

TRƯỜNG ĐẠI HỌC KINH TẾ ĐÀ NẴNG
KHOA THƯƠNG MẠI ĐIỆN TỬ



PHÂN CỤM KHÁCH HÀNG TỪ DỮ LIỆU
GIAO DỊCH BẰNG MÔ HÌNH LRFM &
THUẬT TOÁN
K-MEANS

Đà Nẵng, ngày 03 tháng 05 năm 2025

Mục lục

I. GIỚI THIỆU TỔNG QUAN.....	6
1. Lý do chọn đề tài.....	6
2. Mục tiêu của đề tài.....	6
3. Mô tả dữ liệu.....	7
II. CƠ SỞ LÝ THUYẾT.....	8
1. Phân khúc khách hàng (Customer Segmentation).....	8
2. Phân cụm dữ liệu (Clustering).....	8
3. Mô hình RFM (RFM Model).....	9
4. Mô hình LRFM (LRFM Model).....	11
5. Thuật toán K-Means.....	12
6. Kiểm định giả thuyết (Hypothesis Testing).....	14
7. Phân tích phương sai đa biến (Multivariate Analysis of Variance - MANOVA).....	15
8. Phân tích phương sai một chiều (Analysis of Variance - ANOVA).....	16
9. Phân tích hậu định (Post-hoc Analysis).....	17
10. Phân tích vòng đời giá trị khách hàng (Customer Lifetime Value Analysis).....	18
III. PHƯƠNG PHÁP NGHIÊN CỨU.....	19
1. Data Preparation.....	20
2. Clustering with K-means.....	24
3. Hypothesis testing.....	25
3.1 Phân tích phương sai đa biến (Multivariate Analysis of Variance - MANOVA).....	25
3.2 Phân tích phương sai một chiều (Analysis of Variance - ANOVA).....	26
3.3 Phân tích hậu định (Post-hoc Analysis) với Tukey's Honest Significant Difference.....	28
IV. KẾT QUẢ.....	31

V. KẾT LUẬN.....35

VI. TÀI LIỆU THAM KHẢO.....36

MỤC LỤC HÌNH ẢNH

Hình 1: Quy trình hoạt động của thuật toán K-Means.....	13
Hình 2: Quy trình bài toán (Framework).....	19
Hình 3: Dữ liệu gốc của bài toán.....	21
Hình 4: Dữ liệu sau khi đã tiền xử lý.....	22
Hình 5: Các chỉ số L,R,F,M sau khi tính toán.....	23
Hình 6: Thống kê mô tả của các chỉ số L,R,F,M.....	23
Hình 7:Các chỉ số L,R,F,M sau khi đã được chuẩn hóa.....	24
Hình 8: Đồ thị của Elbow Method và Silhoutte Score.....	24
Hình 9: Kết quả sau khi đã được phân cụm.....	25
Hình 10:Tổng hợp số lượng khách hàng vào từng cụm và các chỉ số của cụm.....	25
Hình 11: Kết quả sau khi thực hiện MANOVA.....	26
Hình 12: Kết quả sau khi thực hiện ANOVA cho biến Length.....	27
Hình 13: Kết quả sau khi thực hiện ANOVA cho biến Recency.....	27
Hình 14: Kết quả sau khi thực hiện ANOVA cho biến Frequency.....	27
Hình 15: Kết quả sau khi thực hiện ANOVA cho biến Monetary.....	28
Hình 16: Kết quả sau khi thực hiện Tukey HSD cho biến Length.....	29
Hình 17: Kết quả sau khi thực hiện Tukey HSD cho biến Recency.....	29
Hình 18: Kết quả sau khi thực hiện Tukey HSD cho biến Frequency.....	30
Hình 19: Kết quả sau khi thực hiện Tukey HSD cho biến Monetary.....	30
Hình 20: Kết quả sau khi chạy thuật toán Kmeans.....	31
Hình 21:Hình ảnh trực quan về kết quả của từng cụm.....	31
Hình 22: Kết quả của Cụm 0.....	31
Hình 23: Kết quả của cụm 1.....	32

Hình 24: Kết quả của cụm 2.....32

Hình 25: Kết quả của cụm 3.....32

I. GIỚI THIỆU TỔNG QUAN

1. Lý do chọn đề tài

Trong bối cảnh dữ liệu ngày càng trở thành tài sản chiến lược của doanh nghiệp, việc phân tích hành vi khách hàng từ dữ liệu giao dịch đã và đang trở thành xu hướng tất yếu. Một trong những ứng dụng quan trọng là phân cụm khách hàng nhằm hỗ trợ cá nhân hóa chiến lược tiếp thị, nâng cao hiệu quả chăm sóc và giữ chân khách hàng (Nguyen et al., 2020). Xuất phát từ nhu cầu này, đề tài lựa chọn mô hình LRFM kết hợp với thuật toán K-Means như một hướng tiếp cận hiệu quả trong việc phân khúc khách hàng.

So với mô hình RFM truyền thống, mô hình LRFM bổ sung chỉ số Length – đại diện cho khoảng thời gian gắn bó giữa khách hàng và doanh nghiệp. Điều này cho phép đánh giá khách hàng không chỉ qua giá trị giao dịch mà còn qua độ trung thành, từ đó phản ánh hành vi tiêu dùng một cách toàn diện hơn (Tsiptsis & Chorianopoulos, 2009). Trong khi đó, thuật toán K-Means được lựa chọn nhờ tính đơn giản, hiệu quả và phù hợp với dữ liệu dạng số trong mô hình LRFM. Thuật toán này đã được ứng dụng rộng rãi trong nhiều nghiên cứu và chứng minh hiệu quả trong việc nhận diện các nhóm hành vi khách hàng tiềm ẩn (Han et al., 2011).

Ngoài phương pháp phân cụm, nghiên cứu còn tích hợp phân tích giá trị vòng đời khách hàng (Customer Lifetime Value – CLV) nhằm xác định nhóm khách hàng có giá trị cao để ưu tiên chăm sóc, góp phần tối ưu hóa chiến lược giữ chân và nâng cao hiệu quả kinh doanh (Venkatesan & Kumar, 2004). Bên cạnh đó, nhóm nghiên cứu cũng thực hiện kiểm định giả thuyết bằng phương pháp phân tích phương sai đa biến (MANOVA), kết hợp với kiểm định hậu nghiệm (post-hoc) nhằm kiểm tra sự khác biệt giữa các cụm khách hàng trên nhiều khía cạnh. Qua đó, đề tài không chỉ nâng cao độ tin cậy cho mô hình mà còn cung cấp cơ sở khoa học để xây dựng các chiến lược phù hợp cho từng phân khúc cụ thể.

2. Mục tiêu của đề tài

2.1. Mục tiêu tổng quát

Nghiên cứu và áp dụng mô hình LRFM kết hợp với thuật toán K-Means để phân cụm khách hàng dựa trên vòng đời và giá trị của họ, từ đó hỗ trợ doanh nghiệp đưa ra chiến lược marketing và chăm sóc khách hàng hiệu quả.

2.2. Mục tiêu cụ thể

- **Xây dựng framework phân cụm khách hàng**
 - Áp dụng mô hình LRFM để xác định các yếu tố quan trọng ảnh hưởng đến hành vi mua sắm.
 - Sử dụng thuật toán K-Means để phân nhóm khách hàng dựa trên dữ liệu số.
- **Phân tích vòng đời và giá trị khách hàng**
 - Xác định các nhóm khách hàng theo vòng đời (khách hàng mới, trung thành, sắp rời bỏ, v.v.).
 - Đánh giá giá trị của từng nhóm để đề xuất chiến lược kinh doanh phù hợp.
- **Kiểm định và đánh giá các phân nhóm:**

- Thực hiện phân tích phương sai đa biến (MANOVA) để kiểm định sự khác biệt giữa các nhóm khách hàng theo các yếu tố như giá trị giao dịch và vòng đời.
- Sử dụng kiểm định hậu nghiệm (post-hoc) để xác định các yếu tố phân biệt rõ ràng giữa các nhóm.
- Kiểm tra độ chính xác của các phân nhóm qua các chỉ số như tỷ lệ rời bỏ, mức độ trung thành, và phản hồi từ chiến lược marketing.
- **Ứng dụng kết quả vào thực tiễn**
 - Đưa ra gợi ý chiến lược giữ chân khách hàng cho từng nhóm.
 - Hỗ trợ doanh nghiệp tối ưu chiến dịch quảng cáo và chương trình khách hàng thân thiết.

3. Mô tả dữ liệu

- Link dataset: [ITBLogisticDataset.xlsx](#)
- Tổng số dòng dữ liệu: 180,519
- Tổng số cột (trường thông tin): 53
- Các trường dữ liệu sử dụng

Trường dữ liệu	Mô tả	Kiểu dữ liệu
Customer Id	Mã khách hàng	Intenger
Order Id	Mã đơn hàng	Intenger
Order Date	Ngày đặt hàng	DateTime
Order Item Total	Tổng giá trị đơn hàng	Float
Days for shipping	Số ngày vận chuyển thực tế của sản phẩm đã mua	Intenger
Customer Segment	Loại khách hàng (Consumer, Corporate, Home Office)	Object
Category Name	Mô tả danh mục sản phẩm	Object
Shipping Mode	Phương thức vận chuyển (Standard Class, First Class, Second Class, Same Day)	Object

II. CƠ SỞ LÝ THUYẾT

1. Phân khúc khách hàng (Customer Segmentation)

Trong bối cảnh cạnh tranh ngày càng gay gắt và nhu cầu tiêu dùng đa dạng, phân cụm khách hàng đã trở thành một chiến lược then chốt giúp doanh nghiệp nâng cao hiệu quả tiếp thị, phát triển sản phẩm phù hợp và tăng trưởng doanh thu (McKinsey, 2016).

