NGHIÊN CỨU PHÂN TÍCH HÀNH VI NGƯỜI DÙNG TRÊN WEBSITE BÁN TRÁI CÂY, PHÁT TRIỂN GIẢI PHÁP HỖ TRỢ QUYẾT ĐỊNH KINH DOANH THÔNG MINH

Nguyễn Hoàng Giang*, Vũ Hữu Hiếu, Nguyễn Tiến Duy
Nguyễn Hữu Hiếu, Nguyễn Đình Hoàng, Nguyễn Tuấn Tú
Vũ Thị Dương, Nguyễn Thái Cường
Trường Đại học Công nghiệp Hà Nội - K15 - Khoa CNTT - KTPM02

Tóm tắt

"Dữ liệu là dầu mới" - hàm ý nói về sự dồi dào của cả hai nguồn tài nguyên: Dầu và dữ liệu, tuy nhiên nếu cả hai ở trạng thái "thô" đều sẽ không có giá trị, do vậy việc thu thập dữ liệu chính xác, đầy đủ và phân tích chúng sẽ tạo nên giá trị cốt lõi của dữ liệu. Đồng thời giúp tổ chức phân loại, làm rõ các nhu cầu khách hàng, từ đó xây dựng được nhiều bước đột phá mới trong kinh doanh. Để tối ưu quá trình phân tích dữ liệu từ những thành phần thô ban đầu, phương pháp phân cụm được sử dụng như một phương pháp để tiếp cận tốt hơn với những giá trị mà nó mang lại. Bài báo sử dụng phương pháp phân cụm thông qua thuật toán K-Means. Phân tích dựa trên hơn 1.000 đơn hàng và thu được các cụm khách hàng có hành vi mua sắm tương tự. Nhờ vậy, thông qua quá trình phân loại các tổ chức có thể phát triển nhiều chiến lược kinh doanh tương ứng với từng nhóm khách hàng riêng biệt, cũng như cách tiếp cận với ngành kinh doanh dịch vụ nói chung.

Từ khóa: Dữ liệu; Phân cụm khách hàng; Thuật toán K-Means; Khách hàng mục tiêu; Trái cây.

Abstract

Researching and analyzing - user behavior on websites that sell fruit, giving solutions for supporting smart business decisions

The phrase "Data is the new oil" implies the abundance of both resources: Oil and data. However, if both remain in their "raw" states, they bring little value. Therefore, accurate data collection and thorough analysis form the core value of data. Simultaneously, this process helps organizations classify and clarify customer needs, leading to innovative breakthroughs in business. To optimize the analysis of data from its raw components, clustering methods are employed as an approach to better engage with the inherent values they offer. This article explores clustering methods using the K-Means algorithm. The analysis is based on over 1,000 orders, resulting in clusters of customers with similar shopping behaviors. Consequently, through this classification process, organizations can develop distinct business strategies tailored to each customer group, as well as adopt effective approaches to the broader service industry.

Keywords: Data; Customer segmentation; K-Means algorithm; Targeted customer; Fruit.

***Tác giả liên hệ, Email:** giangnh@st.haui.edu.vn
DOI: https://doi.org/10.63064/khtnmt.2024.560

1. Giới thiệu

Nghiên cứu và phân tích hành vi người dùng trên các trang mạng toàn cầu bán trái cây là một trong những lĩnh vực quan trọng trong ngành thương mại điện tử hiện đại. Trong thời đại mua sắm trực tuyến trở thành xu hướng và dễ dàng thực hiện hơn bao giờ hết. Việc hiểu rõ hành vi của người tiêu dùng trên các nền tảng trực tuyến trở nên cực kỳ quan trọng để tối ưu hóa trải nghiệm mua sắm, tăng cường doanh số bán và thu thập thông tin thị hiếu khách hàng. Nghiên cứu này tập trung vào việc thu thập và phân tích dữ liệu về các phương pháp người dùng tương tác với trang thông tin trên môi trường mạng toàn cầu (các trang web) bán trái cây, từ quá trình tìm kiếm sản phẩm đến việc chọn mua và hoàn tất giao dịch thanh toán.

Để tổng hợp và phân tích dữ liệu người dùng, nhóm tác giả sử dụng bài toán phân cụm, thuật toán K-Means, cùng với đó là áp dụng phương pháp Elbow, Silhouette để phân khúc khách hàng tại cơ sở kinh doanh trái cây trực tuyến. Nội dung chính của nghiên cứu với mục tiêu giải quyết các yêu cầu đặt ra là: Vận dụng phương pháp phân cụm hành vi khách hàng, sử dụng thuật toán K-Means với nguồn dữ liệu thông qua hơn 1.000 đơn đặt hàng trái cây có trong cơ sở dữ liệu của trang thông tin bán hàng (website), từ đó phân cụm khách hàng dựa vào hành vi (thanh toán và sử dụng), theo giá trị (giá trị trung bình của mỗi đơn đặt hàng, giá trị chi phí vận chuyển hàng hóa, đơn giá của từng loại sản phẩm), sở thích khách hàng (loại trái cây được nhiều lượt mua, phổ biến).

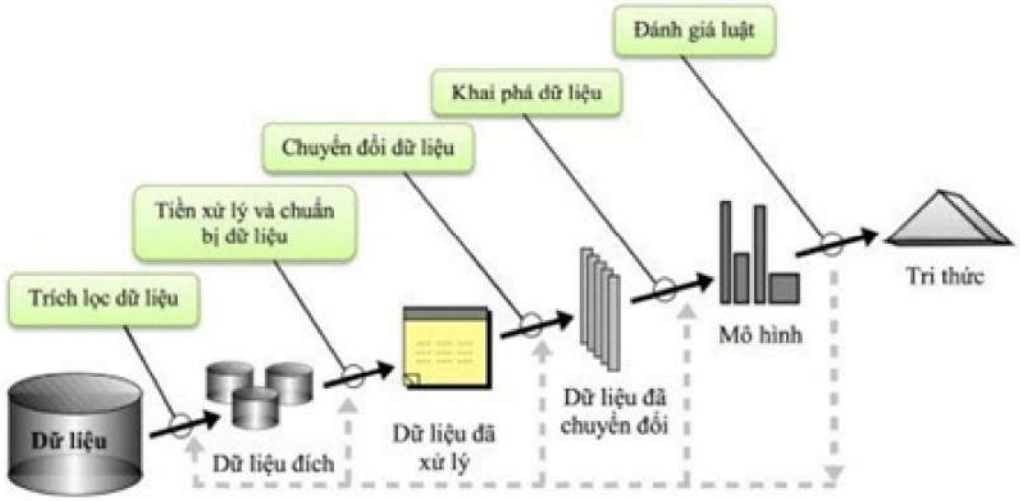
Nghiên cứu nhằm phục vụ cho các doanh nghiệp, tổ chức kinh doanh và từ đó gợi ý cho các cơ sở và ngành sản xuất, trồng trọt trong quá trình lựa chọn các vấn đề trọng tâm. Như vậy yếu tố quan trọng để nghiên cứu được áp dụng rộng rãi thì sự hiệu quả là yếu tố được đặt lên hàng đầu.

