

## Tối ưu hóa phân khúc khách hàng bán lẻ trực tuyến:

### Áp dụng các thuật toán phân cụm dựa trên mô hình RFM và các chỉ số đánh giá đa chiều

NGUYỄN TỊNH GIANG<sup>a</sup>, HOÀNG ANH<sup>\*,a</sup>

Đại học Kinh tế Thành phố Hồ Chí Minh

| THÔNG TIN                                                                                                                                                                                                                                                                                                                                                                                                                         | TÓM TẮT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Ngày nhận: 08/06/2025<br/>Ngày nhận lại: 20/08/2025<br/>Duyệt đăng: 30/08/2025</p> <p><b>Mã phân loại JEL:</b><br/>M31; L81; C38; C52; C55.</p> <p><b>Từ khóa:</b><br/>Phân khúc khách hàng;<br/>Bán lẻ trực tuyến;<br/>Tiếp thị cá nhân hóa;<br/>Khai phá dữ liệu;<br/>Mô hình RFM;<br/>Phân tích cụm.</p> <p><b>Keywords:</b><br/>Customer segmentation;<br/>Online retail;<br/>Personalized marketing;<br/>Data mining;</p> | <p>Nghiên cứu này sử dụng các kỹ thuật khai thác dữ liệu giao dịch để phân khúc khách hàng trong lĩnh vực bán lẻ trực tuyến. Tích hợp khuôn khổ CRISP-DM với mô hình RFM, nghiên cứu đánh giá sáu thuật toán phân cụm bao gồm K-Means, Spectral Clustering, Mean Shift, Gaussian Mixture Model, DBSCAN và Fuzzy C-Means nhằm tối ưu hóa giải pháp phân khúc khách hàng. Các chỉ số định lượng bao gồm Davies-Bouldin, Calinski-Harabasz, Dunn và điểm Silhouette, chứng minh K-Means đạt hiệu quả tốt nhất trên dữ liệu thực nghiệm. Kết quả tối ưu hóa bốn phân khúc khách hàng riêng biệt: khách hàng tốt nhất, khách hàng trung thành, khách hàng tiềm năng và khách hàng rời bỏ dịch vụ. Phát hiện này cung cấp bằng chứng ra quyết định dựa trên dữ liệu, thiết lập và thực hiện các chiến lược tiếp thị có mục tiêu thông qua tích hợp hệ thống CRM, tối ưu hóa phân bổ nguồn lực và giữ chân khách hàng. Đáng chú ý, các phân khúc khách hàng được xác định tạo điều kiện cá nhân hóa, tăng cường tiềm năng doanh thu và đạt lợi thế cạnh tranh. Nghiên cứu cũng chỉ rõ những hạn chế và đề xuất các hướng phát triển trong tương lai, chẳng hạn như kết hợp dữ liệu nhân khẩu học và các kỹ thuật học sâu để nâng cao chất lượng phân khúc khách hàng.</p> <p><b>Abstract</b></p> <p>This study utilizes transactional data mining techniques to segment customers in the online retail sector. Integrating the CRISP-DM framework with the RFM model, the study evaluates six clustering</p> |

\* Tác giả liên hệ.

Biên tập viên: Nguyễn Thị Hồng Cẩm.

Email: [giangnguyen.524102060872@st.ueh.edu.vn](mailto:giangnguyen.524102060872@st.ueh.edu.vn) (Nguyễn Tịnh Giang), [anh@ueh.edu.vn](mailto:anh@ueh.edu.vn) (Hoàng Anh).

Trích dẫn bài viết: Nguyễn Tịnh Giang, & Hoàng Anh. (2025). Tối ưu hóa phân khúc khách hàng bán lẻ trực tuyến: Áp dụng các thuật toán phân cụm dựa trên mô hình RFM và các chỉ số đánh giá đa chiều. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á*, 36(7), 04-21. <https://doi.org/10.24311/jabes/2025.36.7.01>

|                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| RFM Model;<br>Cluster analysis. | algorithms including K-Means, Spectral Clustering, Mean Shift, Gaussian Mixture Model, DBSCAN, and Fuzzy C-Means to optimize the customer segmentation solution. Quantitative metrics including Davies-Bouldin, Calinski-Harabasz, Dunn, and Silhouette scores demonstrate that K-Means performs best on the experimental data. The results optimize four distinct customer segments: best customers, loyal customers, potential customers, and churners. The findings provide evidence for data-driven decision making, establishing and implementing targeted marketing strategies through CRM system integration, optimizing resource allocation, and customer retention. Notably, the identified customer segments facilitate personalization, increase revenue potential, and gain competitive advantage. The study also highlights limitations and suggests future directions, such as combining demographic data and deep learning techniques to improve customer segmentation. |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 1. Giới thiệu

Trong bối cảnh ngành bán lẻ trực tuyến ngày càng phát triển mạnh mẽ, việc hiểu và phân tích hành vi khách hàng trở thành yêu cầu thiết yếu để doanh nghiệp đưa ra các chiến lược marketing hiệu quả, cá nhân hóa trải nghiệm và tối ưu hóa nguồn lực. Phân tích hành vi khách hàng, với mục tiêu chia nhóm khách hàng theo đặc điểm tương đồng về nhân khẩu học và hành vi, đóng vai trò quan trọng trong việc gia tăng giá trị vòng đời khách hàng và nâng cao hiệu quả kinh doanh. Tuy nhiên, với sự gia tăng nhanh chóng về khối lượng, tốc độ, sự đa dạng và tính chính xác của dữ liệu, các phương pháp phân tích hành vi khách hàng truyền thống bộc lộ nhiều điểm hạn chế, đòi hỏi nghiên cứu phát triển và ứng dụng các kỹ thuật khai phá dữ liệu tiên tiến.

Nghiên cứu này thực hiện với quy trình CRISP-DM chuẩn mực (Chapman và cộng sự, 2000), đảm bảo tính hệ thống trong khai phá dữ liệu. Dữ liệu thực nghiệm Online Retail từ nguồn UCI Machine Learning Repository (Chen, 2015), lưu giữ thông tin giao dịch của một doanh nghiệp bán lẻ trực tuyến tại Vương quốc Anh. Dữ liệu được thu thập, làm sạch và chuyển đổi nhằm xử lý giá trị thiếu, trùng lặp và bất thường. Mô hình RFM sau đó được xây dựng và dữ liệu được chuẩn hóa phản ánh hành vi khách hàng qua ba yếu tố: Recency, Frequency và Monetary. Sáu thuật toán phân cụm được thực nghiệm, với kết quả được đánh giá cả về mặt kỹ thuật (các chỉ số đo lường sự phù hợp mô hình) và mục tiêu kinh doanh. Cuối cùng, các nhóm khách hàng riêng biệt được phân tích và gắn với các đề xuất chiến lược marketing cụ thể, hỗ trợ triển khai thực tế trong hệ thống quản lý quan hệ khách hàng.

Nghiên cứu này mang lại những đóng góp quan trọng trong lĩnh vực phân khúc khách hàng bán lẻ trực tuyến, được cấu trúc rõ ràng qua hai khía cạnh chính: lý thuyết và ứng dụng thực tiễn. (1) Về mặt lý thuyết: Nghiên cứu thực hiện một phân tích so sánh toàn diện giữa sáu thuật toán phân cụm (K-Means, Spectral Clustering, Mean Shift, Gaussian Mixture Model, DBSCAN và Fuzzy C-Means) dựa trên mô hình RFM, được tích hợp chặt chẽ trong quy trình CRISP-DM. Thông qua việc áp dụng các chỉ số đánh giá định lượng đa chiều, nghiên cứu xác định K-Means là thuật toán tối ưu, khắc phục hạn chế của các nghiên cứu trước đây, vốn thường chỉ tập trung vào một hoặc hai thuật toán mà thiếu sự so sánh đa dạng trên dữ liệu giao dịch thực tế từ nguồn UCI Machine Learning Repository. Điều

này làm phong phú thêm cơ sở lý thuyết về khai phá dữ liệu, cung cấp nền tảng khoa học cho các nghiên cứu tiếp theo, như mở rộng mô hình với dữ liệu đa nguồn hoặc tích hợp kỹ thuật học sâu. (2) Về mặt ứng dụng: Nghiên cứu đề xuất bốn phân khúc khách hàng có ý nghĩa kinh doanh cao (Best Customers, Loyal Customers, Potential Customers và Lost Customers), được xây dựng từ dữ liệu thực nghiệm với hơn 4.000 khách hàng. Các phân khúc này hỗ trợ tích hợp vào hệ thống CRM, tạo cơ sở cho các chiến lược marketing cá nhân hóa, tối ưu hóa nguồn lực, và nâng cao khả năng giữ chân khách hàng, từ đó mang lại lợi thế cạnh tranh cho các doanh nghiệp bán lẻ trực tuyến trong bối cảnh chuyển đổi số toàn cầu và tại Việt Nam.