Phân cụm khách hàng được định nghĩa là quá trình phân loại khách hàng thành các nhóm có đặc điểm hành vi hoặc nhân khẩu học tương đồng nhằm dễ dàng tiếp cận và phục vụ (Wedel & Kamakura, 2000), dựa trên nền tảng lý thuyết phân khúc thị trường được giới thiệu bởi Smith (1956) và phát triển bởi McDonald & Dunbar (2004). Các tiêu chí phân cụm thường bao gồm biến số chung như độ tuổi, giới tính, thu nhập và biến số đặc thù sản phẩm như hành vi mua sắm, với mô hình RFM (Recency – Frequency – Monetary) là công cụ phân tích hành vi khách hàng phổ biến (Bauer, 1988; Newell, 1997; Tsai & Chiu, 2004).

Sự phát triển của Big Data và AI cho phép việc ứng dụng các thuật toán học máy như K-Means, K-Prototypes, DBSCAN để tự động phát hiện các mẫu hành vi tiềm ẩn (Xu & Wunsch, 2005), hỗ trợ doanh nghiệp khám phá các phân khúc mới mà phương pháp truyền thống khó đạt được. Thực tế cho thấy, các nền tảng thương mại điện tử như Magento (2014) đã xác định phân khúc khách hàng là một phần thiết yếu của chiến lược cá nhân hóa nội dung và dịch vụ. Ngoài ra, nghiên cứu của R. Punhani và cộng sự (2021) cũng chỉ ra rằng việc phân cụm khách hàng hiệu quả, kết hợp với truyền thông đúng thời điểm và đúng đối tượng, góp phần quan trọng vào việc nâng cao tỷ lệ chuyển đổi và xây dựng lợi thế cạnh tranh bền vững cho doanh nghiệp.

2. Phân cụm dữ liệu (Clustering)

Phân cụm dữ liệu (Clustering) là một kỹ thuật trọng yếu trong khai phá dữ liệu (Data Mining), cho phép nhóm các thực thể có đặc điểm tương đồng vào cùng một cụm nhằm phục vụ các mục tiêu như phân đoạn khách hàng, phát hiện bất thường, hoặc tối ưu hóa quy trình vận hành (M. Inaba et al., 1994). Về bản chất, phân cụm là quá trình phân chia tập dữ liệu sao cho các đối tượng trong cùng một cụm có mức độ tương đồng cao nhất có thể, trong khi sự khác biệt giữa các cụm là tối đa. Hai phương pháp phổ biến nhất trong phân cụm là phân cụm phân cấp (Hierarchical Clustering) và phân cụm phân vùng (Partitional Clustering). Phân cụm phân cấp xây dựng cấu trúc cây phân cụm (dendrogram) thông qua việc hợp nhất hoặc phân tách tuần tự các đối tượng, mà không yêu cầu xác định trước số lượng cụm, qua đó giúp khám phá cấu trúc tự nhiên tiềm ẩn trong dữ liệu, mặc dù phương pháp này thường có chi phí tính toán cao khi áp dụng cho tập dữ liệu lớn (Xu & Wunsch, 2005). Ngược lại, phân cụm phân vùng yêu cầu xác định trước số lượng cụm k và tìm cách

tối ưu hóa việc phân nhóm sao cho các đối tượng trong cùng một cụm đồng nhất nhất có thể, với các thuật toán tiêu biểu như K-Means, K-Mode và K-Means++ nhằm cải thiện hiệu quả khởi tạo điểm cụm (Omran et al., 2007). Mặc dù có ưu thế về tốc độ xử lý và khả năng mở rộng đối với dữ liệu lớn, phương pháp này lại nhạy cảm với nhiễu và khó xử lý các cụm có hình dạng phức tạp. Ngoài ra, sự phát triển của dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI) đã thúc đẩy việc áp dụng các kỹ thuật phân cụm tiên tiến hơn như DBSCAN hay Gaussian Mixture Models, vốn cho phép nhận diện các cấu trúc dữ liệu phi tuyến tính và không đồng nhất (Xu & Wunsch, 2005). Hiện nay, phân cụm dữ liệu đóng vai trò thiết yếu trong nhiều lĩnh vực thực tiễn như thương mại điện tử, tài chính, y tế và tiếp thị cá nhân hóa; tiêu biểu như Magento (2014) đã nhấn mạnh tầm quan trọng của phân cụm khách hàng trong việc tối ưu hóa chiến lược tiếp thị, còn nghiên cứu của R. Punhani và cộng sự (2021) khẳng định rằng việc phân cụm chính xác và truyền thông đúng đối tượng, đúng thời điểm có thể nâng cao đáng kể tỷ lệ chuyển đổi và xây dựng lợi thế cạnh tranh dài hạn cho doanh nghiệp.

3. Mô hình RFM (RFM Model)

Mô hình RFM lần đầu tiên được đề xuất bởi Hughes (1996) nhằm phân tích và dự đoán hành vi của khách hàng (Hughes, 1996). Mô hình RFM cơ bản dựa trên ba thuộc tính chính là Recency (R), Frequency (F) và Monetary (M).

Recency là khoảng thời gian kể từ lần mua hàng gần nhất (tính bằng ngày hoặc tháng) và cung cấp thông tin về khả năng khách hàng sẽ mua lại hoặc quay trở lại trong tương lai. Khoảng thời gian này càng ngắn thì khả năng khách hàng tiếp tục mua sắm hoặc truy cập lại càng cao.

$$Recency = date_diff(t_{snapshot}, t_{last})$$

Trong đó:

- + $t_{snapshot}$: thời điểm phân tích - ngày quan sát (thường là ngày cuối trong tập dữ liệu + 1)
- + t_{last} : ngày giao dịch gần nhất của khách hàng trong khoảng thời gian quan sát
- + $date_diff(a, b)$: số ngày giữa hai thời điểm a và b, tức là a-b
- Frequency là số lần mua hàng hoặc truy cập trong một khoảng thời gian nhất định và thể hiện mức độ trung thành của khách hàng. Tần suất càng cao thì mức độ trung thành của khách hàng càng lớn.

$$Frequency = count(Ordery)$$

Trong đó:

- + : số lượng đơn hàng của khách hàng i trong khoảng thời gian quan sát
- + : hàm đếm số giao dịch
- Monetary là tổng số tiền mà khách hàng đã chi tiêu hoặc mức chi tiêu trung bình mỗi lần mua trong một khoảng thời gian nhất định, qua đó phản ánh mức độ đóng

góp của khách hàng vào doanh thu của doanh nghiệp. Số tiền chi tiêu càng lớn thì khách hàng càng đóng góp nhiều vào doanh thu.

$$Monetary = \sum_{j=1}^n Amount_{ij}$$

Trong đó:

- + $Amount_{ij}$: là giá trị của đơn hàng thứ j của khách hàng i
- + n : tổng số đơn hàng của khách hàng đó

Một yếu tố học thuật quan trọng khi triển khai mô hình RFM là vấn đề phân bổ trọng số cho các biến thành phần. Theo Hughes (2005), có quan điểm cho rằng các yếu tố Recency, Frequency và Monetary đều có mức độ ảnh hưởng tương đương đến hành vi khách hàng, do đó nên được gán trọng số như nhau trong quá trình phân tích.

Tuy nhiên, một số nghiên cứu khác lại đề xuất rằng mức độ tác động của từng biến có thể khác nhau tùy theo đặc điểm của ngành hàng hoặc lĩnh vực ứng dụng cụ thể. Theo Stone và Jacobs (2007), trong một số ngành, yếu tố Recency có thể đóng vai trò chi phối hơn Frequency hoặc Monetary, trong khi ở các lĩnh vực khác, thứ tự ưu tiên có thể đảo ngược. Từ đó, việc điều chỉnh trọng số linh hoạt được xem là cần thiết để đảm bảo độ phù hợp và hiệu quả của mô hình trong thực tiễn triển khai. Trong những năm gần đây, mô hình RFM đã được nhiều nghiên cứu mở rộng nhằm khắc phục các giới hạn của cấu trúc gốc và nâng cao khả năng mô tả hành vi tiêu dùng trong các bối cảnh cụ thể. Các mô hình mở rộng này thường kết hợp thêm các biến số bổ sung, áp dụng thuật toán phân cụm hiện đại hoặc sử dụng kỹ thuật gán trọng số nhằm tăng độ chính xác và tính linh hoạt trong phân đoạn khách hàng. Một trong những cải tiến phổ biến là:

- **Mô hình Weighted RFM**, trong đó ba thành phần R, F, M được gán trọng số dựa trên mức độ quan trọng tương đối, thay vì giả định rằng chúng có ảnh hưởng ngang nhau. Wei et al. (2010) đã áp dụng phương pháp phân tích thứ bậc (AHP) để xác định trọng số cho từng thành phần, qua đó cải thiện độ chính xác trong phân nhóm khách hàng so với mô hình RFM nguyên bản.
- **Cheng và Chen (2009)** đã tích hợp mô hình RFM với thuật toán **K-Means**, cho thấy sự kết hợp này mang lại hiệu quả cao hơn trong việc trích xuất quy luật phân đoạn khách hàng, đồng thời cải thiện hiệu quả triển khai chiến lược CRM.
- **Wei et al. (2012)** đã đề xuất mô hình **LRFM**, bổ sung thêm biến **Length** – khoảng cách thời gian giữa lần mua đầu tiên và lần gần nhất của khách hàng. Nhằm khắc phục hạn chế của Recency trong việc phản ánh mối quan hệ dài hạn. Nhờ đó, mô hình có thể phân biệt được khách hàng theo vòng đời quan hệ (short-life vs. long-life customers), từ đó hỗ trợ doanh nghiệp định hướng chiến lược chăm sóc phù hợp hơn.
- Một số nghiên cứu khác đã mở rộng RFM theo hướng kết hợp với yếu tố sản phẩm. **Heldt et al. (2021)** phát triển mô hình **RFM/P**, trong đó hành vi mua hàng được

phân tích riêng trên từng sản phẩm và sau đó tổng hợp lại để đánh giá giá trị khách hàng một cách chi tiết hơn.