Hình 1: Gian hàng trái cây

2. Cơ sở lý thuyết

2.1. Khai phá dữ liệu

Quá trình khai phá dữ liệu là một trong những công việc mở ra cánh cửa cho việc khám phá và hiểu rõ về các mẫu, quy luật ẩn và tri thức chưa bộc lộ trong các tập dữ liệu. Từ dữ liệu không cấu trúc hoặc một lượng lớn thông tin, quá trình này sử dụng một loạt các kỹ thuật phân tích và mô hình hóa để tạo ra những thông tin hữu ích và hành động có ý nghĩa. Khai phá dữ liệu không chỉ là việc khám phá các mẫu và quy luật hiện hữu, mà còn là quá trình tạo ra các dự đoán và đề xuất thông minh dựa trên những phát hiện từ dữ liệu. Điều này có thể áp dụng trong nhiều lĩnh vực, từ dự đoán xu hướng thị trường và hành vi người tiêu dùng đến phát hiện gian lận và cải thiện dịch vụ tiếp cận khách hàng.

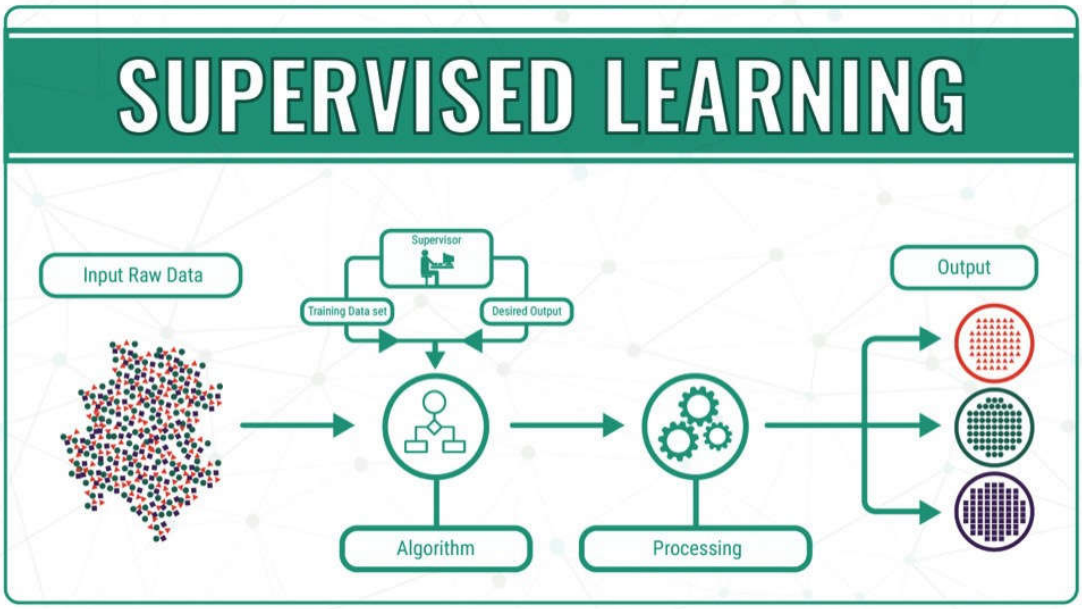


Hình 2: Các bước khai phá dữ liệu

Quá trình khai phá dữ liệu yêu cầu sự kết hợp giữa khoa học dữ liệu, máy học và hiểu biết sâu sắc về ngữ nghĩa của dữ liệu. Nó là công cụ mạnh mẽ giúp các tổ chức hiểu rõ hơn về mình và về thế giới xung quanh, đồng thời tạo ra cơ hội mới để cải thiện quy trình, tối ưu hóa chiến lược và định hướng tương lai.

2.1.1. Học có giám sát

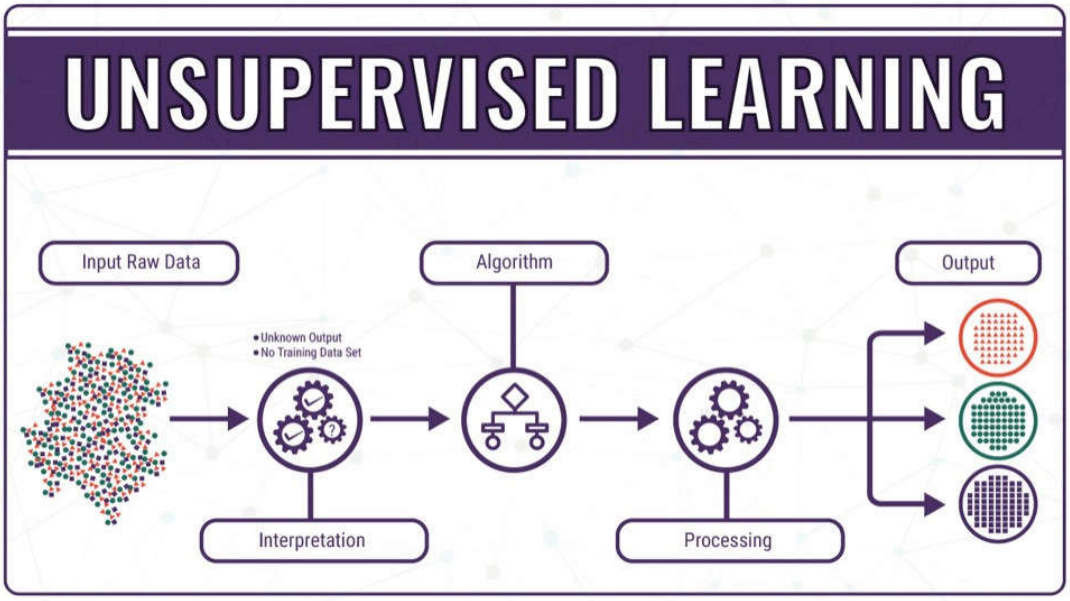
Đây là một lĩnh vực quan trọng của máy học, nơi mà dữ liệu được sử dụng để huấn luyện mô hình dự đoán hoặc phân loại dữ liệu mới. Thay vì chỉ xử lý dữ liệu mà không biết rõ nhãn hay kết quả mong muốn, các thuật toán học từ dữ liệu đã được gán nhãn trước đó và tiếp sau áp dụng những gì đã học vào dữ liệu mới. Điều này giúp chúng ta dự đoán các nhãn hoặc giá trị mong muốn cho dữ liệu mới dựa trên kinh nghiệm từ các dữ liệu đã được huấn luyện trước. Các thuật toán học có giám sát bao gồm nhiều kỹ thuật như phân loại và hồi quy, chúng được ứng dụng rộng rãi trong nhiều lĩnh vực như thị giác máy tính, xử lý ngôn ngữ tự nhiên, tài chính, y tế và nhiều lĩnh vực khác. Điều quan trọng là hiểu rõ cách sử dụng và lựa chọn đúng thuật toán phù hợp với vấn đề cụ thể để tạo ra mô hình dự đoán chính xác và hiệu quả.