## 2. Cơ sở lý thuyết và các nghiên cứu liên quan

### 2.1. Mô hình RFM, thuật toán phân cụm, và khuôn khổ CRISP-DM

Phân khúc khách hàng trong bán lẻ trực tuyến là chủ đề được quan tâm trong hơn hai thập kỷ qua, với nhiều nghiên cứu ứng dụng mô hình RFM và các thuật toán phân cụm để khám phá hành vi khách hàng. Hughes (1994, 2005) là người đầu tiên giới thiệu mô hình RFM để xác định giá trị và phân chia khách hàng thành các nhóm có ý nghĩa. MacQueen (1967) giới thiệu thuật toán K-Means dựa trên khoảng cách Euclidean, mở đường cho nhiều nghiên cứu phân cụm khách hàng. Kotler (1967, 2012) nhấn mạnh vai trò của phân khúc khách hàng trong quản trị marketing. Fayyad và cộng sự (1996) đề xuất quy trình KDD, tiền đề của CRISP-DM, thiết lập khung chuẩn cho khai phá dữ liệu từ tiền xử lý đến phân tích kết quả.

Mô hình RFM (Recency, Frequency, Monetary) là một phương pháp định lượng hành vi khách hàng dựa trên: thời gian kể từ lần mua gần nhất (Recency), tần suất mua hàng (Frequency), và tổng số tiền chi tiêu (Monetary). Theo Hughes (1994, 2005), phân tích RFM là một kỹ thuật tiếp thị dùng để đánh giá và phân nhóm khách hàng một cách định lượng. Kumar và Reinartz (2012) cũng đưa ra định nghĩa tương tự, là một mô hình phân khúc khách hàng dựa trên dữ liệu, đánh giá giá trị của khách hàng dựa trên RFM. Trong nghiên cứu này, mô hình RFM được sử dụng kết hợp với các thuật toán phân cụm để tạo ra các phân khúc khách hàng có ý nghĩa phục vụ mục tiêu kinh doanh.

Phân khúc khách hàng dựa trên nền tảng phân khúc thị trường là việc chia thị trường thành các nhóm người mua riêng biệt có những nhu cầu, đặc điểm hoặc hành vi khác nhau, và có thể đòi hỏi các sản phẩm hoặc chương trình marketing riêng biệt (Kotler & Armstrong, 2017). Thuật toán phân cụm giúp giải quyết vấn đề này. Phân tích phân cụm là quá trình tổ chức một tập hợp các mẫu (thường được biểu diễn dưới dạng vector các giá trị đo lường, hoặc một điểm trong không gian đa chiều) thành các cụm dựa trên mức độ tương đồng (Jain và cộng sự, 1999).

Quy trình CRISP-DM (Cross-Industry Standard Process for Data Mining) (Chapman và cộng sự, 2000) là một quy trình chuẩn gồm 6 bước thực hiện khai phá dữ liệu một cách có hệ thống và hiệu quả. Sáu bước gồm: Hiểu nghiệp vụ, hiểu dữ liệu, chuẩn bị dữ liệu, mô hình hóa, đánh giá và triển khai. Quy trình CRISP-DM giúp đảm bảo quá trình khai phá dữ liệu diễn ra có định hướng, khoa học và phù hợp với mục tiêu thực tế.

## 2.2. Các nghiên cứu liên quan

Nghiên cứu của Cheng và Chen (2009) là một trong những nghiên cứu tiên phong kết hợp mô hình RFM và lý thuyết Rough Set nhằm phân cụm khách hàng một cách chính xác hơn. Nghiên cứu cho thấy việc tích hợp hai phương pháp giúp cải thiện kết quả phân cụm nhờ khả năng xử lý dữ liệu không chắc chắn và không đầy đủ. Chen và cộng sự (2012) đã vận dụng mô hình RFM kết hợp với khai phá dữ liệu để phân khúc khách hàng trong ngành bán lẻ trực tuyến. Kết quả cho thấy RFM là công cụ mạnh để cá nhân hóa marketing và giữ chân khách hàng. Cho và cộng sự (2013) tiếp tục mở rộng RFM bằng cách thêm yếu tố ưu tiên sản phẩm để nâng cao khả năng gợi ý trong môi trường thương mại điện tử, từ đó hỗ trợ các hệ thống khuyến nghị hoạt động hiệu quả hơn.

Zahrotun (2017) đề xuất sử dụng Fuzzy C-Means thay vì K-Means để thực hiện phân cụm khách hàng với hệ thống CRM trong thương mại điện tử, giúp linh hoạt hơn trong việc xác định khách hàng thuộc nhiều nhóm khác nhau. Rizki và cộng sự (2020) đã áp dụng RFM kết hợp với K-Means để phân khúc khách hàng trên hệ thống POS (Point of Sale). Nghiên cứu nhấn mạnh tính thực tiễn của RFM trong môi trường bán lẻ vật lý và cho thấy khả năng phát hiện nhóm khách hàng trung thành. Christy và cộng sự (2021) đề xuất phương pháp xếp hạng RFM nhằm phân biệt vai trò của khách hàng trước khi thực hiện phân cụm. Cách tiếp cận này giúp tăng độ chính xác và hiệu quả của quá trình phân cụm.

Anitha và Patil (2022) tiếp tục sử dụng RFM và K-Means để nghiên cứu hành vi mua sắm, khẳng định rằng phân cụm khách hàng giúp cá nhân hóa chiến lược marketing. Ibrahim và Tyasnurita (2022) mở rộng RFM thành LRFM (thêm yếu tố thời gian L – Length) và cho thấy kết quả phân cụm tốt hơn trong phân tích vòng đời khách hàng. Li và cộng sự (2022) đề xuất mô hình RFM có trọng số kết hợp K-Means để tối ưu hóa hiệu quả phân khúc, có tính ứng dụng cao trong thương mại điện tử. Mensouri và cộng sự (2022) nâng cấp mô hình lên RFMT (thêm yếu tố T – Time) và phân tích thêm điểm hài lòng, cho thấy cách tiếp cận này mang lại bức tranh toàn diện về hành vi khách hàng.

Các nghiên cứu gần đây đã mở rộng mô hình RFM nhằm nâng cao hiệu quả phân khúc khách hàng trong thương mại điện tử. Chẳng hạn, Lang và cộng sự (2023) đề xuất mô hình RFM có trọng số động, kết hợp phương pháp phân tích thứ bậc (AHP) và Entropy để xác định trọng số của các thành phần R - F - M một cách đồng thời khách quan và chủ quan, qua đó cải thiện đáng kể hiệu quả phân cụm (Wong và cộng sự, 2024) áp dụng RFM kết hợp phân cụm phân cấp, đạt chỉ số Silhouette 0,47 và Calinski-Harabasz xấp xỉ 3.787, kết quả vượt trội so với phương pháp K-Means truyền thống. Trong khi đó, Wang (2025) xây dựng khung phân đoạn khách hàng tích hợp Q-learning và thuật toán tiến hóa vi sai với K-Means, giúp cải thiện khả năng xác định nhóm khách hàng trong bối cảnh marketing số. Bên cạnh đó, tổng quan của Alves Gomes và Meisen (2023) cũng khẳng định xu hướng mở rộng RFM và các phương pháp phân cụm trong thương mại điện tử hiện đại.

