ĐỀ XUẤT MÔ HÌNH PHÂN LOẠI KHÁCH HÀNG DỰA TRÊN HOẠT ĐỘNG MUA/BÁN TRỰC TUYẾN

PROPOSE A MODEL TO CLASSIFY CUSTOMERS BASED ON ONLINE BUYING/SELLING ACTIVITIES

Mai Mạnh Trừng, Lê Thị Thu Hiền

Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

Đến Tòa soạn ngày 29/03/2022, chấp nhận đăng ngày 16/05/2022

Tóm tắt: Thế giới đang trải qua cuộc Cách mạng công nghiệp 4.0 với đặc trưng là Dữ liệu lớn (Big data), Trí tuệ nhân tạo (AI), Internet vạn vật (IoT), Điện toán đám mây (Cloud computing) và Khoa học dữ liệu (Data Science). Bài báo theo hướng khoa học dữ liệu, sử dụng thuật toán học máy nhằm xây dựng mô hình phân loại khách hàng thông qua bộ dữ liệu giao dịch thương mại điện tử, phân khúc khách hàng để phân chia cơ sở khách hàng thành nhiều nhóm cá nhân có cùng điểm giống nhau theo những cách khác nhau có liên quan đến hoạt động tiếp thị như giới tính, tuổi tác, sở thích và thói quen chi tiêu không rõ ràng. Đầu tiên, nhóm sẽ thực hiện thu thập và khai phá dữ liệu. Tiếp theo, nhóm sẽ dựa trên thuật toán học máy xây dựng mô hình phân loại khách hàng. Cuối cùng, nhóm sẽ đánh giá dữ liệu khách hàng đầu vào để có được phân loại dữ liệu vào phân khúc khách hàng.

Từ khóa: Mô hình, phân loại khách hàng, khoa học dữ liệu, K-Means.

Abstract: The world is going through the industrial revolution 4.0, characterized by Big Data, Artificial Intelligence (AI), Internet of Things (IoT), Cloud Computing, and Data Science. Data Science. The article is in the direction of data science on data management and analysis, extracting values from data to find insights, actionable knowledge, decisions that lead actions. Use machine learning to build customer classification models through e-commerce data. Customer cluster to divide the customer base into groups of individuals who are similar in different ways that are related to marketing activities such as gender, age, preferences, and unclear spending habits. The team will do data discovery first. Next, the team will import the essential packages needed for this role and then read our data. Finally, the team will evaluate the input data to get the necessary insights about it.

Keywords: Model, Customer classification, Data Science, K-Means.

1. GIỚI THIỆU

Phân loại khách hàng là quá trình phân chia khách hàng thành nhiều nhóm cá nhân có sự giống nhau ở hoạt động, thói quen mua bán dựa trên các đặc điểm như: giới tính, tuổi tác, sở thích và chi tiêu.

Để đạt được mục tiêu này các doanh nghiệp cần nghiên cứu tốt thị trường, phát triển tốt đội ngũ, nâng cao chất lượng sản phẩm, dịch vụ. Chăm sóc khách hàng tốt cũng là một trong các yếu tố mang lại thành công. Nắm bắt được yếu tố này nhiều nhà nghiên cứu và phát triển ứng dụng quan tâm việc thu thập dữ liệu khách hàng. Dựa trên tập dữ liệu khách hàng này kết hợp các thuật toán khai phá dữ

liệu để phân tích và xử lý dữ liệu đưa ra nhưng dự đoán cho tương lai. Có nhiều thuật toán về khai phá dữ liệu nhưng K-Means [1] là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất. Trong thuật toán phân cụm K-means, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau [2]. Như vậy mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm chúng ta muốn tìm, hãy chỉ ra trung tâm của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng. Giả sử thêm rằng mỗi điểm dữ liệu chỉ thuộc vào đúng một nhóm. Do vậy, với nhóm nghiên cứu muốn xây dựng mô hình phân cụm để mang lại những hiểu biết sâu sắc về phân khúc khách hàng. Trên thế giới K-means được sử dụng nhiều trong xây dựng mô hình dự báo cho các bài toán về marketing, chăm sóc khách hàng và chiến lược bán hàng [3].