Hình 3: Học có giám sát

2.1.2. Học không giám sát

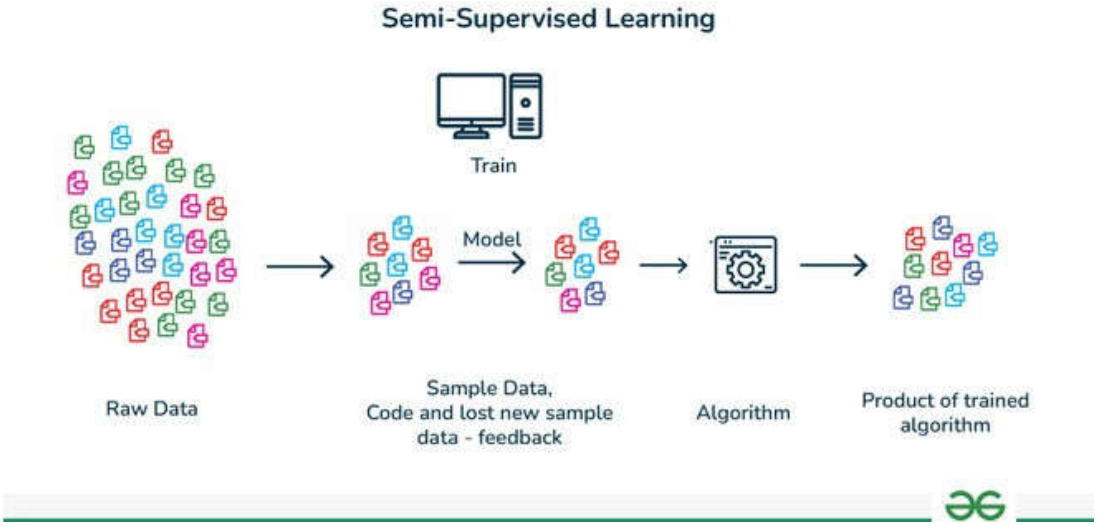
Nhóm thuật toán học không giám sát là một phần quan trọng trong máy học, tập trung vào việc khám phá cấu trúc ẩn trong dữ liệu mà không cần gán nhãn từ con người. Các thuật toán này thông qua các kỹ thuật phân cụm và giảm chiều dữ liệu giúp chúng ta hiểu rõ hơn về các mẫu, cấu trúc và nhóm trong dữ liệu. Các kỹ thuật phổ biến bao gồm kỹ thuật Phân cụm K-Means (K-Means clustering) và Phân tích thành phần chính (Principal Component Analysis - PCA) là các thuật toán, được áp dụng rộng rãi trong nhiều lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên.



Hình 4: Học không giám sát

2.1.3. Học bán giám sát

Học bán giám sát là một nhóm quan trọng trong máy học, kết hợp giữa các yếu tố của học có giám sát và không giám sát. Trong kỹ thuật này, một phần của dữ liệu được gán nhãn và một phần không. Mục tiêu kỹ thuật sử dụng thông tin nhãn có sẵn để học từ dữ liệu không nhãn, giúp cải thiện chất lượng và hiệu suất của mô hình học máy. Các phương pháp trong học bán giám sát bao gồm học chuyển giao (Transfer Learning), học bán giám sát (Semi-Supervised Learning) và học tự giám sát (Self-Supervised Learning). Đây là một lĩnh vực quan trọng trong máy học, đặc biệt khi dữ liệu có sẵn không đủ hoặc tốn kém để thu thập đủ nhãn phục vụ cho việc huấn luyện mô hình.



Hình 5: Học bán giám sát

2.2. Phân cụm dữ liệu - thuật toán K-Means

2.2.1. Khái niệm về phân cụm

Phân cụm dữ liệu là quá trình tự động chia nhóm các điểm dữ liệu thành các nhóm có ý nghĩa dựa trên yếu tố tương đồng. Mục tiêu của phân cụm tạo ra nhiều nhất các nhóm các điểm trong cùng một nhóm có sự tương đồng. Các nhóm khi phân loại càng khác biệt càng tốt (lớn nhất có thể). Các phương pháp phân cụm dữ liệu phổ biến có thể kể đến như K-Means, phân cụm dữ liệu (Hierarchical Clustering) và DBSCAN.

2.2.2. Thuật toán K-Means

Thuật toán K-Means là một trong những phương pháp phân cụm dữ liệu phổ biến nhất. Quá trình thực hiện thuật toán K-means bao gồm các bước sau:

- Bước 1 - Chọn số cụm (K): Trước tiên chọn số lượng cụm mà bạn muốn dữ liệu được phân thành. Số lượng cụm này thường được gọi là K.

- Bước 2 - Khởi tạo các điểm trung tâm ban đầu: Chọn ngẫu nhiên K điểm từ dữ liệu làm các điểm trung tâm ban đầu của các cụm.