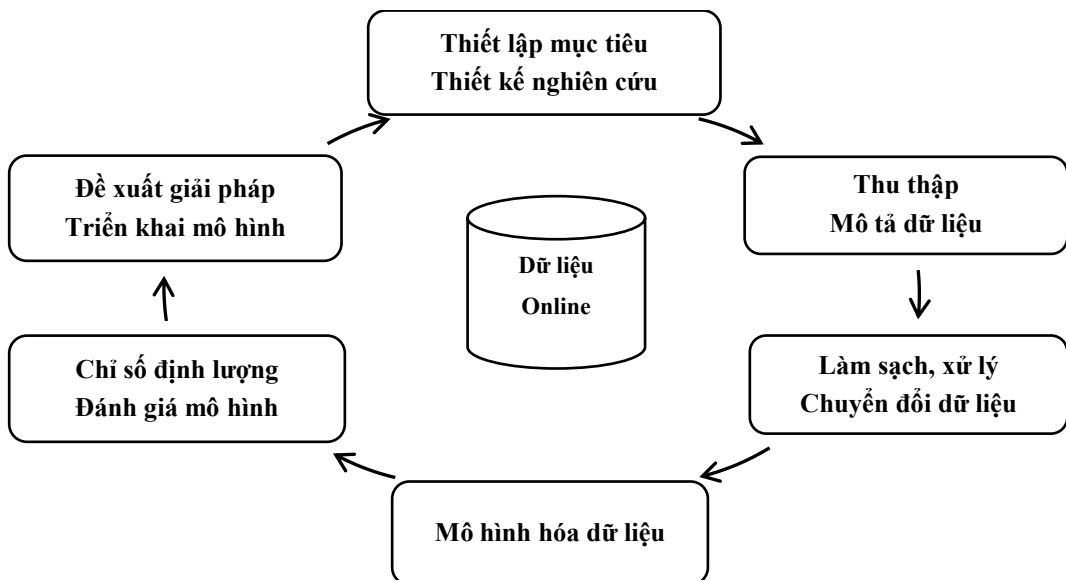
So sánh với các nghiên cứu liên quan, có thể thấy rằng hầu hết các công trình trước đây thường giới hạn ở việc áp dụng một hoặc hai thuật toán phân cụm kinh điển kết hợp với mô hình RFM, dẫn đến thiếu hụt về sự so sánh toàn diện và đánh giá đa chiều. Chẳng hạn, Anitha và Patil (2022) chỉ tập trung vào thuật toán K-Means với hai chỉ số đánh giá cơ bản (Silhouette và Davies-Bouldin), mà không mở rộng so sánh với các thuật toán khác, từ đó tạo ra khoảng trống về việc xác định thuật toán tối ưu trong bối cảnh dữ liệu giao dịch phức tạp; tương tự, Chen và cộng sự (2012) chủ yếu sử dụng K-Means và Fuzzy C-Means (FCM) nhưng chỉ dựa vào chỉ số Silhouette đơn lẻ, thiếu các chỉ số bổ sung để đánh giá độ tách biệt và đồng nhất của các cụm một cách toàn diện. Các nghiên cứu khác như

Cheng và Chen (2009) kết hợp RFM với Rough Set nhưng không so sánh thuật toán phân cụm đa dạng, Zahrotun (2017) ưu tiên FCM cho CRM trong thương mại điện tử mà chưa đánh giá với các thuật toán khác, Rizki và cộng sự (2020) và Li và cộng sự (2022) cũng chỉ sử dụng K-Means với RFM, tập trung vào phân khúc khách hàng nhưng thiếu sự so sánh rộng rãi, trong khi Ibrahim và Tyasnurita (2022) mở rộng thành LRFM và Mensouri và cộng sự (2022) thành RFMT, nhưng vẫn giới hạn ở K-Means hoặc FCM mà không đánh giá đa thuật toán trên cùng bộ dữ liệu. Bài báo này lấp đầy khoảng trống đó bằng cách thực hiện một phân tích so sánh toàn diện sáu thuật toán phân cụm (K-Means, Spectral Clustering, Mean Shift, Gaussian Mixture Model - GMM, DBSCAN và Fuzzy C-Means - FCM) kết hợp với bốn chỉ số đánh giá định lượng đa chiều (Silhouette, Davies-Bouldin, Calinski-Harabasz và Dunn), tất cả được áp dụng trên cùng bộ dữ liệu RFM từ nguồn Online Retail (UCI Machine Learning Repository), đảm bảo tính nhất quán và khả năng so sánh trực tiếp. Không giống các nghiên cứu trước thường áp dụng riêng lẻ các thuật toán kinh điển, bài báo này kết hợp sáu thuật toán một cách có hệ thống trong quy trình CRISP-DM để xác định thuật toán tối ưu (K-Means), từ đó tạo cơ sở lý thuyết mới mẻ và vững chắc hơn cho việc phân khúc khách hàng trong lĩnh vực bán lẻ trực tuyến, hỗ trợ các ứng dụng thực tiễn như cá nhân hóa marketing và tối ưu hóa CRM.

### 3. Phương pháp nghiên cứu

#### 3.1. Thiết lập mục tiêu và thiết kế nghiên cứu

Ba nhóm câu hỏi chính được xác định trong nghiên cứu này, bao gồm: (1) Có thể chia khách hàng thành những phân khúc nào dựa trên dữ liệu giao dịch? (2) Xây dựng chiến lược marketing phù hợp với từng nhóm như thế nào? (3) Xác định thuật toán phân cụm tối ưu trên tập dữ liệu thực nghiệm? Nghiên cứu này được thiết kế theo quy trình CRISP-DM với 6 bước đề xuất như Hình 1 nhằm tìm câu trả lời.



Hình 1. Quy trình thực hiện nghiên cứu

### 3.2. Thu thập và mô tả dữ liệu

Tập dữ liệu Online Retail (Chen, 2015) được thu thập từ một nhà bán lẻ trực tuyến tại Vương quốc Anh, gồm 541.909 giao dịch từ ngày 01/12/2010 đến 09/12/2011. Dữ liệu này được lấy từ UCI Machine Learning Repository, là nơi cung cấp dữ liệu uy tín có chứng thực, và là một nguồn chuẩn tham chiếu trong lĩnh vực khai phá dữ liệu bán lẻ.

Mỗi bản ghi chứa thông tin về mã hóa đơn, sản phẩm, số lượng, đơn giá, thời điểm giao dịch, khách hàng và quốc gia. Phần lớn giao dịch (trên 91%) đến từ Vương quốc Anh, với 4.372 khách hàng, nhưng có đến 24,93% bản ghi thiếu ID khách hàng, ảnh hưởng đến việc xác định hành vi từng cá nhân. Ngoài ra, dữ liệu cũng bao gồm 1,71% giao dịch hủy; 0,97% dữ liệu trùng lặp; 1,96% dữ liệu về số lượng có giá trị âm; 0,46% bản ghi đơn giá có giá trị bằng 0 và đơn giá bất thường, cho thấy sự tồn tại của dữ liệu không hợp lệ và ngoại lai. Đặc trưng của tập dữ liệu nghiên cứu được tổng hợp trong Bảng 1.

**Bảng 1.**

Bảng tổng hợp thuộc tính của dữ liệu Online Retail

| Thuộc tính  | Ý nghĩa                                  | Kiểu dữ liệu | Ghi chú                                                               |
|-------------|------------------------------------------|--------------|-----------------------------------------------------------------------|
| InvoiceNo   | Mã định danh duy nhất cho mỗi giao dịch  | Chuỗi        | Bắt đầu bằng "C" nếu là giao dịch hủy, tính Frequency trong RFM.      |
| StockCode   | Mã định danh duy nhất cho mỗi sản phẩm   | Chuỗi        | Có mã đặc biệt như "POST", ít liên quan đến RFM.                      |
| Description | Mô tả sản phẩm được mua                  | Chuỗi        | Có 0,27% giá trị thiếu, không nhất quán, ít liên quan RFM.            |
| Quantity    | Số lượng sản phẩm trong giao dịch        | Số nguyên    | Giá trị âm biểu thị trả hàng, cần xử lý để tính Monetary trong RFM.   |
| InvoiceDate | Thời gian thực hiện giao dịch            | Ngày giờ     | Chi tiết đến phút, dùng để tính Recency trong RFM.                    |
| UnitPrice   | Giá mỗi đơn vị sản phẩm (bảng Anh)       | Số thực      | Có giá trị âm (lỗi hoặc điều chỉnh), dùng để tính Monetary trong RFM. |
| CustomerID  | Mã định danh duy nhất cho mỗi khách hàng | Chuỗi        | Thiếu 24,93%, cần thiết để xác định khách hàng trong RFM.             |
| Country     | Quốc gia                                 | Chuỗi        | Phân bố không đều (UK chiếm 91,43%), ít liên quan đến RFM.            |

### 3.3. Làm sạch, xử lý và chuyển đổi dữ liệu

Các điểm dữ liệu được loại bỏ bao gồm: bản ghi thiếu CustomerID vì không thể xác định được khách hàng; bản ghi trùng lặp, để đảm bảo tính duy nhất của từng giao dịch; các bản ghi có giá trị Quantity âm, UnitPrice bằng 0 hoặc giao dịch bị hủy, nhằm phản ánh đúng hành vi mua sắm thực tế. Dữ liệu nhiễu và không nhất quán như mô tả sản phẩm sai lệch, mã sản phẩm không chuẩn, hoặc một StockCode tương ứng nhiều mô tả khác nhau cũng được loại bỏ, giúp đơn giản hóa tập dữ liệu và

tránh ảnh hưởng đến quá trình huấn luyện mô hình. Với các giá trị Quantity và UnitPrice quá lớn, chúng được giới hạn bằng kỹ thuật ngưỡng (Capping) theo giá trị IQR, giảm ảnh hưởng tiêu cực đến chỉ số Monetary trong RFM.

Thực hiện chuyển đổi Logarit khắc phục hiện tượng phân phối lệch với các biến Quantity, UnitPrice và TotalPrice và dữ liệu RFM về gần với phân phối chuẩn. Việc tạo ra các biến log cải thiện đáng kể độ lệch, từ đó tăng hiệu quả khi áp dụng các thuật toán học máy.

