doanh phù hợp. Thứ hai, thông tin chi tiết về khách hàng từ kết quả phân cụm đã hỗ trợ việc thiết kế chiến lược marketing và bán hàng tối ưu, tăng cường hiệu quả tiếp cận và tương tác với khách hàng mục tiêu. Thứ ba, việc nhận diện nhu cầu và mong muốn của từng nhóm khách hàng đã nâng cao chất lượng dịch vụ và trải nghiệm người dùng, giữ chân và thu hút khách hàng mới. Thứ tư, phân tích hiệu suất kinh doanh của từng nhóm cửa hàng đã hỗ trợ điều chỉnh và tối ưu hóa các hoạt động kinh doanh từ quản lý kho hàng đến phân phối sản phẩm. Cuối cùng, quá trình thu thập và phân tích dữ liệu đã tạo ra một cơ sở dữ liệu lớn và đa dạng, hỗ trợ cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực thương mại điện tử và học máy.

Để nâng cao hiệu quả và khả năng cạnh tranh, chúng tôi đề xuất một số kiến nghị. Thứ nhất, doanh nghiệp nên tận dụng kết quả phân cụm để phát triển các chiến lược kinh doanh tùy biến, điều chỉnh chiến lược marketing và phát triển sản phẩm phù hợp. Thứ hai, cần đầu tư mạnh mẽ vào hệ thống phân tích dữ liệu và học máy để liên tục cải thiện các chiến lược kinh doanh, nhận diện và phản ứng nhanh chóng với thị trường và hành vi khách hàng. Thứ ba, cần cải thiện trải nghiệm khách hàng bằng cách nâng cao chất lượng dịch vụ và cá nhân hóa trải nghiệm mua sắm. Thứ tư, nghiên cứu và tích hợp các công nghệ mới như trí tuệ nhân tạo, học sâu và IoT để dự đoán nhu cầu khách hàng và tối ưu hóa chuỗi cung ứng. Cuối cùng, cần đầu tư vào việc đào tạo và phát triển nguồn nhân lực trong phân tích dữ liệu và công nghệ thông tin để duy trì và nâng cao năng lực cạnh tranh trong thị trường. Thực hiện các kiến nghị này sẽ giúp tối ưu hóa hiệu quả hoạt động, nâng cao khả năng cạnh tranh và tạo ra giá trị bền vững trong thị trường thương mại điện tử, đồng thời đóng góp vào sự phát triển bền vững và hiệu quả của ngành.

TÀI LIỆU THAM KHẢO

[1] F. Aldi **andothers**. "Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning". **in***JAETS*: 5.1 (**december** 2023), **pages** 401–413.

[2] L. E. K. Huda Hamdan Ali. "K- Means Clustering Algorithm Applications in". **in***International Journal of Science and Research (IJSR)*: (2017).

[3] V. Anh. *Underthesea Documentation*. 2022. URL: https://underthesea.readthedocs.io/en/latest/.

[4] C. Benli. *Medium*. Accessed: 2023-09-21. **september** 2023. URL: https://medium.com/@mcbenli80/machine-learning-beddff9e3f46.

[5] Q. blog. *github*. URL: https://ndquy.github.io/posts/thuat-toan-phan-cum-kmeans/.

[6] Paula Dhiman **andothers**. "Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review". **in***BMC Medical Research Methodology*: 22.1 (2022), **pages** 75–105.

[7] H. W. N. L. T. C. H. H. Y. L. X. M. T. Changyong Feng. "Log-transformation and its implications for data analysis". **in***Shanghai Archives of Psychiatry*: 26.2 (2014), **pages** 105–109.

[8] Daniel Glez-Peña **andothers**. "Web scraping technologies in an API world". **in***Briefings in Bioinformatics*: 15.5 (2014), **pages** 788–797. DOI: 10.1093/bib/bbt026.

[9] *Học không có giám sát*. Accessed: 2021-08-14. URL: https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_kh%C3%B4ng_c%C3%B3_gi%C3%A1m_s%C3%A1t.

[10] International Journal of Science and Research (IJSR). "Machine Learning Algorithms - A Review". **in***International Journal of Science and Research (IJSR)*: 9.1 (2020).

[11] *Interquartile range*. Accessed: 2024-04-20. URL: https://en.wikipedia.org/wiki/Interquartile_range.

[12] P. Mahajan. *Introduction to Artificial Intelligence, Machine Learning, Deep Learning, & Convolutional Neural Networks - Invited International Guest Faculty - XIX Annual Conference on Evidence Based Management of Cancers in India "Technology and Cancer Care - Promise*. 2021.

[13] J. H. M. T. Husein Perez Perez. "Improving the Accuracy of Convolutional Neural Networks by Identifying and Removing Outlier Images in Datasets Using t-SNE". **in***Mathematics*: 8.5 (2020).

[14] T. T. R. **and** F. J. Hastie. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer New York, 2009.

[15] Muhamad Ramdani **and** Nagita Ramdhoni. "K-means cluster method for potential student grouping using elbow optimization". **in***AIP Conference Proceedings*: 2578.1 (2022).

[16] Muhamad Ramdani **and** Nagita Ramdhoni. "K-means cluster method for potential student grouping using elbow optimization". **in***AIP Conference Proceedings*: 2578.1 (2022).

[17] B. W. S. W. W. W. H. L. J. L. Congming Shi. "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm". **in***EURASIP Journal on Wireless Communications and Networking*: 2021 (2021), **page** 31.

[18] N. N. K. Meshal Shutaywi. "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering". **in***Entropy*: 23.6 (2021), **page** 759.

[19] R. T. **and** J. F. T. Hastie. *Unsupervised Learning*. New York, NY, USA: Springer New York, 2009.

[20] A. Tait. *IPBurger*. Accessed: 2023-05-26. URL: https://www.ipburger.com/vi/blog/web-scraping-using-api/#:~:text=API%20scraping%20l%C3%A0%20m%E1%BB%99t%20k%E1%BB%B9,web%20th%C6%B0%C6%A1ng%20m%E1%BA%A1i%20%C4%91i%E1%BB%87n%20t%E1%BB%AD..

[21] *underthesea*. URL: https://github.com/undertheseanlp/underthesea.