- **Allegue et al. (2020)** đề xuất mô hình **RFMC**, bổ sung yếu tố **Category** (nhóm sản phẩm) để phân nhóm khách hàng dựa trên loại hàng tiêu dùng. Kết quả nghiên cứu cho thấy mô hình này cho phép doanh nghiệp cung cấp ưu đãi và dịch vụ cá nhân hóa sát với hành vi thực tế hơn.
- **Moghaddam et al. (2017)** đã giới thiệu mô hình **RFMV**, trong đó biến **V (Variety)** thể hiện độ đa dạng sản phẩm trong giỏ hàng của khách hàng. Nghiên cứu này khẳng định rằng sự đa dạng tiêu dùng là yếu tố quan trọng để mô tả hành vi khách hàng và góp phần nâng cao độ chính xác khi phân đoạn, đặc biệt trong các môi trường B2B.

Tổng thể, các mô hình RFM mở rộng đã và đang chứng minh tính hiệu quả trong việc thích nghi với những đặc thù ngành nghề khác nhau, đồng thời cung cấp công cụ phân tích hành vi khách hàng sâu sắc và linh hoạt hơn cho doanh nghiệp.

4. Mô hình LRFM (LRFM Model)

Mô hình RFM (Recency – Frequency – Monetary) là một công cụ truyền thống được sử dụng rộng rãi trong phân tích hành vi khách hàng. Tuy nhiên, trước yêu cầu ngày càng cao về việc phân khúc khách hàng một cách sâu sắc và chính xác hơn, nhiều nhà nghiên cứu đã mở rộng mô hình này bằng cách bổ sung thêm các biến phản ánh vòng đời và hành vi tiêu dùng đặc thù. Một trong những hướng mở rộng tiêu biểu là mô hình LRFM, với L (Length) là biến mới được thêm vào để đại diện cho vòng đời khách hàng.

- *Length(L)* – Độ dài mối quan hệ (Customer Lifetime Length): Là khoảng cách thời gian giữa lần mua hàng đầu tiên và lần mua hàng gần nhất trong khoảng thời gian quan sát. Biến này cho thấy khách hàng đã duy trì tương tác với doanh nghiệp trong bao lâu – từ đó giúp phân biệt giữa short-life customers (khách hàng chỉ mua trong thời gian ngắn) và long-life customers (khách hàng có sự trung thành lâu dài).

$$Length = date_diff(t_{last}, t_{first})$$

Trong đó:

- + t_{first} : ngày giao dịch đầu tiên của khách hàng trong khoảng thời gian quan sát
- + t_{last} : ngày giao dịch gần nhất của khách hàng trong khoảng thời gian quan sát
- + $date_diff(a,b)$: số ngày giữa hai thời điểm a và b, tức là a-b

Việc bổ sung biến L giúp mô hình LRFM vượt trội hơn mô hình RFM truyền thống trong việc nhận diện khách hàng trung thành và có giá trị dài hạn. Mô hình RFM truyền thống chủ yếu đánh giá hành vi mua hàng của khách hàng tại một thời điểm nhất định (snapshot) (Fader & Hardie, 2009; Procedia Computer Science, 2018), trong đó chỉ số Recency (R) phản ánh khoảng thời gian kể từ lần mua hàng gần nhất đến thời điểm phân tích. Tuy nhiên, cách tiếp cận này chưa thể hiện được đầy đủ vòng đời gắn bó của khách hàng với doanh nghiệp. Do đó, việc bổ sung chỉ số Length (L) – đo khoảng cách giữa lần mua đầu tiên và lần mua gần nhất – trong mô hình LRFM giúp mở rộng phạm vi phân tích từ lát cắt hiện tại sang toàn bộ hành

trình khách hàng, mang lại cái nhìn toàn diện hơn về mức độ trung thành và giá trị lâu dài của họ.

Wei và cộng sự (2012) là một trong những nghiên cứu điển hình đã áp dụng mô hình LRFM trong phân đoạn thị trường nha khoa cho trẻ nhỏ, và cho thấy việc bổ sung chiều Length đã giúp cải thiện đáng kể độ chính xác trong phân khúc khách hàng cũng như nâng cao hiệu quả của các chương trình chăm sóc khách hàng.

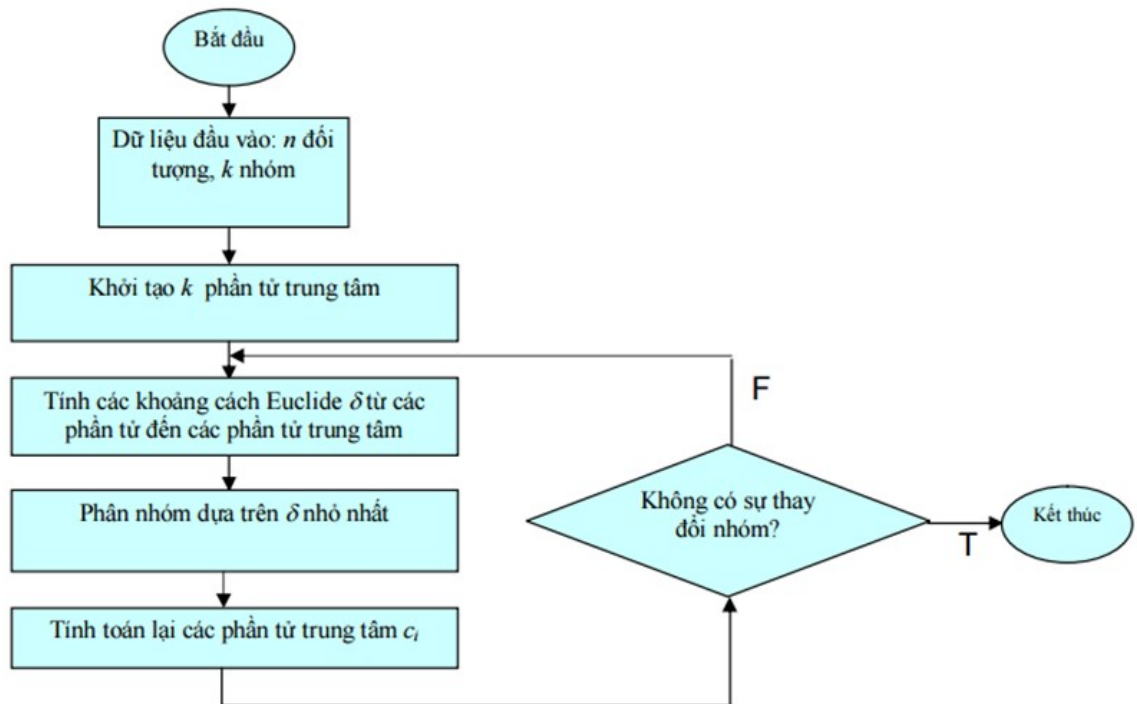
Trong bối cảnh doanh nghiệp cần tập trung vào phân tích vòng đời và giá trị khách hàng (Customer Lifetime Value – CLV), việc sử dụng mô hình LRFM là hoàn toàn phù hợp. Biến Length đóng vai trò quan trọng trong việc đo lường thời gian gắn bó và tiềm năng phát triển dài hạn của từng khách hàng, từ đó giúp doanh nghiệp có chiến lược phù hợp như: giữ chân khách hàng cũ, tái kích hoạt khách hàng ngủ đông, hay ưu tiên chăm sóc các khách hàng dài hạn.

5. Thuật toán K-Means

K-Means là một thuật toán phân cụm phổ biến thuộc nhóm phân vùng (Partitional Clustering), được giới thiệu lần đầu bởi MacQueen vào năm 1967. Thuật toán này được sử dụng rộng rãi trong lĩnh vực khai phá dữ liệu (Data Mining) và học máy (Machine Learning) nhờ tính đơn giản, hiệu quả và khả năng mở rộng với các tập dữ liệu lớn (MacQueen, 1967; Jain, 2010). Mục tiêu chính của K-Means là chia tập dữ liệu thành k cụm sao cho tổng phương sai nội cụm (Within-Cluster Sum of Squares – WCSS) được tối thiểu hóa.