Việc tính các chỉ số RFM được thực hiện như sau: (1) Recency (R) = Ngày hiện tại - Ngày mua gần nhất (dựa trên InvoiceDate, với ngày hiện tại là ngày cuối cùng trong tập dữ liệu: 09/12/2011); (2) Frequency (F) = Số lượng hóa đơn duy nhất (InvoiceNo) của mỗi khách hàng (CustomerID); (3) Monetary (M) = Tổng giá trị giao dịch =  $\text{Sum}(\text{Quantity} * \text{UnitPrice})$  cho mỗi khách hàng. Ví dụ, một khách hàng với lần mua cuối cùng cách 10 ngày, 5 hóa đơn, và tổng chi 1000 GBP sẽ có RFM = (10, 5, 1000). Các chỉ số này sau đó được chuẩn hóa bằng Logarit và StandardScaler để giảm lệch phân phối.

### 3.4. Mô hình hóa dữ liệu

Sáu thuật toán phân cụm (K-Means, Spectral Clustering, Mean Shift, GMM, DBSCAN, FCM) được thực nghiệm trên tập dữ liệu Online Retail đã xử lý trước đó. Các thuật toán được lựa chọn đại diện cho nhiều phương pháp phân tích hành vi khách hàng khác nhau: từ phân cụm cứng như K-Means, DBSCAN đến phân cụm mềm như FCM, cũng như các phương pháp phù hợp với dữ liệu tuyến tính và phi tuyến tính. Các thuật toán yêu cầu xác định số cụm trước (K-Means, Spectral Clustering, GMM, FCM) được tối ưu hóa với  $K = 4$  dựa trên phương pháp Elbow và các chỉ số đánh giá như Silhouette Score, trong khi Mean Shift và DBSCAN xác định số cụm tự động dựa trên mật độ dữ liệu.

### 3.5. Chỉ số định lượng và đánh giá mô hình

Giai đoạn đánh giá mô hình phải đáp ứng được cả yếu tố kỹ thuật lẫn giá trị kinh doanh, đồng thời đảm bảo tính bền vững và khả năng mở rộng trong tương lai. Kết quả đánh giá dựa trên: các chỉ số đánh giá chất lượng cụm như: Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, Dunn Index; và mức độ đáp ứng mục tiêu kinh doanh như phân khúc khách hàng thành các nhóm có ý nghĩa.

### 3.6. Đề xuất giải pháp và triển khai mô hình

Kết quả huấn luyện và đánh giá mô hình cho thấy K-Means là lựa chọn tối ưu, tập dữ liệu thực nghiệm được chia thành 4 cụm rõ ràng. Mỗi cụm phản ánh đặc điểm hành vi mua sắm khác nhau thông qua các chỉ số RFM. Cụ thể, cụm 0 (Lost Customers) là tập khách hàng có thời gian mua gần nhất cách xa, ít giao dịch và tổng chi tiêu thấp; cụm 1 (Loyal Customers) với tần suất mua ổn định, chi tiêu khá, thời gian mua gần đây có giá trị trung bình; cụm 2 (Potential Customers), tập khách hàng mới hoặc mua gần đây nhưng chưa thường xuyên, chi tiêu trung bình; cụm 3 (Best Customers) là những người mua thường xuyên, gần đây và chi tiêu nhiều nhất. Sau khi diễn giải và đặt tên từng cụm, mỗi khách hàng được gán một nhãn cụ thể như Lost, Best, Loyal hoặc Potential Customers. Dữ liệu này được lưu trữ và tích hợp vào hệ thống CRM, sẵn sàng cho các chiến dịch marketing tự động (Bảng 2).

Việc gán nhãn các nhóm khách hàng dựa trên giá trị trung bình RFM và mục tiêu kinh doanh: Cụm 0 (Lost Customers) với R cao (189,73 ngày), F thấp (1,22), M thấp (220,27 GBP) - được gán nhãn vì hành vi rời bỏ; cụm 1 (Loyal Customers) với R trung bình (89,35 ngày), F trung bình (3,37), M khá (1009,38 GBP) - được gán nhãn vì trung thành ổn định; cụm 2 (Potential Customers) với R thấp (20,34 ngày), F thấp (1,92), M trung bình (373,92 GBP) - được gán nhãn vì tiềm năng mới; cụm 3 (Best Customers) với R thấp (15,39 ngày), F cao (11,23), M cao (3525,25 GBP) - được gán nhãn vì giá trị cao nhất. Cơ sở này kết hợp định lượng (RFM) và định tính (diễn giải kinh doanh), phù hợp với các chiến lược và mục tiêu marketing và kinh doanh của doanh nghiệp.

## Bảng 2.

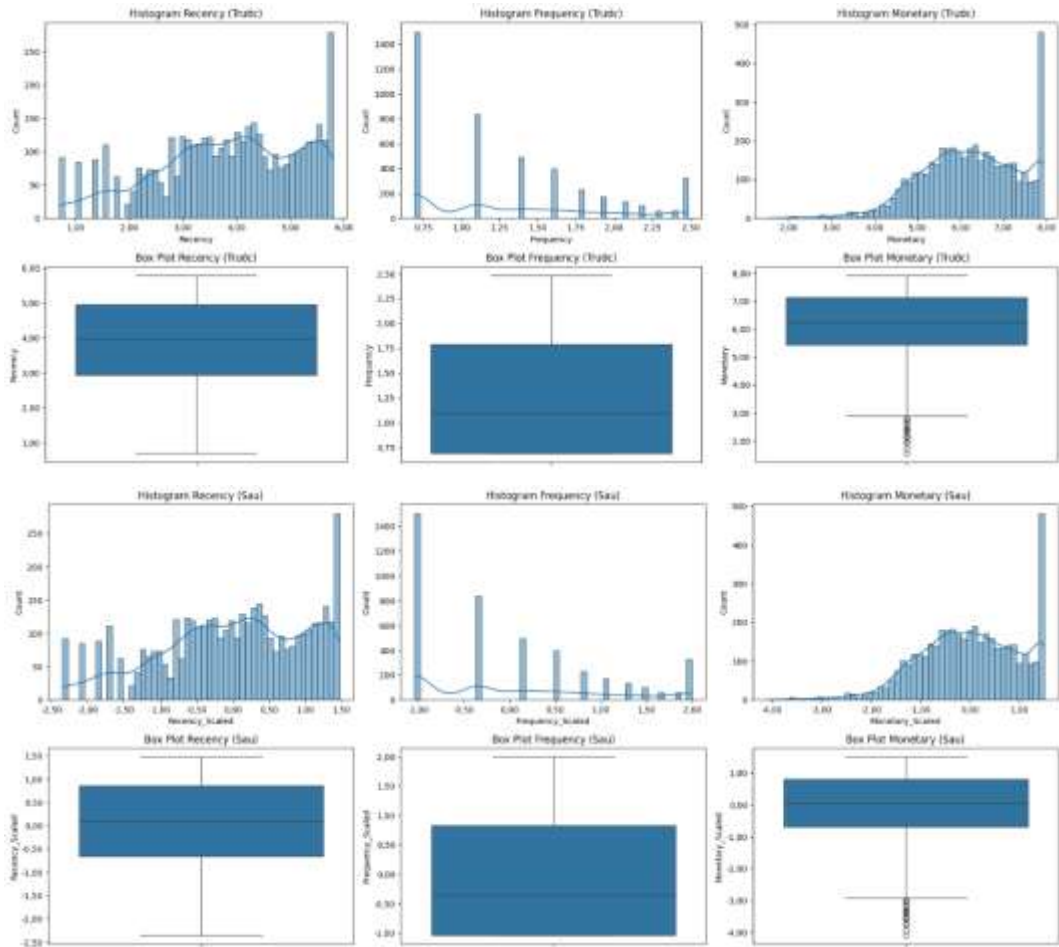
Bảng tổng hợp diễn giải, phân loại, đặt tên và dán nhãn

| Cụm | Recency<br>(Ngày) | Frequency<br>(Đơn hàng) | Monetary<br>(GBP) | Tỷ lệ | Tên cụm                | Diễn giải kinh doanh                                                                                        |
|-----|-------------------|-------------------------|-------------------|-------|------------------------|-------------------------------------------------------------------------------------------------------------|
| 0   | 189,73            | 1,22                    | 220,27            | ~33%  | Lost<br>Customers      | Khách hàng rời bỏ. Lâu không mua, tần suất rất thấp, chi tiêu thấp.                                         |
| 1   | 89,35             | 3,37                    | 1.009,38          | ~26%  | Loyal<br>Customers     | Khách hàng trung thành. Mua hàng lần gần nhất cách đây khoảng (3 tháng), tần suất trung bình, chi tiêu khá. |
| 2   | 20,34             | 1,92                    | 373,92            | ~18%  | Potential<br>Customers | Khách hàng tiềm năng. Mua hàng gần đây (20 ngày) tần suất thấp, chi tiêu trung bình.                        |
| 3   | 15,39             | 11,23                   | 3.525,25          | ~23%  | Best<br>Customers      | Khách hàng giá trị. Mua rất gần đây (15 ngày), tần suất cao, chi tiêu cao.                                  |