Đã có một số kết quả nghiên cứu sử dụng K-means nhằm phân cụm khách hàng trong bài toán kinh doanh [4], bán hàng [5] và chăm sóc khách hàng [6]. Trong nghiên cứu này, nhóm nghiên cứu tập trung vào việc việc nhóm khách hàng thành các phân khúc sẽ dựa trên các khía cạnh về hoạt động của khách hàng trên thị trường thương mại điện tử.

2. CƠ SỞ LÝ THUYẾT

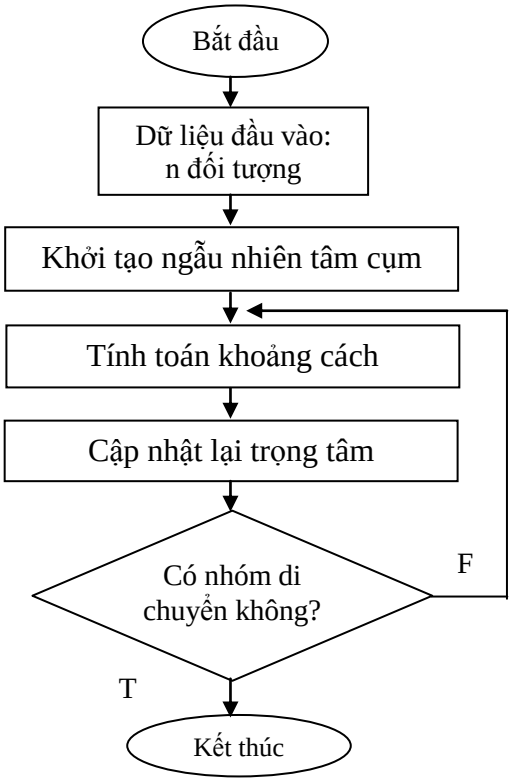
2.1. Phân cụm

Phân cụm có thể hiểu đơn giản rằng bạn có một nắm sôcôla, kẹo đường và kẹo cam thảo. Bạn được yêu cầu để tách thành các phần ăn. Bằng trực giác, bạn có thể phân biệt chúng dựa trên vẻ ngoài của chúng. Quá trình phân tách các đối tượng thành các nhóm dựa trên các đặc điểm tương ứng của chúng được gọi là phân cụm. Trong các cụm, các tính năng của các đối tượng trong một nhóm tương tự như các đối tượng khác có trong cùng một nhóm. Nó là thuật toán học không giám sát trong học máy. Điều này là do các điểm dữ liệu hiện tại không được gắn nhãn và không có ánh xạ rõ ràng giữa đầu vào và đầu ra. Như vậy, dựa trên các mẫu hiện có bên trong, quá trình phân cụm diễn ra.

2.2. Thuật toán phân cụm bằng K-Means

Thuật toán K-Means là một thuật toán lặp lại nhằm cố gắng phân vùng tập dữ liệu thành K nhóm (cụm) con không trùng lặp được xác định trước, trong đó mỗi điểm dữ liệu chỉ thuộc về một nhóm. Nó cố gắng làm cho các điểm dữ liệu trong cụm càng giống nhau càng tốt đồng thời giữ cho các cụm càng khác biệt (càng xa) càng tốt. Nó chỉ định các điểm dữ liệu cho một cụm sao cho tổng khoảng cách bình phương giữa các điểm dữ liệu và trung tâm của cụm là nhỏ nhất. Chúng ta càng có ít biến thể trong các cụm, thì các điểm dữ liệu trong cùng một cụm càng đồng nhất. Sau đó, chúng tôi tiến hành thực hiện phân cụm K-means sẽ tạo ra các nhóm khác nhau để nhóm hoạt động chi tiêu tương tự dựa trên độ tuổi và thu nhập hàng năm của họ. Phân cụm K-Means chọn các giá trị ngẫu nhiên từ dữ liệu và tạo thành các cụm được chỉ định. Các giá trị gần nhất từ tâm của mỗi cụm được lấy để cập nhật cụm và định hình lại biểu đồ (giống như k-NN). Các giá trị gần nhất dựa trên khoảng cách Euclide.