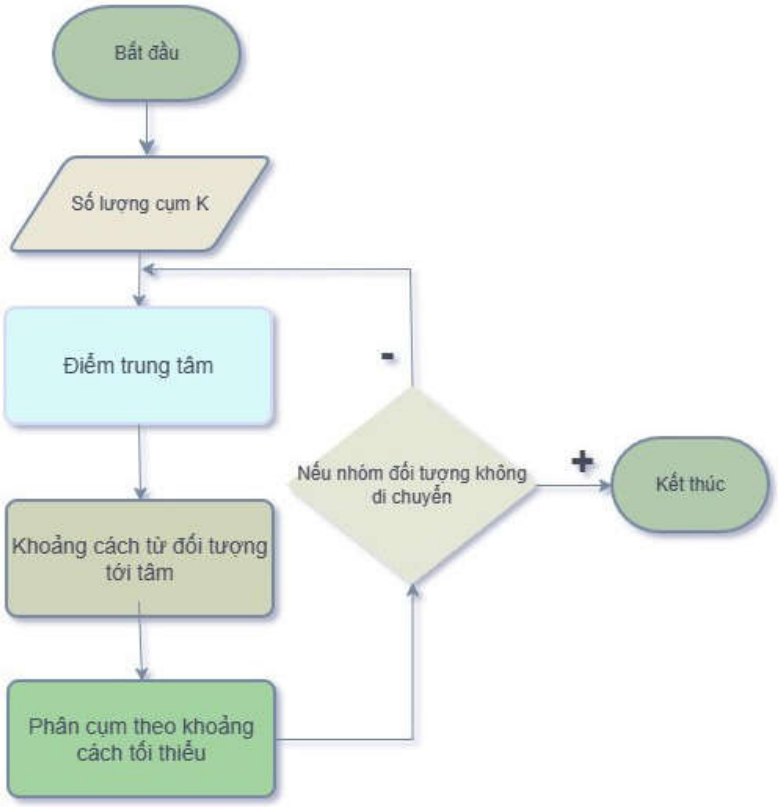
- Bước 3 - Xác định khoảng cách giữa từng điểm dữ liệu tới các điểm trung tâm, sau đó gán từng điểm dữ liệu vào cụm có điểm trung tâm gần với điểm đó nhất.

- Bước 4 - Cập nhật các điểm trung tâm: Tính toán lại các điểm trung tâm của các cụm dựa trên dữ liệu mới được gán vào từ bước trước.

- Bước 5 - Lặp lại quá trình: Thực hiện lặp lại các bước 3 và 4 cho đến khi không có sự thay đổi đáng kể nào trong việc gán các điểm dữ liệu vào các cụm hoặc khi đạt đến số lần lặp đã xác định trước.

- Bước 6 - Kết thúc: Quá trình hội tụ, thuật toán K-Means sẽ dừng lại và trả về các cụm đã phân chia bao gồm cả các điểm trung tâm tương ứng.

Tuy nhiên, việc chọn lựa số lượng cụm đúng (K) và khởi tạo các điểm trung tâm ban đầu tốt có thể ảnh hưởng đến hiệu suất của thuật toán. Do đó, để giảm các sai số ta nên thực hiện nhiều lần các thử nghiệm với các giá trị K khác nhau hoặc phương pháp khởi tạo khác nhau để tìm ra kết quả tốt nhất.



Hình 6: Lưu đồ thuật toán K-Means

Như vậy K-Means là thuật toán đi tìm lời giải cho bài toán tối ưu sau:

$$\min_{u,c} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij} d(i,j)$$

$$s.t \sum_{j=1}^{C} u_{ij} = 1, u_{ij} \in \{0,1\}$$

trong đó:

$$d(i,j) = \left\| x_i - c_j \right\|^2$$

là cự ly bình phương giữa dữ liệu x_i và vector c_j đại diện cho lớp có nhãn j. Và u_{ij} là mức độ phụ thuộc của x_i trong lớp có nhãn j.

Đầu ra của các vector c_j trong thuật toán K-Means được tính bởi công thức sau:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij} x_i}{\sum_{i=1}^{N} u_{ij}}$$

3. Kết quả nghiên cứu

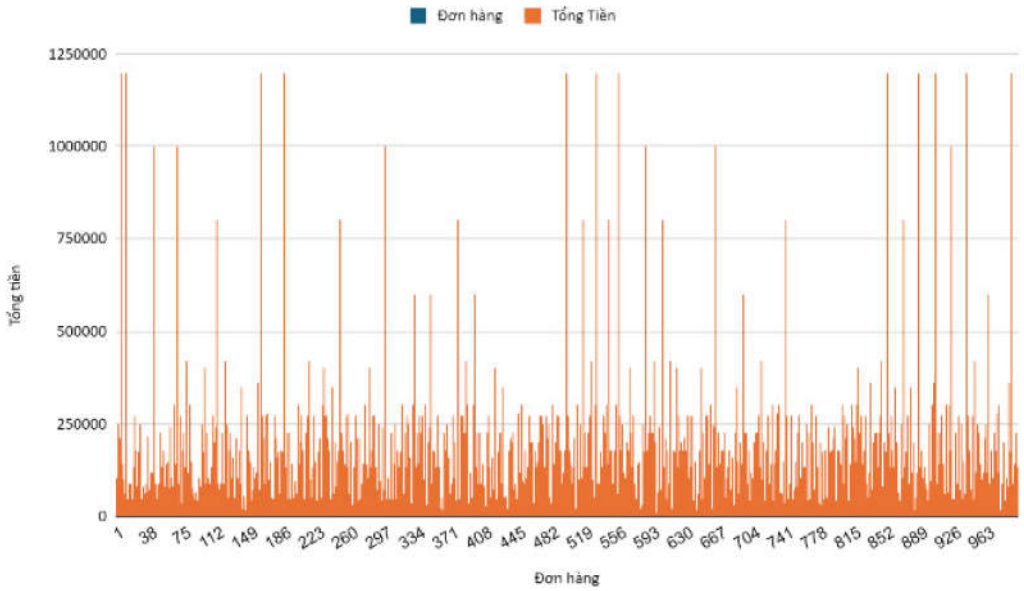
3.1. Mô tả dữ liệu

Nghiên cứu thu thập thông tin của hơn 1.000 khách hàng trên website bán trái cây trực tuyến của doanh nghiệp từ ngày 01/01/2023 - 31/12/2023, các thông tin được tập hợp gồm 10 cột: Mã đơn hàng, tên trái cây, tên khách hàng, giới tính khách mua hàng, đơn giá/kg, trọng lượng hàng hoá, số tiền đã mua sắm, chi phí vận chuyển, hình thức thanh toán, ngày đặt hàng. Hình 7 mô tả một phần dữ liệu.