Thuật toán hoạt động thông qua quá trình lặp gồm hai bước: (1) gán mỗi điểm dữ liệu vào cụm có tâm gần nhất, thường sử dụng khoảng cách Euclid, và (2) cập nhật lại tâm cụm bằng trung bình cộng của tất cả các điểm trong cụm. Quá trình lặp tiếp tục cho đến khi các cụm hội tụ (không thay đổi đáng kể) hoặc đạt đến số vòng lặp tối đa được định trước (Lloyd, 1982).

Quy trình thuật toán:



Hình 1: Quy trình hoạt động của thuật toán K-Means

Bước 1: Chọn số cụm k cần phân chia, tùy theo yêu cầu hoặc được xác định thông qua các phương pháp đánh giá như Elbow Method hoặc Silhouette Score.

Bước 2: Khởi tạo ngẫu nhiên k tâm cụm (centroids) ban đầu. Các tâm cụm này là các điểm trong không gian dữ liệu và sẽ được cập nhật qua từng vòng lặp.

Bước 3: Với mỗi điểm dữ liệu trong tập, gán điểm đó vào cụm có tâm gần nhất, sử dụng khoảng cách Euclid làm thước đo:

$$d(x_i, \mu_j) = \sqrt{(x_i - \mu_j)^2}$$

Trong đó:

- + x_i : điểm dữ liệu
- + μ_j : tâm cụm của thứ j

Bước 4: Sau khi tất cả các điểm dữ liệu được gán vào cụm, tính lại tâm cụm mới bằng cách lấy trung bình cộng của tất cả các điểm trong cụm:

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i$$

Trong đó: n_j : là số điểm trong cụm j

Bước 5: Lặp lại hai bước gán cụm và cập nhật tâm cụm (bước 3 và 4) cho đến khi hội tụ, tức là khi không còn sự thay đổi đáng kể trong việc gán cụm giữa hai lần lặp liên tiếp, hoặc khi đạt đến số lần lặp tối đa được định trước

Bước 6: Tính tổng sai số bình phương trong cụm:

$$WCSS = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

Trong đó:

- + k : số cụm
- + C_j : cụm thứ j
- + x_i : điểm dữ liệu thuộc cụm j
- + $\|x_i - \mu_j\|^2$: bình phương khoảng cách Euclidean giữa điểm dữ liệu và tâm cụm

Ưu điểm của K-Means là đơn giản, dễ triển khai, và có khả năng xử lý hiệu quả với dữ liệu lớn. Tuy nhiên, thuật toán cũng có những hạn chế: phải xác định trước số cụm k , nhạy cảm với điểm khởi tạo, dễ rơi vào cực trị cục bộ và không phù hợp với dữ liệu có cụm hình dạng phức tạp hoặc có outliers (Jain, 2010; Arthur & Vassilvitskii, 2007).

Để khắc phục các hạn chế này, một số biến thể đã được đề xuất như K-Means++ nhằm cải thiện quá trình khởi tạo tâm cụm để tăng tính ổn định và chất lượng phân cụm (Arthur & Vassilvitskii, 2007).

Hiện nay, K-Means được ứng dụng rộng rãi trong nhiều lĩnh vực như: phân đoạn khách hàng (customer segmentation), phân tích ảnh (image segmentation), phát hiện bất thường (anomaly detection), và mô hình hóa chủ đề trong xử lý ngôn ngữ tự nhiên (text clustering) (Tan et al., 2018).

6. Kiểm định giả thuyết (Hypothesis Testing)

Kiểm định giả thuyết là một phương pháp thống kê cơ bản được sử dụng để đưa ra suy luận về đặc điểm của một quần thể thông qua dữ liệu mẫu thu thập được.

Phương pháp này liên quan đến việc đánh giá hai giả thuyết cạnh tranh: giả thuyết không (H_0) và giả thuyết thay thế (H_1). Giả thuyết không (H_0) đóng vai trò là giả định ban đầu, thường phát biểu rằng không có sự khác biệt hoặc không tồn tại mối liên hệ đáng kể giữa các biến trong quần thể, trong khi giả thuyết thay thế (H_1) lại phản ánh lập trường ngược lại, cho rằng có sự khác biệt hoặc mối quan hệ thực sự tồn tại (Moore et al., 2018).

Quy trình kiểm định giả thuyết bắt đầu bằng việc xây dựng các giả thuyết rõ ràng dựa trên câu hỏi nghiên cứu. Tiếp theo, người nghiên cứu cần xác định mức ý nghĩa thống kê (thường ký hiệu là α), đại diện cho xác suất chấp nhận sai lầm loại I – tức là bác bỏ giả thuyết không trong khi nó đúng. Sau đó, dữ liệu được thu thập và xử lý, từ đó tính toán các giá trị thống kê phù hợp như giá trị Z , t , F hoặc chi bình phương, tùy thuộc vào loại dữ liệu và mô hình nghiên cứu. Cuối cùng, dựa trên kết quả so sánh giữa thống kê kiểm định và giá trị tới hạn hoặc giá trị p , người nghiên cứu sẽ đưa ra quyết định có nên bác bỏ giả thuyết không hay không (Field, 2013). Tùy vào cấu trúc dữ liệu và mục tiêu nghiên cứu, nhiều loại kiểm định giả thuyết khác nhau đã được phát triển và ứng dụng. Trong đó, kiểm định một mẫu được sử

dùng khi so sánh đặc điểm của một mẫu với một giá trị chuẩn hoặc giá trị kỳ vọng đã biết. Kiểm định hai mẫu độc lập thường áp dụng để đánh giá sự khác biệt giữa hai nhóm không liên quan, trong khi kiểm định mẫu cặp được dùng cho các trường hợp dữ liệu có mối liên hệ, chẳng hạn như trước và sau khi can thiệp. Ngoài ra, kiểm định chi bình phương được ứng dụng phổ biến trong việc phân tích dữ liệu phân loại (categorical), giúp kiểm tra mối quan hệ giữa hai hoặc nhiều biến định danh. Trong bối cảnh nghiên cứu so sánh trung bình giữa nhiều nhóm, phân tích phương sai (ANOVA) là một công cụ mạnh mẽ cho phép xác định xem có sự khác biệt đáng kể giữa các nhóm hay không mà không cần thực hiện nhiều kiểm định t riêng lẻ (Sharma, 1996).

Việc lựa chọn loại kiểm định phù hợp cần dựa trên đặc điểm của dữ liệu, số lượng nhóm so sánh, bản chất của biến phụ thuộc (định lượng hay định tính), và giả định về phân phối chuẩn của dữ liệu. Kiểm định giả thuyết, xét về bản chất, không chỉ là công cụ thống kê mà còn là nền tảng logic để biến các quan sát mô tả trở thành kết luận khoa học có giá trị thực tiễn. Chính vì thế, kiểm định giả thuyết thường được sử dụng như một bước bắt buộc trong các nghiên cứu ứng dụng mô hình học máy hoặc phân tích hành vi khách hàng, để xác minh tính hợp lý của các phân nhóm, dự đoán hoặc giải thích hành vi trong bối cảnh dữ liệu thực tế (Hair et al., 2010; Sorkun et al., 2022).

7. Phân tích phương sai đa biến (Multivariate Analysis of Variance - MANOVA)

Phân tích phương sai đa biến (MANOVA – Multivariate Analysis of Variance) là một phương pháp thống kê mở rộng từ ANOVA, cho phép kiểm định đồng thời nhiều biến phụ thuộc định lượng nhằm đánh giá sự khác biệt giữa các nhóm thuộc một hoặc nhiều biến độc lập. MANOVA đặc biệt phù hợp trong các tình huống nghiên cứu mà các biến phụ thuộc có thể tồn tại mối tương quan với nhau, nhờ đó không chỉ tăng độ chính xác của kiểm định mà còn hạn chế nguy cơ mắc sai lầm loại I khi phải thực hiện nhiều phép ANOVA riêng lẻ (Tabachnick & Fidell, 2007). Về nguyên lý, MANOVA kiểm định giả thuyết không (H_0) rằng trung bình của tất cả các biến phụ thuộc là đồng nhất giữa các nhóm, và giả thuyết thay thế (H_1) rằng có ít nhất một sự khác biệt có ý nghĩa giữa các nhóm này. Để đưa ra quyết định thống kê, MANOVA sử dụng một số chỉ số như Wilks' Lambda, Pillai's Trace hoặc Roy's Largest Root – trong đó Wilks' Lambda là phổ biến nhất và thường được chọn làm căn cứ chính để đánh giá mức độ khác biệt tổng thể (Stevens, 2009). Kết quả của MANOVA cung cấp cơ sở thống kê để xác định liệu biến độc lập (chẳng hạn như nhân cụm trong bài toán phân nhóm khách hàng) có ảnh hưởng đồng thời đến các biến phụ thuộc hay không.

Trong lĩnh vực phân tích hành vi khách hàng, MANOVA đã được áp dụng như một phương pháp kiểm định hiệu lực của mô hình phân cụm thông qua việc đánh giá mức độ khác biệt có ý nghĩa thống kê giữa các nhóm khách hàng trên tổ hợp các chỉ số hành vi. Đáng chú ý, nghiên cứu của Wei et al. (2012) trong bối cảnh phân đoạn khách hàng y tế đã sử dụng MANOVA để kiểm định sự khác biệt giữa các cụm

khách hàng trên bốn biến chính: Length, Recency, Frequency và Monetary. Kết quả cho thấy mô hình phân cụm là có ý nghĩa thống kê, và từ đó trở thành cơ sở khoa học để đề xuất các chiến lược tiếp thị phân khúc hiệu quả. Một hướng ứng dụng tương tự cũng xuất hiện trong nghiên cứu của Ghasemzadeh và cộng sự (2021), khi nhóm tác giả sử dụng MANOVA để kiểm định các cụm khách hàng trong phân tích giá trị vòng đời (CLV) dựa trên mô hình mở rộng RFM.