Thuật toán K-Means lặp đi lặp lại quá trình phân các ví dụ vào cụm có tâm gần nhất, sau đó là điều chỉnh tâm cụm, cho tới khi điều kiện hội tụ được thỏa mãn. Cụ thể hơn, thuật toán được biểu diễn thông qua lưu đồ thuật toán như sau:



Hình 1. Lưu đồ thuật toán K- Means

Thuật toán k-means bao gồm các bước cơ bản sau:

Đầu vào: **Số cụm k và hàm E**

$$E = \sum_{1}^{k} \sum_{x \in c_i} \left| x - m_i \right|^2$$

Đầu ra: Các cụm C[i] $(1 \le i \le k)$ và hàm tiêu chuẩn E đạt giá trị tối thiểu.

Bắt đầu

Bước 1: Khởi tạo

Chọn ngẫu nhiên k tâm ban đầu trong không gian R_d (d là số chiều của dữ liệu). Mỗi cụm được đại diện bằng các tâm của cụm.

Bước 2: Tính toán khoảng cách

$$D_{j=1}^{k} = \sqrt{\sum_{i=1}^{n} \left(x_i - m_j \right)^2}$$

Đối với mỗi điểm x_i $(1 \le i \le n)$ tính toán khoảng cách của nó tới mỗi trọng tâm m_j $(1 \le j \le k)$ Sau đó tìm trọng tâm gần nhất đối với mỗi điểm và nhóm chúng vào các nhóm gần nhất. Bước 3: Cập nhật lại trọng tâm.

Đối với mỗi $1 \le j \le k$, cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng các vectơ đối tượng dữ liệu, gán lại điểm gần trung tâm nhóm mới.

Điều kiện dừng:

Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

Kết thúc.

3. MÔ HÌNH VÀ DỮ LIỆU HUẤN LUYỆN

3.1. Mô tả tập dữ liệu

Dữ liệu phân khúc khách hàng của trung tâm mua sắm, dữ liệu được cung cấp bởi Phòng thí nghiệm dữ liệu Exposys (bộ dữ liệu chứa khoảng 54.000 bản ghi). Nó có các ID khách hàng duy nhất riêng lẻ, một biến phân loại ở dạng Giới tính và ba cột Tuổi, Thu nhập hàng năm và Điểm chi tiêu sẽ là các mục tiêu chính của chúng tôi để xác định các kiểu mua sắm và chi tiêu của khách hàng.

Bảng 1. Dữ liệu phân khúc khách hàng

ID	*Giới tính*	*Tuổi*	*Thu nhập (k$)*	*Điểm chi tiêu (1-100)*
1	Nam	19	15	39
2	Nam	21	15	81
3	Nữ	20	16	6
4	Nữ	23	16	77
5	Nữ	31	17	40
6	Nữ	22	17	76

ID	*Giới tính*	*Tuổi*	*Thu nhập (k$)*	*Điểm chi tiêu (1-100)*
7	Nữ	35	18	6
8	Nữ	23	18	94
…	…	…	…	…

3.2. Phân tích dữ liệu

Công cụ cài đặt thực nghiệm:

- Jupyter Notebook được sử dụng là nền tảng thực thi mã lệnh.
- Ngôn ngữ lập trình Python với các thư việc:

Pandas 1.1.5

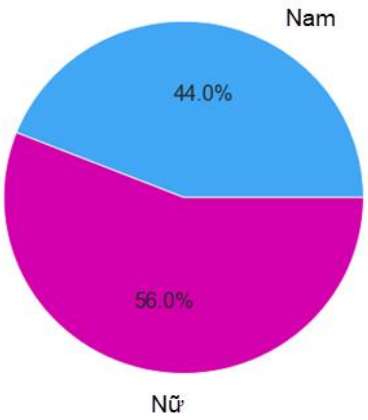
Numpy 1.19.2

Matplotlib 3.3.2

Scikit Learn 0.23.2

Seaborn 0.11.1

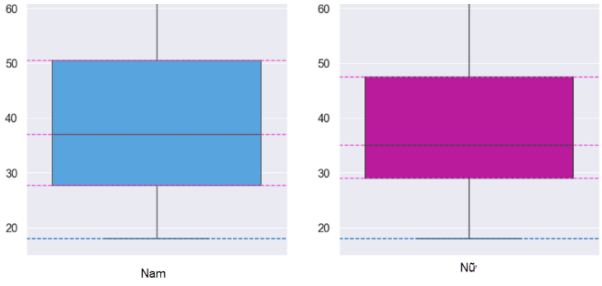
- Phân bố giới tính nam và nữ:



Hình 2. Phân bố giới tính nam và nữ

Từ biểu đồ trên, chúng ta quan sát thấy số lượng nữ (112) nhiều hơn nam (88). Tỷ lệ dân số theo giới tính là 56% nữ và 44% nam. Bằng cách này, chúng ta có thể tạm giả định rằng đa số khách hàng đến thăm trung tâm mua sắm là nữ giới.

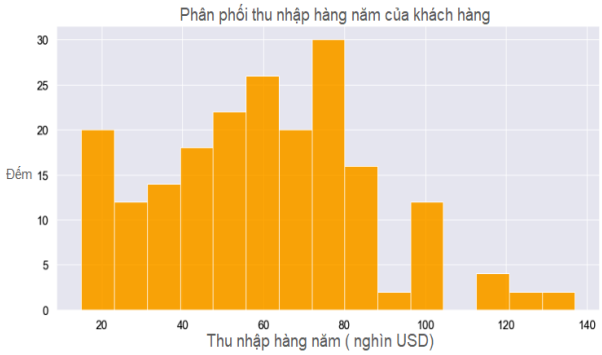
- Phân tích độ tuổi của khách hàng



Hình 3. Phân tích theo độ tuổi khách hàng

Từ hình 3 ở trên, chúng ta có thể kết luận rằng một lượng lớn độ tuổi nằm trong khoảng từ 30 đến 35. Độ tuổi tối thiểu là 18, độ tuổi tối đa là 70. Bằng cách so sánh phân bố độ tuổi của khách hàng, chúng tôi có thể kết luận rằng hầu hết khách hàng đều nằm trong trong khoảng từ 30 đến 50, trong đó trung bình là khoảng 35 tuổi.

- Phân tích điểm thu nhập và chi tiêu hàng năm

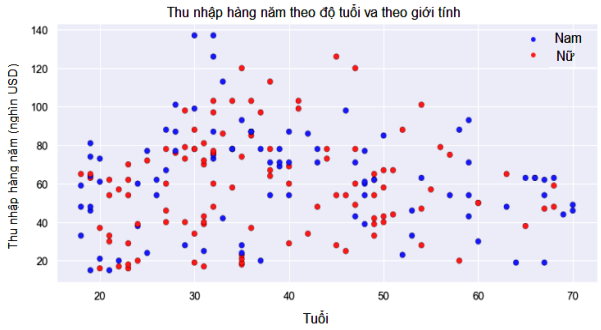


Hình 4. Mối quan hệ thu nhập và chi tiêu (giá trị ở cột đếm được nhân với tỷ lệ 270 lần)

Phân phối Thu nhập cuối kỳ và Điểm chi tiêu thể hiện sự xấp xỉ của phân phối chuẩn, với mật độ cao nhất xung quanh giá trị trung bình của các biến. Thu nhập hàng năm tối đa và tối thiểu lần lượt là 137 và 15, với mức trung bình là 60, 56. Từ đó, chúng ta có thể thấy rằng đỉnh của phân phối đã giảm trong vùng từ 60 đến 75. Đối với điểm Chi tiêu, tối đa và tối thiểu là 99 và 1, trong khi biểu đồ 10 chỉ ra rằng số lượng khách hàng cao nhất có điểm chi tiêu nằm trong khoảng từ 40 đến 60.