## 4. Kết quả nghiên cứu và thảo luận

### 4.1. Kết quả tiền xử lý và chuyển đổi dữ liệu

Kết quả sau quá trình làm sạch, từ tập dữ liệu Online Retail ban đầu, đã có 352.746 giao dịch được tiền xử lý tạo ra bộ dữ liệu RFM gồm 4.319 khách hàng, mỗi khách hàng có ba chỉ số RFM. Các biến RFM được kiểm tra để xác định giá trị ngoại lai bằng phương pháp IQR. Những khách hàng có chỉ số quá lớn (mua hàng quá thường xuyên hoặc chi tiêu rất cao) – các giá trị ngoại lai – sẽ được xử lý bằng kỹ thuật giới hạn để tránh ảnh hưởng xấu đến kết quả mô hình. Cuối cùng, để giảm thiểu độ lệch và đảm bảo các biến có cùng thang đo, kết quả chuyển đổi Logarit được chuẩn hóa bằng công cụ StandardScaler (trung bình: 0, độ lệch chuẩn: 1) (Hình 2).



**Hình 2.** Biểu đồ Histogram và Boxplot biểu diễn kết quả trước và sau khi chuẩn hóa dữ liệu bằng phương pháp StandardScaler

#### 4.2. Kết quả huấn luyện và đánh giá mô hình

Hiệu suất của các thuật toán được đánh giá bằng các chỉ số đo lường (Bảng 3) như: Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, Dunn Index.

**Bảng 3.**

Bảng tổng hợp đánh giá chất lượng phân cụm

| Thuật toán          | Số cụm | Silhouette Score | Davies Bouldin Index | Calinski Harabasz Index | Dunn Index | Nhận xét                                                                     |
|---------------------|--------|------------------|----------------------|-------------------------|------------|------------------------------------------------------------------------------|
| K-Means             | 4      | 0,343            | 1,011                | 3.596,578               | 0,005      | Tách biệt tốt, chất lượng cao, phù hợp mục tiêu kinh doanh.                  |
| Spectral Clustering | 4      | 0,016            | 1,435                | 1.616,374               | 0,012      | Kém tách biệt, khó diễn giải về mặt kinh doanh.                              |
| Mean Shift          | 1      | -                | -                    | -                       | -          | Không phân đoạn hiệu quả, không phù hợp.                                     |
| GMM                 | 4      | 0,269            | 1,180                | 2.587,816               | 0,005      | Tương tự K-Means nhưng kém tách biệt.                                        |
| DBSCAN              | 2      | 0,324            | 1,014                | 2.803,268               | 0,124      | Ít cụm, thiếu các nhóm quan trọng, không đầy đủ.                             |
| FCM                 | 4      | 0,316            | 1,092                | 3.495,526               | 0,003      | Tách biệt tốt, hỗ trợ phân cụm mềm, nhưng kém K-Means về diễn giải đơn giản. |

Cụ thể, Silhouette Score của K-Means (0,343) cao hơn FCM (0,316) khoảng 8,5% và GMM (0,269) khoảng 27,5%, cho thấy độ gắn kết nội cụm và tách biệt liên cụm tốt hơn, phù hợp với dữ liệu RFM đã chuẩn hóa. Davies-Bouldin Index thấp nhất (1,011) so với Spectral Clustering (1,435) và GMM (1,180) chứng tỏ tỷ lệ tương đồng giữa các cụm thấp, giảm thiểu chồng lấp. Calinski-Harabasz Index cao nhất (3.596,578) phản ánh phân tán giữa cụm lớn hơn phân tán nội cụm, lý tưởng cho phân khúc khách hàng đa dạng. Dunn Index (0,005) tuy không cao nhất (DBSCAN: 0,124), nhưng kết hợp với số cụm tối ưu (4 cụm) đảm bảo tính thực tiễn, trong khi Mean Shift chỉ tạo 1 cụm và DBSCAN chỉ 2 cụm, không đáp ứng yêu cầu kinh doanh. Tổng thể, K-Means vượt trội về cân bằng giữa chất lượng kỹ thuật và khả năng diễn giải kinh doanh, hỗ trợ triển khai CRM hiệu quả.

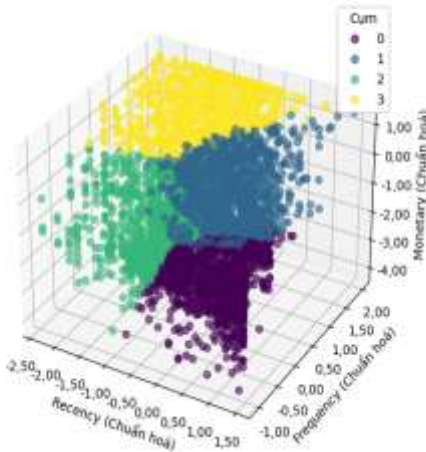
Việc xử lý các giá trị ngoại lai (Outliers) bằng kỹ thuật ngưỡng (Capping) dựa trên khoảng tứ phân vị (IQR) đã tác động tích cực đến kết quả phân tích tổng thể, nâng cao chất lượng phân cụm thông qua việc giảm nhiễu và độ lệch của dữ liệu. Cụ thể, trước khi xử lý, dữ liệu chứa các giá trị ngoại lai về chi tiêu và tần suất mua sắm, gây ra phân phối lệch cao, làm giảm độ tách biệt và đồng nhất của các cụm, từ đó ảnh hưởng đến hiệu suất của các thuật toán phân cụm. Sau khi áp dụng kỹ thuật Capping IQR để giới hạn các giá trị bất thường trong các chỉ số RFM, kết hợp với chuẩn hóa dữ liệu bằng StandardScaler, chất lượng phân cụm được cải thiện đáng kể, với Silhouette Score của thuật toán K-Means đạt 0,343 (Bảng 3). Điều này phản ánh sự gia tăng trong độ tách biệt giữa các cụm và tính đồng nhất trong mỗi cụm, đảm bảo bốn phân khúc khách hàng (Best Customers, Loyal Customers, Potential Customers, Lost Customers) được xác định rõ ràng và có ý nghĩa kinh doanh cao. Quá trình xử lý outliers không chỉ loại bỏ nhiễu từ các giao dịch không điển hình mà còn bảo toàn các đặc điểm hành vi mua sắm quan trọng, duy trì tính đại diện của dữ liệu. Kết quả này khẳng

định vai trò quan trọng của việc tiền xử lý dữ liệu trong việc nâng cao độ chính xác và giá trị thực tiễn của phân cụm, phù hợp với mục tiêu kinh doanh và các tiêu chuẩn phân tích dữ liệu quốc tế.

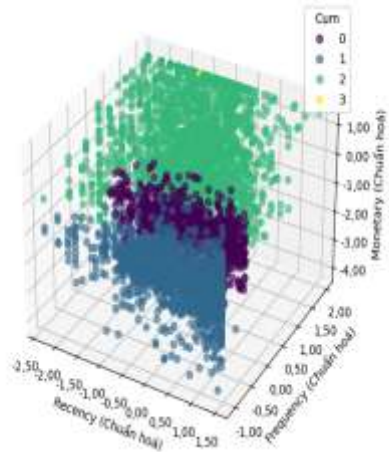
#### 4.3. Trực quan hóa kết quả phân cụm

Kết quả phân cụm bằng các thuật toán K-Means, Spectral Clustering, Mean Shift, GMM, DBSCAN, FCM trên dữ liệu thực nghiệm được trực quan hóa (Hình 3).