Tóm lại, MANOVA là một công cụ thống kê mạnh mẽ, vừa giúp xác nhận sự khác biệt tổng thể giữa các nhóm theo nhiều chiều biến số, vừa tạo nền tảng thống kê vững chắc để tiến hành các bước kiểm định sâu hơn như ANOVA hay phân tích hậu định trong các nghiên cứu ứng dụng dữ liệu thực tiễn.

8. Phân tích phương sai một chiều (Analysis of Variance - ANOVA)

Phân tích phương sai một chiều (ANOVA – Analysis of Variance) là một phương pháp thống kê được phát triển nhằm mục tiêu so sánh giá trị trung bình của một biến phụ thuộc định lượng giữa hai hay nhiều nhóm khác nhau. ANOVA giúp xác định liệu sự khác biệt về trung bình giữa các nhóm là kết quả của yếu tố ngẫu nhiên hay là sự khác biệt có ý nghĩa thống kê, xuất phát từ tác động của biến độc lập phân loại. Về nguyên lý, ANOVA kiểm định giả thuyết không (H_0) rằng tất cả các nhóm có trung bình tổng thể bằng nhau, đối lập với giả thuyết thay thế (H_1) cho rằng có ít nhất một cặp nhóm có sự khác biệt thực sự. Phương pháp này dựa trên việc phân tích tổng biến thiên của dữ liệu thành hai phần: biến thiên giữa các nhóm (between-group variance) và biến thiên trong nhóm (within-group variance), từ đó hình thành chỉ số F – là tỷ số giữa phương sai giữa nhóm và phương sai trong nhóm. Nếu giá trị F đủ lớn, giả thuyết không sẽ bị bác bỏ tại một mức ý nghĩa α xác định trước (Field, 2013).

Trước khi áp dụng ANOVA, cần đảm bảo một số giả định cơ bản như: các quan sát là độc lập với nhau, phân phối chuẩn trong mỗi nhóm, và phương sai giữa các nhóm là đồng nhất (homogeneity of variance). Nếu các giả định này bị vi phạm nghiêm trọng, kết quả kiểm định có thể bị sai lệch, đặc biệt là trong các mẫu nhỏ. Tuy vậy, trong nhiều ứng dụng thực tiễn với kích thước mẫu lớn, ANOVA được đánh giá là khá bền vững với các vi phạm giả định nhẹ, đặc biệt nếu sử dụng kèm với các biện pháp kiểm tra như Levene's Test hoặc Welch's ANOVA trong trường hợp phương sai không đồng nhất (Sharma, 1996).

Trong nghiên cứu hành vi khách hàng, ANOVA thường được sử dụng như bước tiếp theo sau MANOVA nhằm kiểm định riêng từng biến hành vi như Length, Recency, Frequency, hay Monetary giữa các nhóm khách hàng được phân cụm. Cách tiếp cận này cho phép xác định cụ thể biến nào là yếu tố phân biệt chính giữa các cụm, từ đó cung cấp dữ liệu đầu vào quan trọng cho việc xây dựng hồ sơ khách hàng mục tiêu và đề xuất chiến lược marketing. Trong nghiên cứu của Wei et al. (2012), ANOVA được sử dụng để kiểm định từng chỉ số trong mô hình LRFM sau khi phân cụm, cho phép xác định rằng sự khác biệt giữa các nhóm khách hàng không chỉ là tổng thể mà còn mang tính đơn biến. Một hướng tiếp cận tương tự cũng được thể hiện trong

ngiên cứu của Christy và cộng sự (2021), khi nhóm tác giả sử dụng ANOVA để kiểm định sự khác biệt hành vi giữa các nhóm khách hàng dựa trên giá trị chi tiêu trung bình, tần suất mua sắm và thời gian gắn bó với thương hiệu.

9. Phân tích hậu định (Post-hoc Analysis)

Trong các nghiên cứu phân tích sự khác biệt giữa nhiều nhóm, việc thực hiện kiểm định ANOVA chỉ cho biết tồn tại sự khác biệt tổng thể nhưng không chỉ ra cụ thể cặp nhóm nào khác biệt với nhau. Để khắc phục hạn chế này, phân tích hậu định (Post Hoc Analysis) được sử dụng nhằm xác định rõ các cặp nhóm có sự khác biệt có ý nghĩa thống kê. Các kỹ thuật phân tích hậu định được thiết kế để kiểm soát tốt hơn xác suất mắc sai lầm loại I khi thực hiện nhiều phép so sánh, từ đó đảm bảo độ tin cậy cho các kết luận rút ra từ dữ liệu thực nghiệm (Field, 2013).

Có nhiều phương pháp phân tích hậu định đã được phát triển, phổ biến nhất gồm Tukey's Honest Significant Difference (Tukey HSD), Bonferroni Correction, Scheffé Test và Duncan's Multiple Range Test. Trong đó, Tukey HSD là phương pháp được sử dụng rộng rãi nhất trong các nghiên cứu ứng dụng vì tính hiệu quả và sự cân bằng giữa kiểm soát chặt chẽ sai lầm loại I và khả năng phát hiện sự khác biệt thực sự. Phương pháp này đặc biệt phù hợp trong các tình huống cần so sánh tất cả các cặp nhóm có thể có, với giả định rằng phương sai giữa các nhóm là đồng nhất và kích thước mẫu tương đối cân đối (Abdi & Williams, 2010).

Kiểm định Tukey HSD hoạt động dựa trên nguyên lý so sánh hiệu số giữa hai trung bình mẫu bất kỳ với sai số chuẩn của hiệu trung bình, thông qua chỉ số kiểm định Q – được tính dựa trên phân phối Studentized Range. Công thức thống kê được sử dụng trong phép kiểm định này được trình bày như sau:

$$Q = \frac{\bar{X}_j - \bar{X}_k}{\sqrt{MSW / n}}$$

Trong đó:

$\bar{X}_j - \bar{X}_k$: là trung bình mẫu của hai nhóm được so sánh

MSW : là phương sai trong nhóm (Mean Square Within Groups) được lấy từ kết quả ANOVA

n : là số lượng quan sát trung bình trên mỗi nhóm.

Sau khi tính được giá trị Q thực nghiệm, ta so sánh nó với giá trị tới hạn lý thuyết $Q_{critical}$ từ bảng phân phối Studentized Range tại mức ý nghĩa α xác định. Nếu Q thực nghiệm lớn hơn $Q_{critical}$, ta kết luận rằng có sự khác biệt có ý nghĩa thống kê giữa hai nhóm được so sánh.

Quy trình thực hiện kiểm định Tukey HSD bao gồm các bước: đầu tiên thực hiện ANOVA tổng thể để xác định sự khác biệt trung bình giữa các nhóm là có ý nghĩa; nếu ANOVA cho kết quả $p\text{-value} < 0.05$, tiến hành phân tích hậu định với Tukey HSD; tính toán giá trị HSD cho từng cặp so sánh; cuối cùng, kết luận những cặp nhóm có sự khác biệt ý nghĩa dựa trên kết quả so sánh này (Sharma, 1996).

Trong nghiên cứu phân tích hành vi khách hàng sau phân cụm, kiểm định Tukey HSD thường được áp dụng để xác định cụ thể các cụm khách hàng khác biệt nhau về từng biến hành vi như Length, Recency, Frequency và Monetary. Cách tiếp cận này đã được chứng minh hiệu quả trong nghiên cứu của Christy et al. (2021) khi đánh giá sự khác biệt chi tiêu trung bình giữa các nhóm khách hàng e-retail, cũng như trong nghiên cứu của Ghasemzadeh et al. (2021) khi phân tích các biến số liên quan đến giá trị vòng đời khách hàng.

10. Phân tích vòng đời giá trị khách hàng (Customer Lifetime Value Analysis)

Customer Lifetime Value (CLV) là một chỉ số quan trọng trong lĩnh vực marketing và quản trị quan hệ khách hàng, phản ánh tổng giá trị lợi nhuận kỳ vọng mà một khách hàng có thể mang lại cho doanh nghiệp trong toàn bộ vòng đời tương tác với thương hiệu (Kotler & Keller, 2016). Phân tích CLV không chỉ giúp doanh nghiệp hiểu được giá trị dài hạn của từng khách hàng mà còn cung cấp nền tảng định lượng để đưa ra các quyết định chiến lược liên quan đến chăm sóc, giữ chân, và phân bổ ngân sách marketing (Gupta & Lehmann, 2003).