- Quan hệ đặc trưng: Phân tích thu nhập

hàng năm so với tuổi và giới tính.

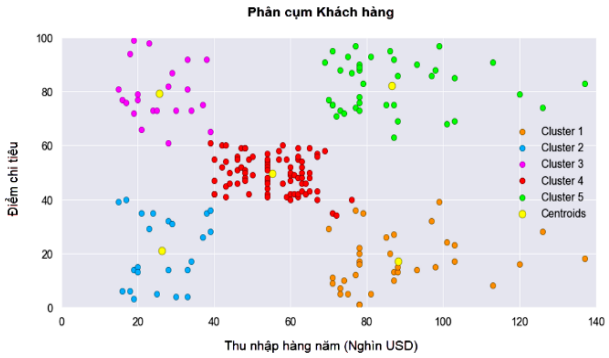


Hình 5. Phân tích thu nhập so với tuổi và giới tính

Như hình 5 thì thu nhập cao phần đa là nam giới, tuổi cao thì cũng phần đa giới tính là nam có thu nhập cao.

3.3. Phân tích phân cụm (cluster)

Các cụm sau được tạo bởi mô hình: cụm màu cam, cụm màu xanh than, cụm màu tím, cụm màu đỏ, cụm màu xanh lá cây.



Hình 6. Phân cụm khách hàng

- Cụm màu cam - khách hàng "cân bằng":

Họ kiếm được ít hơn và chi tiêu ít hơn. Chúng ta có thể thấy những người có thu nhập hàng năm thấp và điểm số chi tiêu thấp, điều này khá hợp lý vì những người có mức lương thấp thích mua sắm ít hơn, trên thực tế, đây là những người biết cách chi tiêu và tiết kiệm. Các cửa hàng / trung tâm mua sắm sẽ ít quan tâm nhất đến những người thuộc nhóm này.

- Cụm màu xanh than - khách hàng chi tiêu ít:

Thu nhập cao và chi tiêu ít hơn. Chúng tôi thấy rằng mọi người có thu nhập cao nhưng điểm chi tiêu thấp, điều này thật thú vị. Có thể đây là những người không hài lòng hoặc không đủ khả năng tài chính với các dịch vụ của trung tâm mua sắm. Đây có thể là những mục tiêu chính của trung tâm mua sắm, vì chúng có khả năng tiêu tiền. Vì vậy, các nhà chức trách trung tâm mua sắm sẽ cố gắng bổ sung các cơ sở mới để có thể thu hút những người này và có thể đáp ứng nhu cầu của họ.

- Cụm màu tím – khách hàng thông thường:

Khách hàng ở mức trung bình về thu nhập và chi tiêu. Một người tiêu dùng ở hạn mức trung bình về tỷ lệ chi tiêu với thu nhập hàng năm, chúng tôi thấy rằng đây là những người có thu nhập trung bình và điểm chi tiêu trung bình, những người này cũng không phải là mục tiêu chính của các cửa hàng hoặc trung tâm mua sắm.