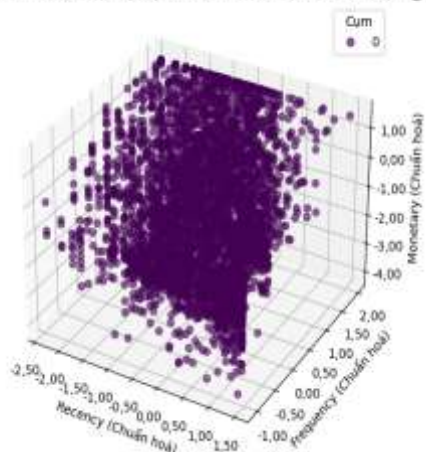
Biểu đồ phân tán 3D của kết quả phân cụm theo thuật toán KMeans\_Cl



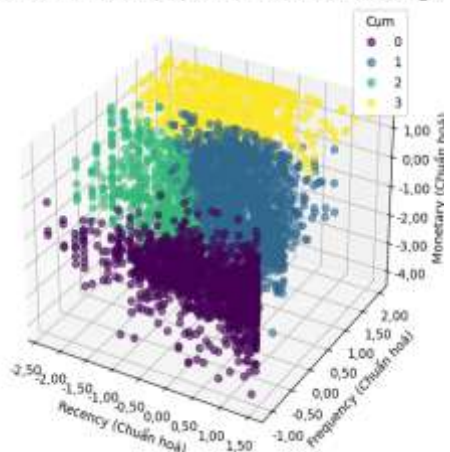
Biểu đồ phân tán 3D của kết quả phân cụm theo thuật toán SpectralClustering\_C

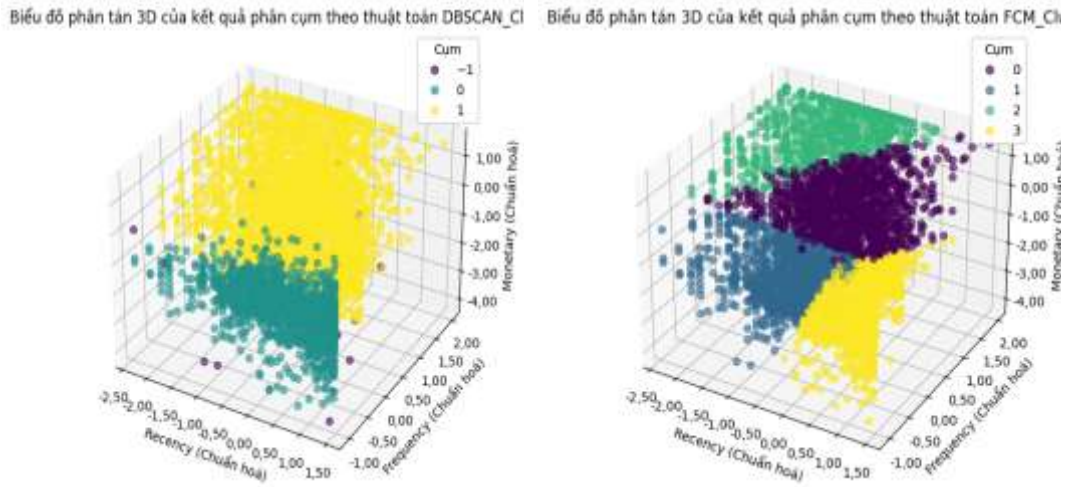


Biểu đồ phân tán 3D của kết quả phân cụm theo thuật toán MeanShift\_Cl



Biểu đồ phân tán 3D của kết quả phân cụm theo thuật toán GMM\_Cl





**Hình 3.** Biểu đồ phân tán 3D minh họa kết quả phân cụm bằng các thuật toán K-Means, Spectral Clustering, MeanShift, GMM, DBSCAN, FCM trên dữ liệu RFM

Đáp ứng mục tiêu kinh doanh (Bảng 4), K-Means phân cụm khách hàng thành bốn nhóm rõ ràng tương ứng với hành vi và giá trị cụ thể, hỗ trợ tốt cho các chiến lược marketing như ưu đãi, giữ chân khách hàng, hoặc tái kích hoạt. FCM cũng phân nhóm tương tự và phù hợp với hành vi khách hàng chồng lấp, nhưng khó triển khai hơn. Các thuật toán còn lại hoặc không đủ cụm (Mean Shift, DBSCAN) hoặc tạo ra các cụm không rõ ràng (Spectral Clustering, GMM).

Cụ thể, các cụm từ K-Means có phân bố tỷ lệ hợp lý (~33% Lost, ~26% Loyal, ~18% Potential, ~23% Best) (Bảng 2), phản ánh cấu trúc dữ liệu thực tế và dễ dàng gắn với chiến lược marketing (ví dụ: ưu tiên Best Customers với Monetary trung bình 3.525,25 GBP). Ngược lại, Spectral Clustering tạo cụm chồng lấp (Silhouette Score chỉ 0,016), dẫn đến diễn giải kinh doanh phức tạp; Mean Shift và DBSCAN thiếu đa dạng cụm, không hỗ trợ phân khúc toàn diện như yêu cầu của mô hình RFM.

**Bảng 4.**

Bảng tổng hợp đánh giá dựa trên mục tiêu kinh doanh

| Thuật toán | Cụm | Giá trị trung bình     |                              |                        | Diễn giải kinh doanh | Phù hợp mục tiêu              |
|------------|-----|------------------------|------------------------------|------------------------|----------------------|-------------------------------|
|            |     | R<br>Recency<br>(Ngày) | F<br>Frequency<br>(Đơn hàng) | M<br>Monetary<br>(GBP) |                      |                               |
| K-Means    | 0   | 189,73                 | 1,22                         | 220,27                 | Lost Customers       | Phù hợp, phân biệt rõ 4 nhóm. |
|            | 1   | 89,35                  | 3,37                         | 1.009,38               | Loyal Customers      |                               |
|            | 2   | 20,34                  | 1,92                         | 373,92                 | Potential Customers  |                               |
|            | 3   | 15,39                  | 11,23                        | 3.525,25               | Best Customers       |                               |
|            | 0   | 99,44                  | 2,02                         | 505,24                 | At Risk              |                               |

|                     |   |        |       |          |                     |                                             |
|---------------------|---|--------|-------|----------|---------------------|---------------------------------------------|
| Spectral Clustering | 1 | 153,22 | 1,00  | 262,93   | Lost Customers      | Phức tạp, thiếu nhóm quan trọng.            |
|                     | 2 | 42,49  | 5,82  | 1.452,91 | Loyal Customers     |                                             |
|                     | 3 | 9,00   | 11,00 | 2.783,28 | Best Customers      |                                             |
| Mean Shift          | 0 | 91,39  | 3,46  | 868,52   | At Risk             | Không phù hợp, chỉ có 1 cụm.                |
| GMM                 | 0 | 153,26 | 1,00  | 263,75   | Lost Customers      | Phù hợp nhưng kém tách biệt so với K-Means. |
|                     | 1 | 98,20  | 3,28  | 835,16   | At Risk             |                                             |
|                     | 2 | 15,67  | 2,70  | 618,86   | Loyal Customers     |                                             |
|                     | 3 | 22,66  | 8,79  | 2.210,12 | Best Customers      |                                             |
| DBSCAN              | 0 | 153,45 | 1,00  | 264,06   | Lost Customers      | Không phù hợp, thiếu nhóm.                  |
|                     | 1 | 58,24  | 4,77  | 1.197,61 | Loyal Customers     |                                             |
| FCM                 | 0 | 217,16 | 1,29  | 244,10   | Lost Customers      | Phù hợp, hỗ trợ phân cụm mềm.               |
|                     | 1 | 12,62  | 8,31  | 2.101,64 | Best Customers      |                                             |
|                     | 2 | 39,71  | 1,53  | 314,93   | Potential Customers |                                             |
|                     | 3 | 65,85  | 3,81  | 1.085,86 | Loyal Customers     |                                             |

*Chú thích:* Tỷ lệ phân bố được trình bày chi tiết cho K-Means trong Bảng 2, vì đây là thuật toán tối ưu; Bảng 4 tập trung vào so sánh giá trị RFM và diễn giải kinh doanh của 6 thuật toán.