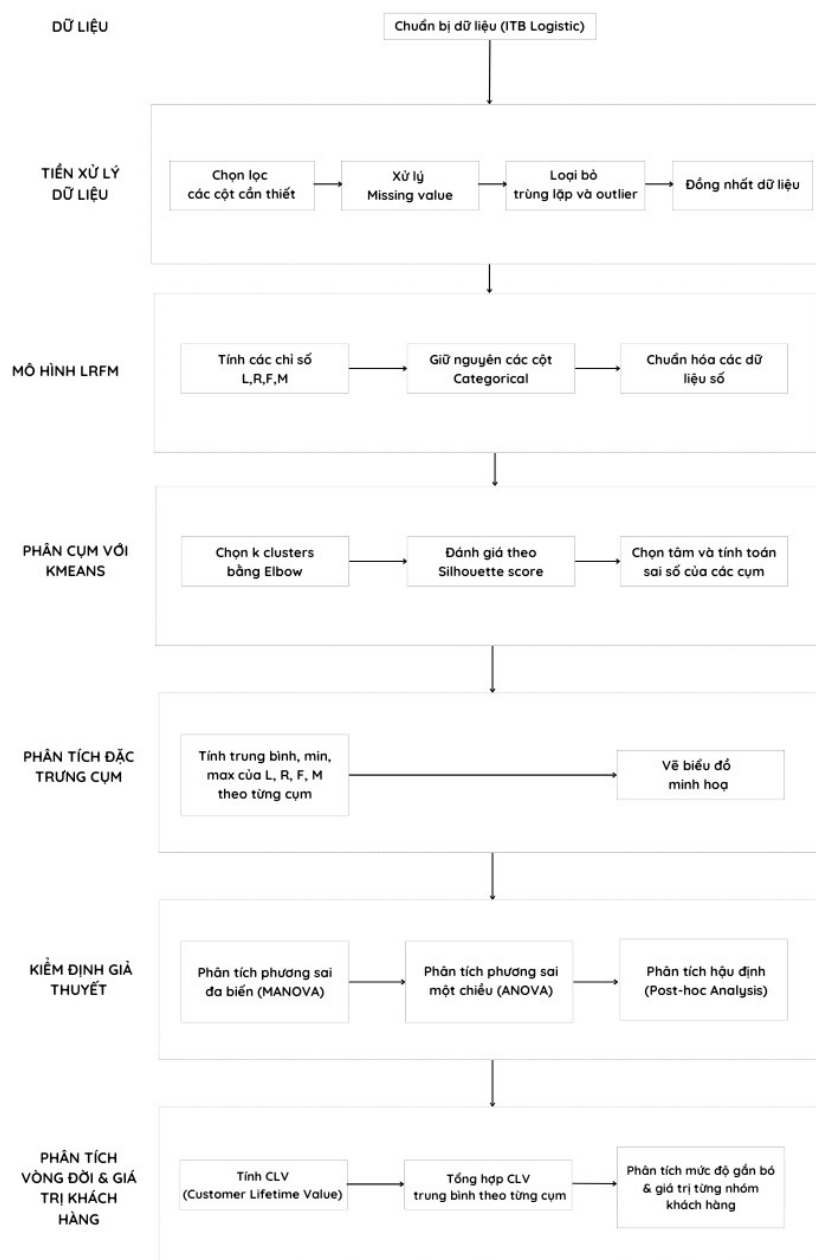
Trong bối cảnh thị trường ngày càng cạnh tranh khốc liệt và chi phí thu hút khách hàng mới không ngừng gia tăng, việc duy trì khách hàng hiện tại – đặc biệt là những khách hàng có giá trị cao – trở thành ưu tiên hàng đầu của doanh nghiệp (Reinartz & Kumar, 2003). Việc phân tích CLV giúp các tổ chức tập trung nguồn lực vào các phân khúc khách hàng có tiềm năng sinh lời cao, giảm tỷ lệ rời bỏ (churn rate), và cá nhân hóa trải nghiệm nhằm gia tăng mức độ trung thành. Ngoài ra, CLV còn là cơ sở để thiết kế chương trình khách hàng thân thiết (loyalty program), xác định nhóm khách hàng ưu tiên trong các chiến dịch bán hàng, cũng như đánh giá hiệu quả đầu tư vào từng nhóm đối tượng.

Trong nghiên cứu này, nhóm thực hiện phân tích CLV dựa trên mô hình **LRFM** (Loyalty – Recency – Frequency – Monetary), một biến thể mở rộng của mô hình RFM truyền thống, có bổ sung thêm yếu tố Loyalty để phản ánh mức độ gắn bó dài hạn của khách hàng. Phân tích được triển khai thông qua ba bước chính:

1. Tính toán CLV cá nhân cho từng khách hàng trên cơ sở các chỉ số trong mô hình LRFM;
2. Phân tích giá trị vòng đời khách hàng theo từng cụm được phân nhóm bằng thuật toán K-means, nhằm nhận diện sự khác biệt về giá trị giữa các phân khúc;
3. Đề xuất các chiến lược chăm sóc và giữ chân khách hàng, tùy theo từng nhóm CLV, từ nhóm khách hàng có giá trị cao đến nhóm có giá trị thấp, để tối ưu hóa hiệu quả nguồn lực và nâng cao lợi nhuận dài hạn cho doanh nghiệp.

Cách tiếp cận này cho phép doanh nghiệp vừa có cái nhìn toàn cảnh về cơ cấu giá trị khách hàng, vừa có khả năng thiết kế các chiến lược marketing cá nhân hóa và hiệu quả hơn trong việc tăng trưởng bền vững.

III. PHƯƠNG PHÁP NGHIÊN CỨU



Hình 2: Quy trình bài toán (Framework)

- Workflow:

- Quy trình bài toán:

(1) Chuẩn bị và tiền xử lý dữ liệu

Dữ liệu được thu thập từ hệ thống bán hàng ITB Logistics, bao gồm thông tin về khách hàng, đơn hàng, sản phẩm và các biến định lượng – định tính. Giai đoạn này bao gồm các bước chọn lọc các cột cần thiết, xử lý giá trị thiếu, loại bỏ trùng lặp và outlier, cũng như đồng nhất định dạng dữ liệu. Đây là bước quan trọng giúp đảm bảo độ sạch và sự phù hợp của dữ liệu trước khi đưa vào mô hình học máy.

(2) Xây dựng mô hình LRFM

Dữ liệu sau khi được làm sạch sẽ được nhóm theo từng khách hàng để tính toán các chỉ số hành vi: Loyalty (L), Recency (R), Frequency (F), và Monetary (M). Các biến phân loại như khu vực địa lý (Region) hoặc phân khúc khách hàng (Segment) được giữ nguyên. Các biến định lượng được chuẩn hóa nhằm đảm bảo tính nhất quán khi đưa vào phân tích phân cụm.

(3) Phân cụm bằng K-Means

Thuật toán K-Means được sử dụng để phân nhóm khách hàng dựa trên các chỉ số hành vi LRFM. Trước tiên, phương pháp Elbow được áp dụng để xác định số lượng cụm tối ưu. Sau đó, thuật toán được triển khai để phân bổ khách hàng vào các cụm dựa trên độ tương đồng. Đánh giá chất lượng phân cụm được thực hiện thông qua chỉ số Silhouette, kết hợp với việc tính toán tâm cụm và sai số nội cụm.

(4) Phân tích đặc trưng cụm

Sau khi phân cụm hoàn tất, các đặc trưng của từng nhóm khách hàng được mô tả thông qua các chỉ số trung bình, giá trị lớn nhất và nhỏ nhất của L, R, F, M. Các biểu đồ minh họa giúp trực quan hóa đặc điểm hành vi của từng cụm khách hàng, làm cơ sở cho bước phân tích sâu hơn.

(5) Kiểm định giả thuyết thống kê

Nhằm kiểm tra sự khác biệt giữa các cụm khách hàng, nhóm nghiên cứu thực hiện kiểm định phương sai. Phân tích phương sai đa biến (MANOVA) được sử dụng để đánh giá sự khác biệt tổng thể, tiếp theo là ANOVA để phân tích từng biến độc lập. Cuối cùng, kiểm định hậu định (Post-hoc Analysis) giúp xác định cặp cụm nào có sự khác biệt có ý nghĩa thống kê.

(6) Phân tích vòng đời và giá trị khách hàng

Giai đoạn cuối cùng tập trung vào phân tích giá trị vòng đời khách hàng (Customer Lifetime Value – CLV). CLV được tính cho từng khách hàng, sau đó tổng hợp theo cụm để xác định nhóm khách hàng có giá trị cao nhất. Trên cơ sở đó, nhóm thực hiện đánh giá mức độ gắn bó và đề xuất chiến lược chăm sóc phù hợp cho từng nhóm, nhằm tối ưu hóa hiệu quả kinh doanh và nâng cao khả năng giữ chân khách hàng.

1. Data Preparation

Tập dữ liệu được sử dụng trong nghiên cứu được thu thập từ hệ thống giao dịch bán hàng trực tuyến của công ty DataCo Global trong giai đoạn từ năm 2011 đến 2014. Dữ liệu ghi nhận hơn 180.000 giao dịch liên quan đến hoạt động mua bán và vận chuyển hàng hóa tại nhiều quốc gia khác nhau. Các sản phẩm trong bộ dữ liệu được phân loại thành ba nhóm chính gồm: Điện tử (Electronics), Đồ gia dụng (Home Appliances) và Thời trang (Fashion).

Mỗi bản ghi trong tập dữ liệu bao gồm các thông tin chi tiết như: ngày đặt hàng, ngày vận chuyển, phương thức thanh toán, phương thức vận chuyển, cũng như thông tin khách hàng. Mặc dù dữ liệu không chứa giá trị bị thiếu nghiêm trọng, một

số trường có thể cần được xử lý bổ sung để chuẩn hóa định dạng và đảm bảo tính nhất quán cho các bước phân tích sau đó.