- Cụm màu đỏ - khách hàng "chi tiêu nhiều":

Loại khách hàng này thu nhập ít hơn nhưng chi tiêu cao, vì vậy cũng có thể được coi là khách hàng mục tiêu tiềm năng, chúng ta có thể thấy rằng những người có thu nhập thấp nhưng điểm chi tiêu cao hơn, đây là những người vì một lý do nào đó thích mua hàng sản phẩm thường xuyên hơn mặc dù họ có thu nhập thấp. Có thể là do những người này hài lòng hơn với các dịch vụ của trung tâm mua sắm. Các cửa hàng / trung tâm thương mại có thể không nhắm mục tiêu đến những người này một cách hiệu quả nhưng vẫn sẽ không đánh mất họ.

- Cụm màu xanh lá cây – khách hàng mục tiêu:

Phân khúc khách hàng có thu nhập cao và cũng chi tiêu cao, đây chính là khách hàng mục tiêu cao. Họ có thu nhập hàng năm cao cũng như điểm chi tiêu cao. Chúng tôi thấy rằng mọi người có thu nhập cao và điểm chi

tiêu cao, đây là trường hợp lý tưởng cho trung tâm mua sắm hoặc cửa hàng vì những người này là nguồn lợi nhuận chính. Những người này có thể là khách hàng thường xuyên của trung tâm mua sắm và bị thuyết phục bởi cơ sở vật chất của trung tâm mua sắm.

4. ĐÁNH GIÁ KẾT QUẢ

Về bài toán phân lớp khách hàng tiềm năng thì năm 2015 Md Saif Ali, một kỹ sư học máy nghiên cứu tại Altimetrik India Pvt Ltd đã sử dụng thuật toán K-Means và bộ dữ liệu các giao dịch trực tuyến của khách hàng tại trung tâm thương mại từ 01/12/2010 đến 09/12/2011 nhằm đưa ra mô hình phân cụm khách hàng [7]. Ông cũng đưa ra phân cụm với 5 nhóm khách hàng, nhưng phân cụm khách hàng của ông đưa ra các kết quả về thói quen sản phẩm mua sắm, cụ thể trong 5 cụm của ông thì 1 cụm chứa khách hàng chủ yếu mua sản phẩm liên quan trang trí và quà tặng, 1 cụm liên quan khách hàng mua sản phẩm là các mặt hàng xa xỉ, các sản phẩm cổ điển, lưu niệm thì xuất hiện ở mọi phân cụm khách hàng. Còn đối với nghiên cứu của chúng tôi đã khám phá năm phân khúc dựa trên Thu nhập hàng năm và Điểm chi tiêu của khách hàng, được cho là những yếu tố / thuộc tính tốt nhất để xác định các phân khúc của khách hàng trong trung tâm mua sắm. Chúng bao gồm: Chốt lại khách hàng tiềm năng, khách hàng cân bằng, khách hàng mục tiêu, người lớn và khách hàng bình thường. Chúng tôi có thể đưa khách hàng mục tiêu vào một số hệ thống cảnh báo nơi có thể gửi tin nhắn SMS và email cho họ hàng ngày về các ưu đãi và giảm giá mà họ có thể nhận được tại trung tâm mua sắm; trong khi phần còn lại, chúng tôi có thể đặt một lần mỗi tuần trong một tháng cho các tin nhắn SMS bùng nổ để thông báo cho họ về sản phẩm của chúng tôi. Tương tự, bây giờ chúng tôi biết hành vi của khách hàng tùy thuộc vào thu nhập hàng năm và điểm chi tiêu của họ. Có thể có nhiều chiến lược tiếp thị được áp dụng cho khách hàng trên phân tích cụm này. Khách hàng có thu nhập cao và có điểm chi tiêu cao là khách hàng mục tiêu của chúng tôi và chúng tôi luôn muốn giữ chân họ vì họ mang lại tỷ suất lợi nhuận cao nhất cho tổ chức của chúng tôi. Điểm thu nhập cao và ít chi tiêu hơn những khách hàng có thể bị thu hút với nhiều loại sản phẩm phù hợp với nhu cầu phong cách sống của họ và điều đó có thể thu hút họ đến với siêu thị mall. Thu nhập ít hơn điểm chi tiêu ít hơn có thể được cung cấp thêm các ưu đãi và liên tục gửi cho họ các ưu đãi và giảm giá sẽ thu hút họ tham gia chi tiêu. Chúng tôi cũng có thể thực hiện phân tích cụm về loại sản phẩm mà khách hàng có xu hướng mua và có thể đưa ra các chiến lược tiếp thị khác cho phù hợp. Tập dữ liệu không có đủ dữ liệu để thực hiện nhiều phân tích hơn trên cùng một tập dữ liệu.