InvoiceNo	ProductName	CustomerID	Customer	Sex	UnitPrice	Quantity	Tổng Tiền	Phí vận chuyển	PT Thanh Toán	InvoiceDate
35	Đu đủ	KH35	Nguyễn Văn Hoàng	Nam	16000	4	64.000 đ	55000	Khi nhận hàng	2023-03-04 06:24:40
36	Nho	KH19	Lê Thị Thùy Linh	Nữ	43000	5	215.000 đ	60000	Ví điện tử MOMO	2023-11-20 15:47:43
37	Dưa lưới	KH38	Phạm Minh Quân	Nam	35000	2	70.000 đ	30000	Khi nhận hàng	2023-03-30 04:51:14
38	Lê	KH18	Lê Thị Thanh Thảo	Nữ	35000	2	70.000 đ	30000	Khi nhận hàng	2023-06-30 00:11:38
39	Kiwi	KH08	Hoàng Minh Tuấn	Nam	20000	6	120.000 đ	90000	VNPAY	2023-03-06 10:45:41
40	Măng cụt	KH27	Nguyễn Thị Hương Giang	Nữ	50000	1	50.000 đ	15000	Zalo Pay	2023-12-30 08:00:08
41	Bưởi	KH100	Vũ Văn Hiếu	Nam	60000	2	120.000 đ	30000	Ví điện tử MOMO	2023-03-05 09:59:20
42	Quýt	KH17	Lê Thị Quỳnh Anh	Nữ	35000	5	175.000 đ	60000	Ví điện tử MOMO	2023-04-19 23:12:35
43	Dâu tây	KH53	Trần Văn Thanh	Nam	200000	5	1.000.000 đ	60000	Khi nhận hàng	2023-02-14 14:36:04
44	Nhãn	KH15	Lê Thị Kim Ngân	Nữ	45000	2	90.000 đ	30000	Ví điện tử MOMO	2023-08-06 16:02:22
45	Vải	KH06	Đỗ Văn Long	Nam	45000	1	45.000 đ	15000	Zalo Pay	2023-07-05 07:04:29
46	Thanh long	KH29	Nguyễn Thị Mai Phương	Nữ	45000	1	45.000 đ	15000	Ví điện tử MOMO	2023-10-16 20:40:02
47	Dừa	KH16	Lê Thị Phương Linh	Nữ	45000	2	90.000 đ	30000	Khi nhận hàng	2023-05-29 10:51:53
48	Ổi	KH52	Trần Văn Nam	Nam	35000	2	70.000 đ	30000	VNPAY	2023-04-13 04:01:24
49	Lựu	KH33	Nguyễn Thị Thu Trang	Nữ	45000	2	90.000 đ	30000	Ví điện tử MOMO	2023-05-04 06:06:28
50	Bơ	KH41	Phạm Văn Hoàng	Nam	45000	5	225.000 đ	60000	Zalo Pay	2023-04-11 07:33:35
51	Mít	KH02	Bùi Thị Kim Ngọc	Nữ	45000	3	135.000 đ	45000	Zalo Pay	2023-06-15 17:29:46
52	Mận	KH21	Lê Văn Đức	Nam	45000	3	135.000 đ	45000	Zalo Pay	2023-06-03 22:58:06
53	Khế	KH26	Nguyễn Thị Bích Thảo	Nữ	45000	4	180.000 đ	55000	Khi nhận hàng	2023-02-02 04:55:19
54	Nhót	KH95	Trần Văn Tuấn Anh	Nam	70000	1	70.000 đ	15000	Ví điện tử MOMO	2023-06-15 15:53:35
55	Me	KH32	Nguyễn Thị Thu Hương	Nữ	20000	4	80.000 đ	55000	Zalo Pay	2023-03-26 07:24:04
56	Dưa hấu	KH07	Hoàng Minh Trí	Nam	35000	4	140.000 đ	55000	Ví điện tử MOMO	2023-07-22 00:04:10

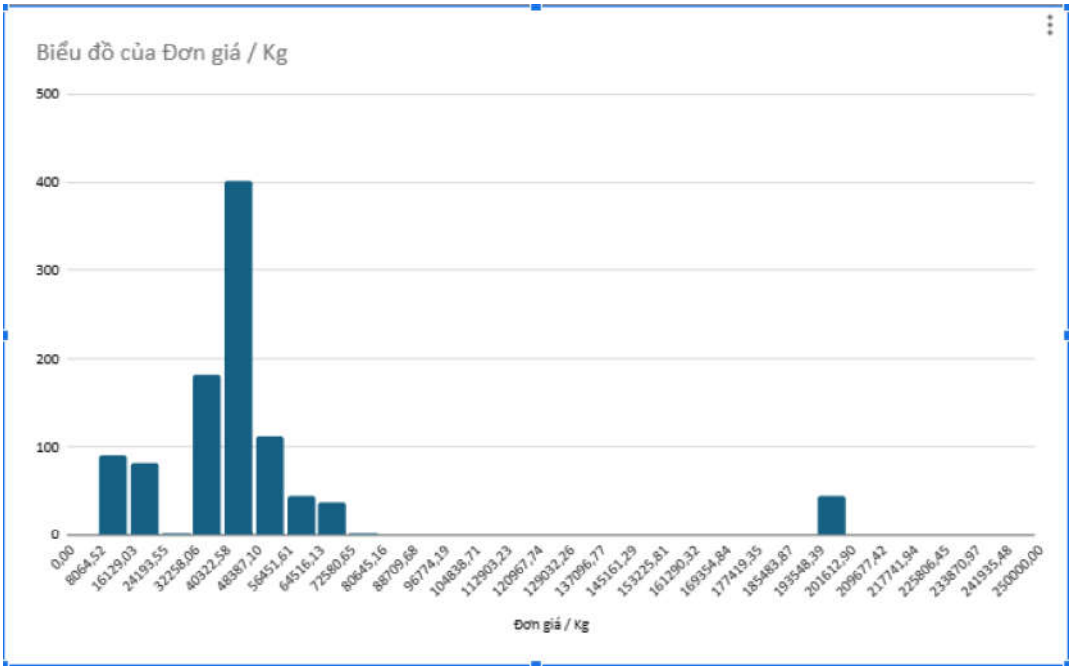
Hình 7: Mô tả dữ liệu

Mô tả tổng số tiền/đơn hàng mà khách hàng thanh toán. Tỷ lệ này đa phần sẽ giao động từ 0 - 250.000 đồng. Hình 8 thể hiện phân bố tổng tiền qua biểu đồ dạng đường.