- 5 dòng dữ liệu đầu tiên của bộ dữ liệu:

	Type	Days for shipping (real)	Days for shipment (scheduled)	\
0	DEBIT	3	4	
1	TRANSFER	5	4	
2	CASH	4	4	
3	DEBIT	3	4	
4	PAYMENT	2	4	

	Benefit per order	Sales per customer	Delivery Status	\
0	91.250000	314.640015	Advance shipping	
1	-249.089996	311.359985	Late delivery	
2	-247.779999	309.720001	Shipping on time	
3	22.860001	304.809998	Advance shipping	
4	134.210007	298.250000	Advance shipping	

	Late_delivery_risk	Category Id	Category Name	Customer City	...	\
0	0	73	Sporting Goods	Caguas	...	
1	1	73	Sporting Goods	Caguas	...	
2	0	73	Sporting Goods	San Jose	...	
3	0	73	Sporting Goods	Los Angeles	...	
4	0	73	Sporting Goods	Caguas	...	

	Order	Zipcode	Product Card Id	Product Category Id	Product Description	\
0		NaN	1360	73	NaN	
1		NaN	1360	73	NaN	
2		NaN	1360	73	NaN	
...						
3		0	2018-01-16 11:45:00	Standard Class		
4		0	2018-01-15 11:24:00	Standard Class		

[5 rows x 53 columns]

Hình 3: Dữ liệu gốc của bài toán

Sau đó sẽ tới bước tiền hành và chọn lọc các cột dữ liệu cần thiết cho bài toán LRFM, kiểm định và CLV. Nhóm đã quyết định và chọn 8 trường cần thiết và dưới đây là kết quả:

```

Dữ liệu sau khi chọn cột:
  Customer Id  Order Id  order date (DateOrders)  Order Item Total  \
0      20755      77202      2015-01-01 00:00:00      314.640015
1      19492      75939      2015-01-01 00:21:00      311.359985
2      19491      75938      2015-01-01 00:21:00      309.720001
3      19490      75937      2015-01-01 00:21:00      304.809998
4      19489      75936      2015-01-01 01:03:00      298.250000

  Days for shipping (real)  Customer Segment  Category Name  Shipping Mode
0                      3      Consumer  Sporting Goods  Standard Class
1                      5      Consumer  Sporting Goods  Standard Class
2                      4      Consumer  Sporting Goods  Standard Class
3                      3  Home Office  Sporting Goods  Standard Class
4                      2  Corporate   Sporting Goods  Standard Class
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180519 entries, 0 to 180518
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Customer Id                          180519 non-null  int64
1   Order Id                             180519 non-null  int64
2   order date (DateOrders)              180519 non-null  object
3   Order Item Total                     180519 non-null  float64
4   Days for shipping (real)             180519 non-null  int64
5   Customer Segment                    180519 non-null  object
6   Category Name                       180519 non-null  object
7   Shipping Mode                       180519 non-null  object
dtypes: float64(1), int64(3), object(4)
Checking missing values:
Customer Id      0
Order Id         0
order date (DateOrders)  0
Order Item Total  0
Days for shipping (real)  0
Customer Segment  0
Category Name     0
Shipping Mode     0
dtype: int64

```

Hình 4: Dữ liệu sau khi đã tiền xử lý

Sau đó để phân tích hành vi khách hàng và phục vụ cho các bước phân cụm, mô hình **LRFM** (Length – Recency – Frequency – Monetary) được áp dụng. Đây là mô hình mở rộng từ RFM truyền thống, trong đó chỉ số Length đại diện cho độ dài mối quan hệ giữa khách hàng và doanh nghiệp, được tính dựa trên thời gian giữa lần mua đầu tiên và lần mua gần nhất.

Tập dữ liệu sau khi lọc giữ lại 8 cột chính: Customer Id, Order Id, order date (DateOrders), Order Item Total, Days for shipping (real), Customer Segment, Category Name và Shipping Mode, với tổng cộng **180.519 giao dịch**. Mỗi dòng trong dữ liệu gốc phản ánh một đơn hàng cụ thể được thực hiện bởi một khách hàng.

Đầu tiên, dữ liệu được tiền xử lý bằng cách chuẩn hóa cột order date (DateOrders) sang định dạng thời gian (datetime) để phục vụ cho việc tính toán các chỉ số thời gian. Tiếp theo, dữ liệu được nhóm theo Customer Id để chuyển đổi từ cấp độ đơn hàng sang cấp độ khách hàng. Từ đây, bốn chỉ số hành vi được tính toán như sau:

- **Length (L):** Khoảng cách thời gian giữa đơn hàng đầu tiên và đơn hàng gần nhất của mỗi khách hàng. Chỉ số này phản ánh độ dài mối quan hệ mà khách hàng duy trì với doanh nghiệp.
- **Recency (R):** Khoảng thời gian từ lần mua gần nhất của khách hàng đến **ngày tham chiếu** (ngày cuối cùng trong tập dữ liệu). Chỉ số này biểu thị mức độ cập nhật trong hành vi mua hàng.
- **Frequency (F):** Tổng số đơn hàng mà khách hàng đã thực hiện, được tính bằng số lần xuất hiện của Order Id theo từng Customer Id.
- **Monetary (M):** Tổng giá trị mà khách hàng đã chi tiêu, tính bằng tổng Order Item Total của từng khách hàng trong toàn bộ giai đoạn phân tích.

Các biến định tính như Customer Segment, Category Name và Shipping Mode cũng được giữ lại và mã hóa (encoding) để phục vụ cho các phân tích thống kê và thuật toán phân cụm sau này. Bên cạnh đó, các khách hàng có tổng giá trị giao dịch bằng 0 (Monetary = 0) được loại khỏi tập dữ liệu nhằm đảm bảo chất lượng phân tích và tránh sai lệch trong việc đánh giá giá trị khách hàng.

Kết quả của quá trình xử lý là một bảng dữ liệu tổng hợp, trong đó mỗi khách hàng được biểu diễn bằng một dòng duy nhất với đầy đủ các chỉ số LRFM và thông tin định tính đã chuẩn hóa. Dữ liệu này là đầu vào cho các bước phân cụm, kiểm định và phân tích vòng đời khách hàng trong các phần tiếp theo của nghiên cứu.

- Dưới đây là bảng dữ liệu sau khi đã tính LRFM:

Analysis date: 2018-02-01

LRFM Table Sample:

	CustomerID	Recency	Length	Frequency	Monetary
0	1	1074	0	1	472.450012
1	2	286	787	4	1618.660042
2	3	14	991	5	3189.200037
3	4	33	1039	4	1480.709993
4	5	262	826	3	1101.919998

Hình 5: Các chỉ số L,R,F,M sau khi tính toán

	Recency	Length	Frequency	Monetary
count	20652.000000	20652.000000	20652.000000	20652.000000
mean	410.239008	430.696252	3.183808	1600.535175
std	322.033628	402.827280	2.430699	1508.414500
min	1.000000	0.000000	1.000000	8.470000
25%	161.000000	0.000000	1.000000	254.940002
50%	310.000000	495.000000	3.000000	1294.504997
75%	622.000000	798.000000	5.000000	2621.140007
max	1127.000000	1122.000000	15.000000	9436.610088

Hình 6: Thống kê mô tả của các chỉ số L,R,F,M

Sau khi hoàn thành mô hình LRFM, nhóm sẽ chuẩn hóa dữ liệu và dưới đây là kết quả:

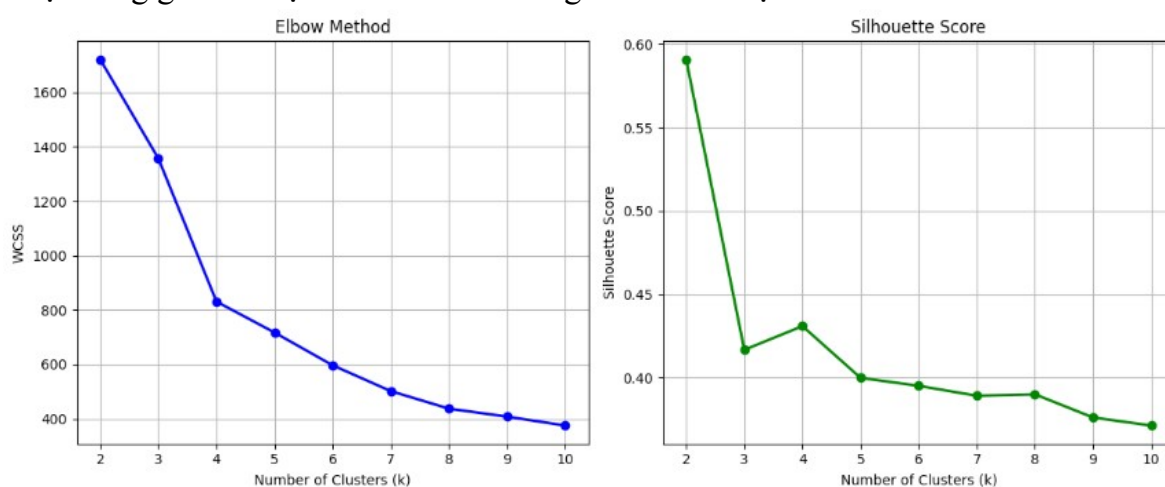
	CustomerID	Recency	Length	Frequency	Monetary
0	1	0.952931	0.000000	0.000000	0.049212
1	2	0.253108	0.701426	0.214286	0.170786
2	3	0.011545	0.883244	0.285714	0.337366
3	4	0.028419	0.926025	0.214286	0.156154
4	5	0.231794	0.736185	0.142857	0.115977

Hình 7: Các chỉ số L,R,F,M sau khi đã được chuẩn hóa

Sau khi đã chuẩn hóa dữ liệu và đồng nhất dữ liệu sẽ được đưa vào mô hình K-means để phân cụm và từ đó xác định được từng cụm khách hàng khác nhau.