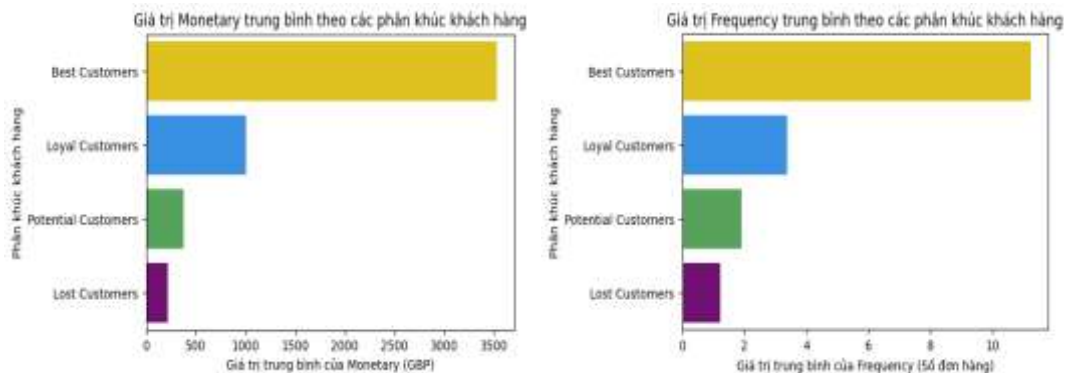
#### 4.4. Thảo luận

Bảng 3 cung cấp so sánh hiệu suất định lượng của K-Means với các thuật toán khác, củng cố lý do K-Means là lựa chọn tối ưu. Cụ thể, K-Means đạt Silhouette Score cao nhất (0,343), vượt trội hơn FCM (0,316) khoảng 8,5% và GMM (0,269) khoảng 27,5%, phản ánh độ gắn kết nội cụm và tách biệt liên cụm tốt hơn, phù hợp với dữ liệu RFM đã chuẩn hóa. Davies-Bouldin Index thấp nhất (1,011) so với Spectral Clustering (1,435) và GMM (1,180) chứng tỏ tỷ lệ tương đồng giữa các cụm thấp, giảm thiểu chồng lấp. Calinski-Harabasz Index cao nhất (3.596,578) phản ánh phân tán giữa cụm lớn hơn phân tán nội cụm, lý tưởng cho phân khúc khách hàng đa dạng. Dunn Index (0,005) tuy không cao nhất (DBSCAN: 0,124), nhưng kết hợp với số cụm tối ưu (4 cụm) đảm bảo tính thực tiễn, trong khi Mean Shift chỉ tạo 1 cụm và DBSCAN chỉ 2 cụm, không đáp ứng yêu cầu phân đoạn kỹ thuật. Kết quả này khẳng định khả năng của K-Means trong việc đáp ứng mục tiêu nghiên cứu, đặc biệt trong khai phá dữ liệu bán lẻ trực tuyến.

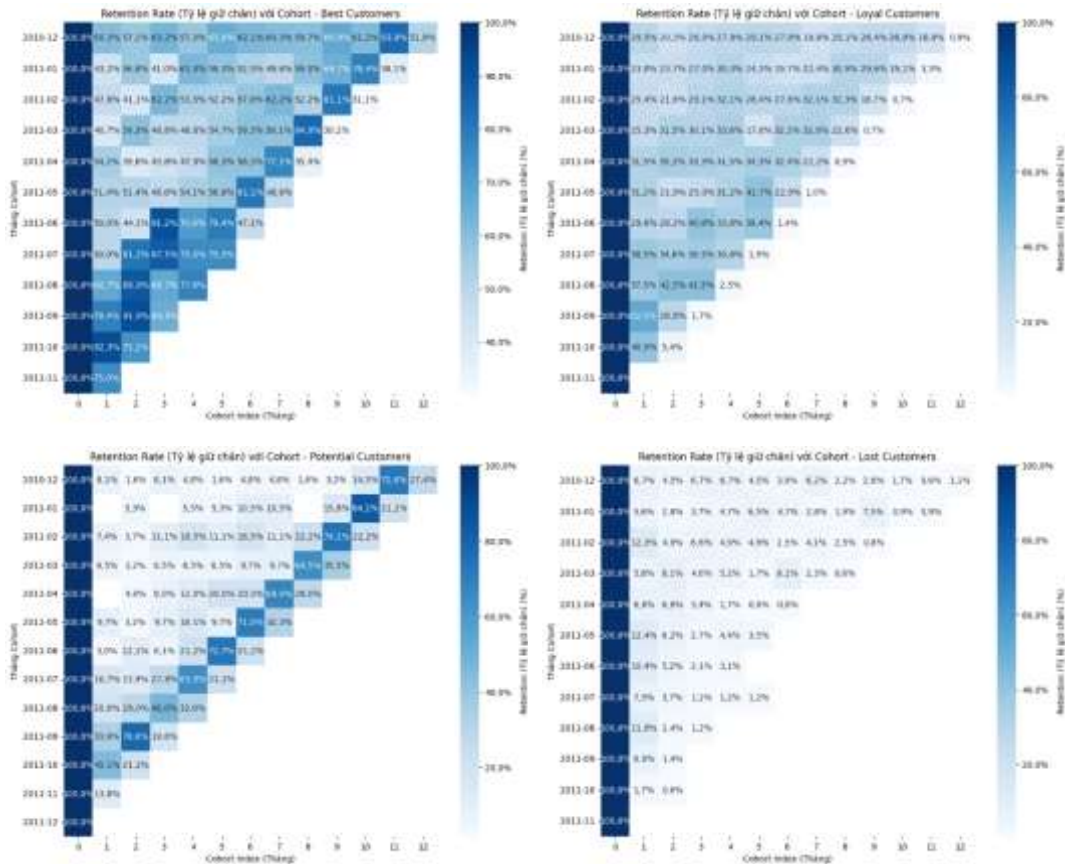
Kết quả thực nghiệm của bài báo chứng minh rằng cách tiếp cận đa chiều với sáu thuật toán phân cụm (K-Means, Spectral Clustering, Mean Shift, GMM, DBSCAN, FCM) và bốn chỉ số đánh giá (Silhouette Score, Davies-Bouldin, Calinski-Harabasz, Dunn Index) cung cấp cái nhìn toàn diện về chất lượng phân cụm trên dữ liệu Online Retail. Cụ thể, K-Means đạt Silhouette Score 0,343 với bốn cụm rõ ràng (Bảng 3), phản ánh hành vi khách hàng đa dạng và phù hợp với mục tiêu kinh doanh như cá nhân hóa marketing. Cách tiếp cận này khắc phục hạn chế của các phương pháp đơn lẻ bằng cách

đánh giá đa chiều, từ đó tăng tính ứng dụng thực tiễn khi tích hợp vào hệ thống CRM, hỗ trợ tối ưu hóa chiến lược giữ chân khách hàng và nâng cao doanh thu.

Với kết quả nghiên cứu là khai thác mô hình RFM và phân loại khách hàng thành bốn nhóm: Best Customers (23%), Loyal Customers (26%), Potential Customers (18%) và Lost Customers (33%), dựa trên dữ liệu thực nghiệm (Bảng 2; Hình 4), được ứng dụng để đề xuất chiến lược marketing cá nhân hóa cho thị trường bán lẻ trực tuyến Việt Nam, ngành được dự báo đạt 144,9 tỷ USD vào năm 2032 với tốc độ tăng trưởng kép hàng năm (CAGR) 18,9% trong giai đoạn 2025–2032 (Verified Market Research, 2025). Cụ thể, Best Customers, chiếm tỷ lệ khoảng 23% và ước tính đóng góp khoảng 33,3 tỷ USD, phù hợp với các chương trình ưu đãi VIP, quyền lợi độc quyền và trải nghiệm mua sắm cao cấp trên Shopee, Lazada. Loyal Customers, chiếm tỷ lệ khoảng 26% và ước tính đóng góp khoảng 37,7 tỷ USD, nên được duy trì và mở rộng qua chương trình tích điểm, bán chéo sản phẩm, gói ưu đãi theo mùa vụ và các chiến dịch trò chơi hóa để khuyến khích mua lặp lại. Potential Customers, chiếm tỷ lệ khoảng 18% và ước tính đóng góp khoảng 26,1 tỷ USD, cần nhắm mục tiêu qua quảng cáo cá nhân hóa trên Facebook và TikTok (hai nền tảng chiếm 70% lưu lượng thương mại điện tử tại Việt Nam), kết hợp ưu đãi dùng thử và remarketing. Lost Customers, chiếm tỷ lệ khoảng 33% và ước tính đóng góp khoảng 47,8 tỷ USD, nên được tái kích hoạt thông qua khảo sát nguyên nhân rời bỏ, chiến dịch hoàn tiền, ưu đãi mạnh vào dịp cao điểm như 11.11 hoặc Tết Nguyên đán, nhằm giảm tỷ lệ rời bỏ, qua đó tiết kiệm chi phí marketing và tăng doanh thu đóng góp thật sự. Các ứng dụng marketing rút ra từ đề tài này phù hợp với bối cảnh Việt Nam, nơi thương mại điện tử phát triển nhanh và cạnh tranh mạnh. Doanh nghiệp có thể: (1) Xây dựng chiến lược phân bổ ngân sách marketing theo giá trị khách hàng thay vì phân bổ đồng đều; (2) phát triển công cụ cá nhân hóa nội dung dựa trên dữ liệu RFM và hành vi mua sắm; (3) áp dụng tự động hóa tiếp thị qua email, SMS và chatbot để duy trì tương tác liên tục; (4) triển khai chiến dịch đa kênh kết hợp online - offline, đặc biệt trong dịp khuyến mại lớn; và (5) sử dụng thử nghiệm A/B để tối ưu thông điệp, hình ảnh và kênh truyền thông cho từng phân khúc. Lưu ý rằng các số liệu thị trường và doanh thu tiềm năng của các phân khúc khách hàng trong nghiên cứu này đều được ước lượng và phỏng chiếu dựa trên báo cáo ngành, tỷ lệ phân khúc RFM từ dữ liệu thực nghiệm, cùng với giả định về mức tăng trưởng và hành vi tiêu dùng trong bối cảnh Việt Nam. Do đó, khi áp dụng vào thực tiễn, các doanh nghiệp cần kết hợp thêm dữ liệu CRM địa phương, kết quả thử nghiệm A/B và theo dõi phản ứng thị trường để điều chỉnh chiến lược cho phù hợp.