5. KẾT LUẬN

Các công ty, trung tâm thương mại, siêu thị trên doanh nghiệp kinh doanh nhỏ nên thực hiện phân tích giỏ thị trường cho doanh nghiệp của họ. Điều này sẽ cho phép các công ty nhắm mục tiêu các nhóm khách hàng cụ thể, mô hình phân khúc khách hàng cho phép phân bổ hiệu quả các nguồn lực tiếp thị và tối đa hóa cơ hội bán chéo và bán thêm. Khi một nhóm khách hàng được gửi thông điệp được cá nhân hóa như một phần của hỗn hợp tiếp thị được thiết kế theo nhu cầu của họ, thì các công ty sẽ dễ dàng gửi cho những khách hàng đó những ưu đãi đặc biệt nhằm khuyến khích họ mua nhiều sản phẩm hơn. Phân khúc khách hàng cũng có thể cải thiện dịch vụ khách hàng và hỗ trợ sự trung thành và duy trì khách hàng. Các tài liệu tiếp thị, tiếp cận khách hàng được gửi đi bằng cách sử dụng kết quả phân khúc khách hàng sẽ có xu hướng được khách

hàng đón nhận và đánh giá cao hơn thay vì các thông điệp mạo danh thương hiệu, không ghi nhận lịch sử mua hàng hoặc bất kỳ loại quan hệ khách hàng nào. Cuối cùng với phân khúc khách hàng các công ty sẽ luôn đi trước các đối thủ cạnh tranh trong các khu vực cụ thể của thị trường và xác định các sản phẩm mới tồn tại hoặc khách hàng tiềm năng có thể quan tâm hoặc cải tiến sản phẩm để đáp ứng kỳ vọng của khách hàng.

TÀI LIỆU THAM KHẢO

[1] Bhatia, A., Jain, V. Sharma, Y. and Arora, V "*Crime Prediction using K-means Algorithm*" GRD Journals- Global Research and Development Journal for Engineering 2 (5) pp. 206-208 GRDJ (2018).

[2] Hafiz, B., Mousa, M. and Waheed, M. *"Fast and Efficient K-means based Algorithm to Contentbased Image Clustering*" International Journal of Computer Applications 152 (5) pp.8-12 Semanticscholar (2016).

[3] Ansari, A. & Riasi, A. "*Taxonomy of marketing strategies using bank customers' clustering*", International Journal of Business and Management, Vol. 11, No. 7, 106-119 (2016).

[4] Bolton, R.N., Kannan, P.K. & Bramlett, M.D, "*Implications of loyalty program membership and service experiences for customer retention and value",* Journal of the Academy of Marketing Science, 28(1), 95–108, 2000.

[5] Bowen, J.T. and Chen, S, "*The relationship between customer loyalty and customer satisfaction*", International Journal of Contemporary Hospitality Management, 13(5), 213-217, 30 Aug 2021.

[6] Bulut, Z.A, *"Determinants of repurchase intention in online shopping: a Turkish consumers' perspective",* International Journal of Business and Social Science, Vol. 6, No. 10, 55-63 (2020).

[7] Md Saif Ali, *"Customer Segmentation for Retail Shop",* Altimetrik India Pvt Ltd, Bangalore, 2015.

Thông tin liên hệ: **Mai Mạnh Trừng**

Điện thoại: 09123.55.022 - Email: mmtrung@uneti.edu.vn

Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.