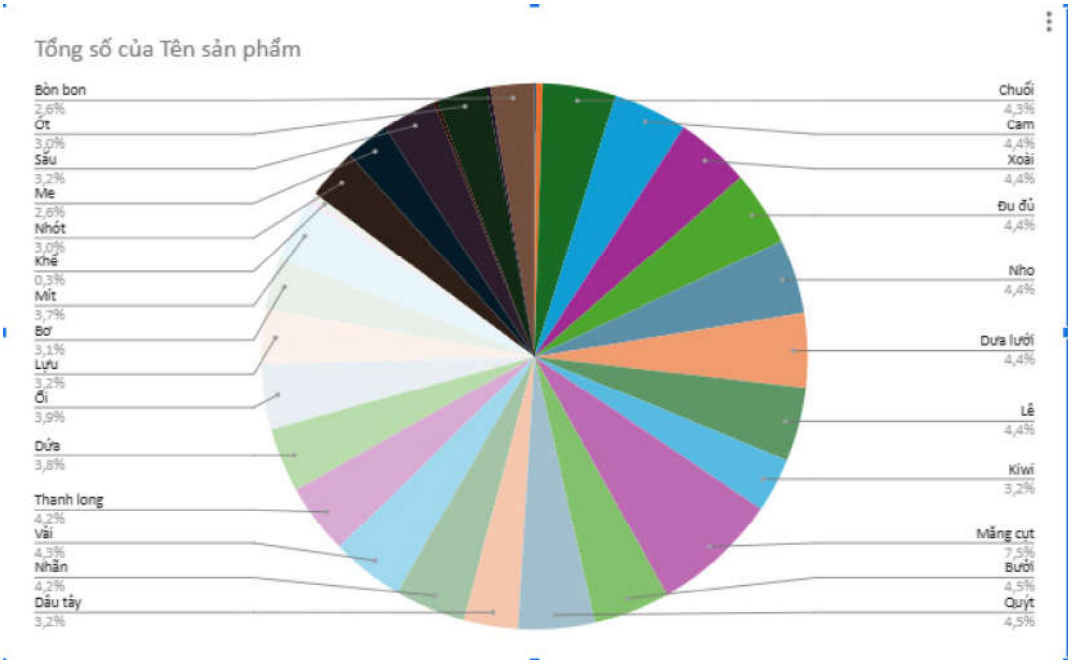
Hình 8: Biểu đồ dạng đường theo tổng tiền

Với Hình 9 cho thấy sản phẩm chủ yếu được lựa chọn theo đơn giá từ 7.000 - 70.000 đồng, thể hiện biểu đồ phân bổ đơn giá của sản phẩm theo dạng cột.



Hình 9: Biểu đồ phân bố đơn giá/kg

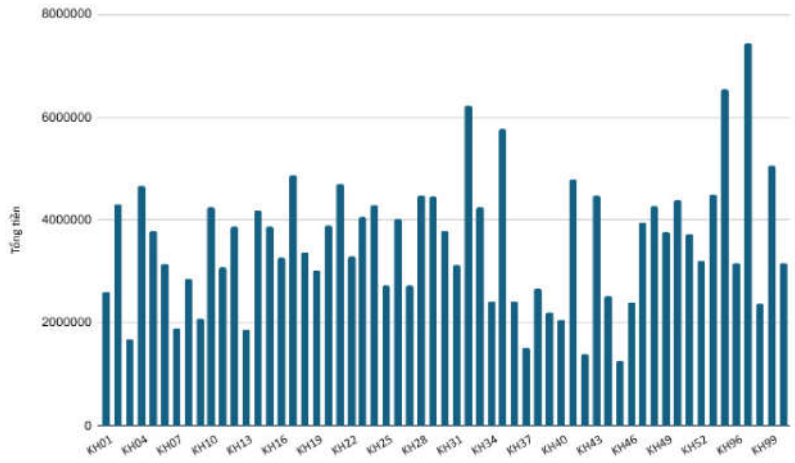
Dựa vào dữ liệu tên sản phẩm đi kèm thông tin khách hàng, mức độ tiêu thụ của từng loại sản phẩm có thể được mô tả thông qua biểu đồ Hình 10.



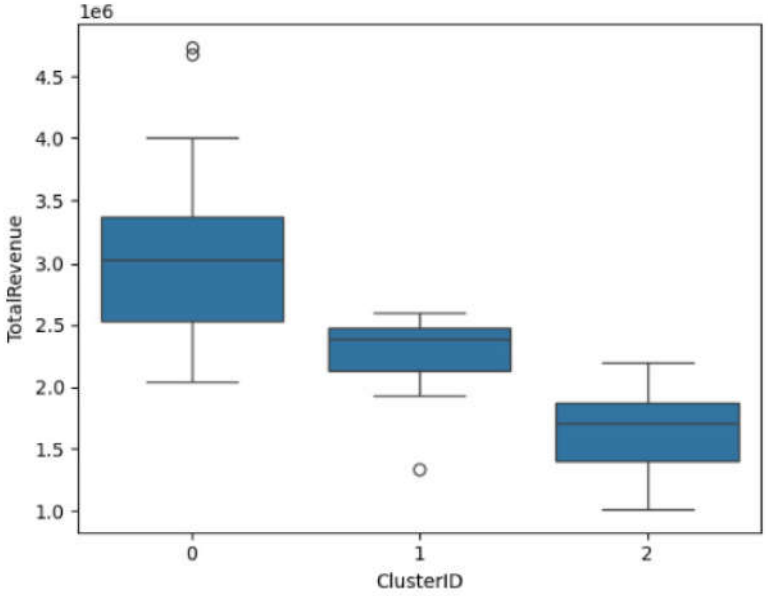
Hình 10: Biểu đồ mức độ tiêu thụ của từng loại sản phẩm

3.2. Kết quả nghiên cứu dữ liệu

Bằng phương pháp Elbow Method: Nghiên cứu xác định số cụm tối ưu để phân bổ khách hàng là 3 cụm như Hình 12, 13 và 14. Đây là số cụm nên phân bổ theo phương pháp này. Tuy nhiên, nếu cần doanh nghiệp có thể phân cụm nhiều hơn lên tới k = 3, k = 4,…cũng có thể nhiều hơn thế.