**Hình 4.** Biểu đồ mô tả giá trị trung bình của Monetary và Frequency theo từng phân khúc



Hình 5. Biểu đồ mô tả Retention Rate theo thời gian của từng phân khúc

## 5. Kết luận và hướng phát triển

Nghiên cứu đã chứng minh rằng mô hình RFM kết hợp với thuật toán phân cụm là một phương pháp hiệu quả để phân khúc khách hàng trong lĩnh vực bán lẻ trực tuyến. Dựa trên dữ liệu giao dịch của khách hàng từ tập dữ liệu Online Retail, mô hình RFM được xây dựng từ ba chỉ số: Recency, Frequency và Monetary, thể hiện rõ hành vi mua sắm của từng khách hàng. Sau khi chuẩn hóa dữ liệu và áp dụng 6 thuật toán phân cụm, kết quả cho thấy thuật toán K-Means vượt trội nhất với các chỉ số chất lượng phân cụm dẫn đầu và khả năng tạo ra bốn nhóm khách hàng có ý nghĩa kinh doanh rõ ràng: Best Customers, Loyal Customers, Potential Customers và Lost Customers.

Kết quả thực nghiệm đã giải đáp toàn diện các câu hỏi và mục tiêu nghiên cứu, đồng thời đáp ứng được cả yếu tố kỹ thuật lẫn giá trị kinh doanh. Các phân khúc khách hàng thu được tích hợp vào hệ thống CRM, làm cơ sở đề xuất các chiến lược marketing cá nhân hóa. Những chiến lược này không chỉ hỗ trợ doanh nghiệp nâng cao hiệu quả quản lý quan hệ khách hàng và tối ưu hóa hoạt động marketing, mà còn góp phần gia tăng doanh thu, nâng cao năng lực công nghệ và củng cố lợi thế cạnh tranh trong bối cảnh chuyển đổi số hiện nay.

Nghiên cứu này cung cấp những đóng góp thiết thực, đặc biệt trong việc ứng dụng phân khúc khách hàng vào thực tiễn ngành bán lẻ trực tuyến tại Việt Nam, nơi thị trường đang phát triển mạnh mẽ nhưng còn thiếu các giải pháp dựa trên dữ liệu hiệu quả. Kết quả nghiên cứu xác định bốn phân khúc khách hàng rõ ràng: Best Customers (nhóm giá trị cao với tần suất mua sắm thường xuyên và chi tiêu lớn), Loyal Customers (nhóm trung thành với tần suất ổn định và chi tiêu khá), Potential Customers (nhóm tiềm năng với hành vi mua mới nhưng chưa thường xuyên), và Lost Customers (nhóm rời bỏ với tần suất và chi tiêu thấp). Các phân khúc này, được xây dựng từ dữ liệu thực nghiệm, có thể được tích hợp vào hệ thống CRM để triển khai các chiến lược marketing cá nhân hóa, đặc biệt phù hợp với bối cảnh Việt Nam, nơi thị trường thương mại điện tử dự kiến đạt quy mô đáng kể trong vài năm tới. Cụ thể, doanh nghiệp có thể áp dụng chương trình tri ân và ưu đãi độc quyền cho Best Customers để duy trì doanh thu cao trên các nền tảng như Tiki, Shopee hoặc Lazada; triển khai chiến lược bán chéo và tích điểm cho Loyal Customers để gia tăng giá trị vòng đời khách hàng; sử dụng quảng cáo nhắm mục tiêu và email marketing để nuôi dưỡng Potential Customers, tận dụng xu hướng tiêu dùng trực tuyến trẻ hóa; và áp dụng các biện pháp tái kích hoạt chọn lọc cho Lost Customers thông qua khảo sát phản hồi hoặc ưu đãi đặc biệt để giảm tỷ lệ rời bỏ, một thách thức lớn trong môi trường cạnh tranh khốc liệt tại Việt Nam. Những ứng dụng này không chỉ tối ưu hóa nguồn lực và tăng cường doanh thu tiềm năng mà còn củng cố lợi thế cạnh tranh, đặc biệt trong bối cảnh chuyển đổi số quốc gia, nơi các chính sách hỗ trợ doanh nghiệp nhỏ và vừa đang thúc đẩy việc áp dụng công nghệ dữ liệu tiên tiến.

Mặc dù bộ dữ liệu Online Retail giai đoạn 2010–2011 đã được sử dụng trong nhiều nghiên cứu trước (Chen và cộng sự, 2012), và hành vi khách hàng trực tuyến đã thay đổi đáng kể đến năm 2025 (do sự phát triển của thương mại điện tử di động và AI), kết quả vẫn có giá trị lý thuyết như một chuẩn tham chiếu để so sánh thuật toán. Tuy nhiên, điều này có thể giảm tính liên quan ứng dụng trực tiếp. Do đó, nghiên cứu đề xuất mở rộng với dữ liệu mới hơn từ các nền tảng thương mại điện tử Việt Nam để tăng tính thuyết phục cho quản trị marketing hiện đại.

---

## Tài liệu tham khảo

- Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21(3), 527-570.
- Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), 1785-1792.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.
- Chen, D. (2015). *Online Retail* (UCI Machine Learning Repository). <https://doi.org/10.24432/C5BW33>
- Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19, 197-208.

- Cheng, C.-H., & Chen, Y.-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, 36(3), 4176-4184.
- Cho, Y. S., Moon, S. C., Jeong, S.-p., Oh, I.-B., & Ryu, K. H. (2013). Clustering method using item preference based on RFM for recommendation system in u-commerce. In Jeong, YS., Park, YH., Hsu, CH., Park, J. (eds), *Ubiquitous Information Technologies and Applications. Lecture Notes in Electrical Engineering* (vol 280). Berlin, Heidelberg: Springer. [https://doi.org/10.1007/978-3-642-41671-2\\_18](https://doi.org/10.1007/978-3-642-41671-2_18)
- Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, 33(10), 1251-1257.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-37.
- Hughes, A. M. (1994). *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable Customer-Based Marketing Program*. Probus Pub. Co.
- Hughes, A. M. (2005). *Strategic Database Marketing* (3<sup>rd</sup> Ed.). McGraw-Hill Pub. Co.
- Ibrahim, M. R. K., & Tyasnurita, R. (2022). *LRFM model analysis for customer segmentation using K-means clustering*. 2022 International Conference on Electrical and Information Technology (IEIT), Malang, Indonesia, 2022, pp. 383-391.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, 31(3), 264-323.
- Kotler, P. (1967). *Marketing Management: Analysis, Planning, and Control*. Prentice-Hall.
- Kotler, P. (2012). *Kotler on Marketing*. UK: Simon and Schuster.
- Kotler, P., & Armstrong, G. (2017). *Principles of Marketing* (17<sup>th</sup> ed., Global Ed.). Pearson Education Limited.
- Kumar, V., & Reinartz, W. J. (2012). *Customer Relationship Management: Concept, Strategy, and Tools*. Springer.
- Lang, L., Zhou, S., Zhong, M., Sun, G., Pan, B., & Guo, P. (2023). A big data based dynamic weight approach for RFM segmentation. *Computers, Materials & Continua*, 74(2).
- Li, P., Wang, C., Wu, J., & Madleňák, R. (2022). *An E-commerce customer segmentation method based on RFM weighted K-means*. 2022 International Conference on Management Engineering, Software Engineering and Service Sciences (ICMSS), Wuhan, China, 2022, pp. 61-68.
- MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, 281-297.
- Mensouri, D., Azmani, A., & Azmani, M. (2022). K-Means customers clustering by their RFMT and score satisfaction analysis. *International Journal of Advanced Computer Science and Applications*, 13(6), 469-476.
- Rizki, B., Ginasta, N. G., Tamrin, M. A., & Rahman, A. (2020). Customer loyalty segmentation on point of sale system using recency-frequency-monetary (RFM) and K-means. *Jurnal Online Informatika*, 130-136.

- Verified Market Research. (2025). *Vietnam E-commerce market size by device (mobile, desktop), by application (beauty and personal care, consumer electronics, fashion and apparel, food and beverage), by consumer demographics (age, income, location), & region for 2025-2032*. Retrieved August 15, 2025 from <https://www.verifiedmarketresearch.com/product/vietnam-ecommerce-market/>
- Wang, G. (2025). Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-Means clustering. *PloS One*, 20(2), e0318519.
- Wong, C.-G., Tong, G.-K., & Haw, S.-C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy*, 12(3), 97-125.
- Zahrotun, L. (2017). *Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering*. 2017 2<sup>nd</sup> International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 299-303.