Hình 11: Biểu đồ cột mô tả mức thanh toán của khách hàng



Hình 12: Phân cụm dữ liệu theo tổng doanh thu

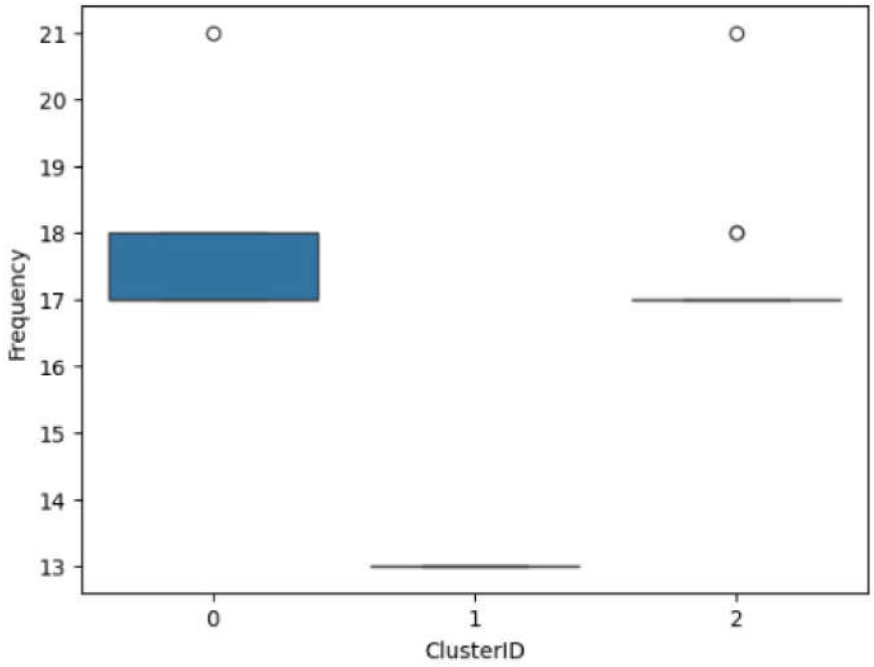
Sau khi sử dụng thuật toán K-Means, kết hợp với phương pháp Elbow với trường dữ liệu là tổng chi tiêu của khách hàng đã chi tiêu (Hình 12), thuật toán K-Means đã chia dữ liệu thành 3 cụm (Hình 14) với các cụm như sau:

- Cụm 0: Tổng chi tiêu trong khoảng từ 2.500.000 - 3.500.000 đồng.

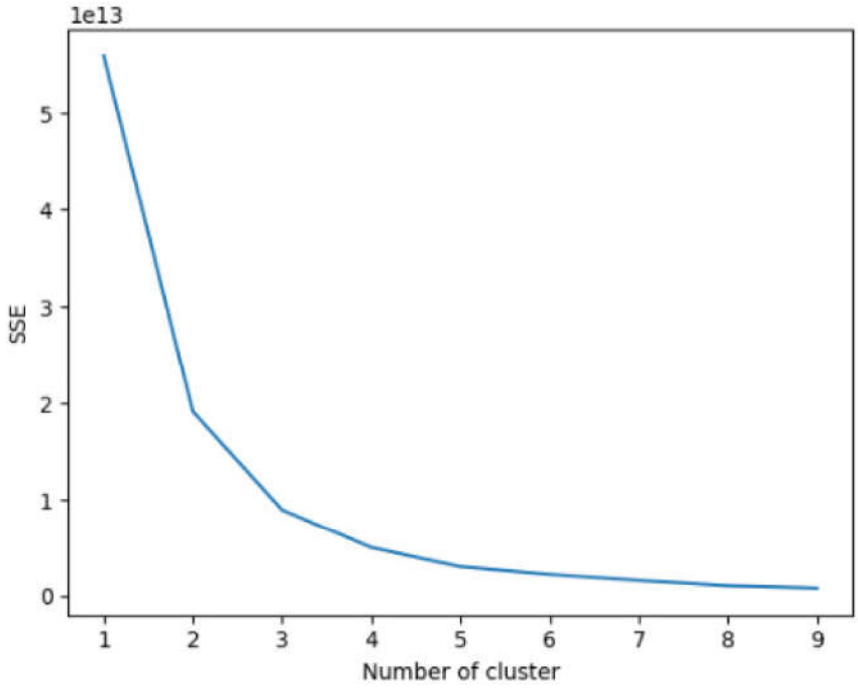
- Cụm 1: Tổng chi tiêu trong khoảng từ 2.000.000 - 2.500.000 đồng.

- Cụm 2: Tổng chi tiêu trong khoảng từ 1.500.000 - 2.000.000 đồng.

Qua đó, có thể đánh giá được lượng khách hàng quay trở lại mua hàng tại công ty chưa nhiều khi tổng chi tiêu của khách hàng nhiều nhất chỉ trong khoảng 2.500.000 - 3.500.000 đồng, từ đó công ty nên đưa ra nhiều chính sách ưu đãi hơn với khách hàng trung thành và cải thiện dịch vụ chăm sóc khách hàng, để từ đó có thể thu hút lượng khách hàng quay trở lại mua hàng tại công ty ngày một nhiều hơn.



Hình 13: Phân cụm dựa trên tần suất mua hàng



Hình 14: Xác định số lượng cụm tối ưu bằng phương pháp Elbow

Như vậy với cùng một trường dữ liệu đầu vào giống nhau là tổng chi tiêu của khách hàng, Hình 11 được tổng hợp theo phương pháp thông thường và Hình 12 được phân cụm bằng thuật toán K-Means, có thể thấy ưu điểm vượt trội của K-Means bởi thuật toán đã phân được ra thành các cụm với số lượng giá trị tương ứng, từ đó công ty có thể phát triển nhiều chiến lược kinh doanh tương ứng với từng nhóm khách hàng riêng biệt.

4. Thảo luận

Phân cụm dữ liệu bằng thuật toán K-Means cho ra những kết quả cụ thể tương ứng với từng tệp dữ liệu:

- Hình 12, 13 cho thấy nhóm khách hàng có nhãn 0 là khách hàng đem lại nhiều doanh thu cho công ty nhất với tần suất mua hàng thường xuyên. Các số liệu chỉ ra khách mua hàng gần đây nhất, chứng tỏ đây là khách hàng trung thành.

- Nhóm khách hàng có nhãn 1 là khách hàng đem lại ít doanh thu cho công ty, tần suất mua hàng thấp và thời gian gần nhất tiếp xúc với khách khá xa tính thời điểm hiện tại. Vậy đây là nhóm khách hàng ít trung thành.

- Nhóm khách hàng có nhãn 2 là nhóm khách hàng mới mua hàng ở công ty. Do vậy doanh thu và tần suất mua hàng của nhóm này chưa cao. Tuy nhiên, vì là khách hàng mới nên có thể ta cần có các chính sách chăm sóc, quảng bá tới nhóm này nhiều hơn để họ trở thành những khách hàng tiềm năng.

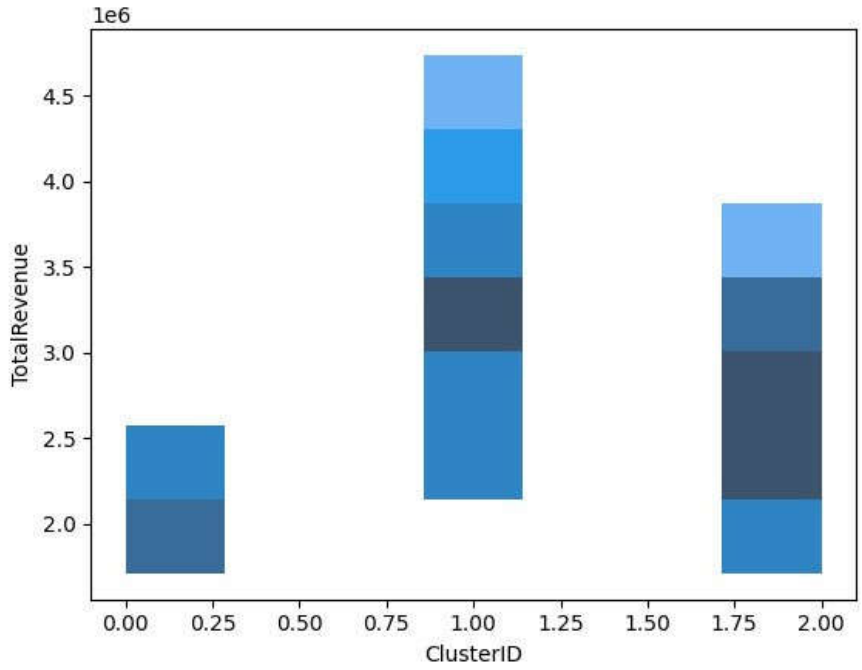
Việc phân cụm thành công tạo tiền đề thuận lợi cho công đoạn phân tích hành vi người dùng. Từ đó, dễ dàng hơn trong việc đưa ra chiến lược kinh doanh phù hợp cho từng nhóm khách hàng, cụ thể như:

- Với nhóm khách hàng trung thành: Nâng cao dịch vụ chăm sóc khách hàng, tạo ưu đãi thường niên, gia tăng hạn mức trung thành.

- Với nhóm khách hàng mới: Cung cấp những ưu đãi về giá, các ưu đãi hấp dẫn để có thể nâng cao hiệu suất mua hàng của nhóm đối tượng này.

- Với nhóm khách hàng đã từng mua hàng: Tạo sự kiện quảng bá để khách trở lại kèm theo nhiều phần quà hấp dẫn, thu hút sự tin tưởng cho lần mua hàng tiếp theo sau một thời gian.

Dữ liệu người dùng là tài sản quý giá và việc nghiên cứu, tận dụng nó đóng vai trò quan trọng trong quá trình phát triển của tổ chức và kinh doanh.



Hình 15: Minh họa phân cụm

5. Kết luận

Với sự trợ giúp của việc phân cụm, chúng ta có thể hiểu các thông tin khách hàng tốt hơn, trợ giúp bộ phận chăm sóc khách hàng đưa ra quyết định phù hợp. Ngoài ra, cùng với việc xác định đối tượng khách hàng, các công ty có thể đưa ra các sản phẩm và dịch vụ nhằm tới các khách hàng mục tiêu dựa trên các thông số như giới tính, tuổi tác, mô hình chi tiêu, mức thu nhập, xu hướng tiêu dùng và các yếu tố khác.

Trong bài nghiên cứu này, nhóm tác giả đã sử dụng phương pháp phân cụm dữ liệu khách hàng, ứng dụng thuật toán K-Means và phương pháp Elbow để phân khúc được các cụm dữ liệu dựa trên hành vi khách hàng, tổng giá trị đơn hàng, sở thích mua sắm của khách và đơn giá của từng loại sản phẩm. Với việc sử dụng dữ liệu đầu vào là bộ cơ sở dữ liệu gồm hơn 1.000 đơn hàng được thu thập.

Trong thuật toán K-Means tại mỗi lần thực hiện cần sử dụng lại toàn bộ dữ liệu, xác định khoảng cách giữa từng điểm dữ liệu tới các điểm trung tâm; sau đó gán từng điểm dữ liệu vào cụm có điểm trung tâm gần với điểm đó nhất. Điều này cho thấy tại mỗi vòng lặp của thuật toán cần phải sử dụng lại toàn bộ dữ liệu, nên với trường hợp dữ liệu khổng lồ sẽ không thể lưu vào bộ nhớ của máy tính vì vậy trong trường hợp này cần sử dụng bộ nhớ ngoài. Do đó với dữ liệu lớn, việc đọc lại các dữ liệu với bộ nhớ ngoài thì tốc độ các bước trong thuật toán K-Means sẽ có tốc độ chậm hơn.

Kết quả thực nghiệm cho thấy, với việc phân cụm khách hàng theo từng tiêu chí có thể giúp tổ chức kinh doanh phát triển chiến lược giới thiệu và quảng bá, chăm sóc khách hàng hiệu quả, chú trọng tới nhiều khác biệt và dễ dàng hơn. Do đó, với việc sử dụng phương pháp phân cụm dữ liệu và ứng dụng thuật toán K-Means có thể là cơ sở để gợi ý các ngành nghề mới và hữu ích hơn là mở rộng trong các ngành sản xuất và trồng trọt.

TÀI LIỆU THAM KHẢO

[1]. Nguyễn Xuân Lãn, Phạm Thị Lan Hương, Đường Thị Liên Hà (2010). *Giáo trình hành vi người tiêu dùng*. Nxb. Tài chính.

[2]. A. K. Jain (2010). *Data clustering: 50 years beyond K-Means*. Pattern recognitionletters, Vol.31, No.8, p. 651 - 666.

[3]. Chapman, C., & Feit, E. M., (2019). *R for marketing research and analytics*. New York, NY: Springer.

[4]. Daniel T. Larose (2005). *Discovering knowledge in data - An introduction to data mining*. Wiley, Hoboken, New Jersey.

[5]. M. Jordan, J. Kleinberg, B. Scholkopf (2006). *Pattern recognition and Machine learning - Part 1*.

[6]. Parul Agarwal (2011). *Issues, challenges and tools of Clustering Algorithms*. Department of Computer Science, Jamia Hamdard.

BBT nhận bài: 12/3/2024; Phản biện xong: 19/3/2024; Chấp nhận đăng: 28/3/2024