2. Clustering with K-means

Biểu đồ Elbow thể hiện sự thay đổi của tổng phương sai nội cụm (Within-Cluster Sum of Squares – WCSS) theo số lượng cụm từ 2 đến 10. Đồng thời, biểu đồ Silhouette Score mô tả mức độ gắn kết và phân tách giữa các cụm theo từng giá trị k. Giá trị silhouette càng cao (gần 1) thì chất lượng phân cụm càng tốt – các điểm dữ liệu càng gần với cụm của mình và càng xa với các cụm khác.



Hình 8: Đồ thị của Elbow Method và Silhouette Score

Dựa trên biểu đồ Elbow, có thể thấy điểm gãy (elbow point) xuất hiện rõ tại k = 4 hoặc k = 5, tuy nhiên mức giảm WCSS bắt đầu chậm lại rõ rệt từ k = 4, cho thấy đây là số cụm hợp lý để đảm bảo mô hình ổn định.

Song song đó, biểu đồ Silhouette Score cho thấy giá trị cao nhất ở k = 2, tuy nhiên số cụm này quá ít và không phản ánh đầy đủ hành vi khách hàng. Trong khoảng từ k = 3 đến k = 5, k = 4 có silhouette score cao hơn các mức còn lại, cho thấy khả năng phân tách và đồng nhất giữa các cụm tốt hơn.

Do đó, nhóm quyết định chọn số lượng nhóm tối ưu là k = 4, đảm bảo sự cân bằng giữa chất lượng phân cụm và khả năng giải thích mô hình

Sau khi áp dụng mô hình K-Means để phân cụm khách hàng dựa trên các chỉ số LRFM đã được chuẩn hóa, dữ liệu được chia thành bốn cụm chính.

Phân cụm K-Means hoàn tất! Dữ liệu mẫu:

	CustomerID	Recency	Length	Frequency	Monetary	Cluster
0	1	1074	0	1	472.450012	1
1	2	286	787	4	1618.660042	2
2	3	14	991	5	3189.200037	0
3	4	33	1039	4	1480.709993	0
4	5	262	826	3	1101.919998	2

Hình 9: Kết quả sau khi đã được phân cụm

Bảng tổng hợp LRFM + số lượng khách:

	Cluster	Num_Customers	Mean_Recency	Mean_Length	Mean_Frequency	Mean_Monetary
0	0	5231	128.560	913.752	6.210	3481.299
1	1	5969	852.257	12.599	1.072	305.145
2	2	6052	286.644	650.269	3.787	1939.514
3	3	3400	287.609	30.669	1.162	377.714

Hình 10: Tổng hợp số lượng khách hàng vào từng cụm và các chỉ số của cụm

3. Hypothesis testing

3.1 Phân tích phương sai đa biến (Multivariate Analysis of Variance - MANOVA)

- Xây dựng giả thuyết:

- + Giả thuyết H_0 : Không có sự khác biệt ý nghĩa nào giữa các nhóm khách hàng đối với các biến phụ thuộc (Length, Recency, Frequency, Monetary)
- + Đối thuyết H_1 : Có sự khác biệt ý nghĩa giữa ít nhất hai nhóm đối với ít nhất một biến phụ thuộc.

Multivariate linear model

=====						

	Intercept	Value	Num DF	Den DF	F Value	Pr > F

	Wilks' lambda	0.0356	4.0000	20645.0000	139716.7541	0.0000
	Pillai's trace	0.9644	4.0000	20645.0000	139716.7541	0.0000
	Hotelling-Lawley trace	27.0703	4.0000	20645.0000	139716.7541	0.0000
	Roy's greatest root	27.0703	4.0000	20645.0000	139716.7541	0.0000

	C(Cluster)	Value	Num DF	Den DF	F Value	Pr > F

	Wilks' lambda	0.0194	12.0000	54621.8273	15638.2536	0.0000
	Pillai's trace	1.6579	12.0000	61941.0000	6376.6092	0.0000
	Hotelling-Lawley trace	18.2657	12.0000	36124.4377	31423.3959	0.0000
	Roy's greatest root	16.4818	4.0000	20647.0000	85074.7997	0.0000
=====						

Hình 11: Kết quả sau khi thực hiện MANOVA


- Nhận xét:
 - + Các chỉ số kiểm định như Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace và Roy's Greatest Root đều cho thấy tồn tại sự khác biệt có ý nghĩa thống kê giữa các nhóm khách hàng đối với ít nhất một biến phụ thuộc
 - + Giá trị p-value rất nhỏ (tiệm cận 0) ở tất cả các phép đo càng củng cố bằng chứng về sự khác biệt đáng kể giữa các nhóm
 - + Với kết quả đó, ta có cơ sở để bác bỏ giả thuyết (H_0) và chấp nhận đối thuyết (H_1), tức là tồn tại sự khác biệt có ý nghĩa thống kê giữa ít nhất hai nhóm khách hàng xét trên một hoặc nhiều biến phụ thuộc.

3.2 Phân tích phương sai một chiều (Analysis of Variance - ANOVA)

- Xây dựng các cặp giả thuyết thống kê cần kiểm định:

1. Length (L)

- + Giả thuyết H_0 : Không có sự khác biệt có ý nghĩa thống kê về giá trị trung bình của biến Length giữa các nhóm khách hàng.
- + Đối thuyết H_1 : Có ít nhất một cặp nhóm khách hàng có giá trị trung bình Length khác biệt có ý nghĩa thống kê.

 ANOVA cho biến: Length

	sum_sq	df	F	PR(>F)
C(Cluster)	3.099887e+09	3.0	84952.027834	0.0
Residual	2.511475e+08	20648.0	NaN	NaN


Hình 12: Kết quả sau khi thực hiện ANOVA cho biến Length

- Nhận xét:

- + Vì $p\text{-value} = 0.0 < 0.05$ (với mức ý nghĩa 5%), ta có cơ sở để bác bỏ giả thuyết (H_0) và chấp nhận đối thuyết (H_1), tức là có sự khác biệt có ý nghĩa thống kê giữa các cụm về Length

2. Recency (R)

- + Giả thuyết H_0 : Không có sự khác biệt có ý nghĩa thống kê về giá trị trung bình của biến Recency giữa các nhóm khách hàng.
- + Đối thuyết H_1 : Có ít nhất một cặp nhóm khách hàng có giá trị trung bình Recency khác biệt có ý nghĩa thống kê.

 ANOVA cho biến: Recency

	sum_sq	df	F	PR(>F)
C(Cluster)	1.724849e+09	3.0	28484.275878	0.0
Residual	4.167761e+08	20648.0	NaN	NaN

Hình 13: Kết quả sau khi thực hiện ANOVA cho biến Recency

- Nhận xét:

- + Vì $p\text{-value} = 0.0 < 0.05$ (với mức ý nghĩa 5%), ta có cơ sở để bác bỏ giả thuyết (H_0) và chấp nhận đối thuyết (H_1), tức là có sự khác biệt có ý nghĩa thống kê giữa các cụm về Recency

3. Frequency (F)

- + Giả thuyết H_0 : Không có sự khác biệt có ý nghĩa thống kê về giá trị trung bình của biến Frequency giữa các nhóm khách hàng.
- + Đối thuyết H_1 : Có ít nhất một cặp nhóm khách hàng có giá trị trung bình Frequency khác biệt có ý nghĩa thống kê.

ANOVA cho biến: Frequency				
	sum_sq	df	F	PR(>F)
C(Cluster)	90639.349018	3.0	19884.680773	0.0
Residual	31372.916331	20648.0	NaN	NaN

Hình 14: Kết quả sau khi thực hiện ANOVA cho biến Frequency

- Nhận xét:

- + Vì $p\text{-value} = 0.0 < 0.05$ (với mức ý nghĩa 5%), ta có cơ sở để bác bỏ giả thuyết (H_0) và chấp nhận đối thuyết (H_1), tức là có sự khác biệt có ý nghĩa thống kê giữa các cụm về Frequency

4. Monetary (M)

- + Giả thuyết H_0 : Không có sự khác biệt có ý nghĩa thống kê về giá trị trung bình của biến Monetary giữa các nhóm khách hàng.
- + Đối thuyết H_1 : Có ít nhất một cặp nhóm khách hàng có giá trị trung bình Monetary khác biệt có ý nghĩa thống kê.

ANOVA cho biến: Monetary				
	sum_sq	df	F	PR(>F)
C(Cluster)	3.429906e+10	3.0	18605.025769	0.0
Residual	1.268845e+10	20648.0	NaN	NaN

Hình 15: Kết quả sau khi thực hiện ANOVA cho biến Monetary

- Nhận xét:

- + Vì $p\text{-value} = 0.0 < 0.05$ (với mức ý nghĩa 5%), ta có cơ sở để bác bỏ giả thuyết (H_0) và chấp nhận đối thuyết (H_1), tức là có sự khác biệt có ý nghĩa thống kê giữa các cụm về Monetary

3.3 Phân tích hậu định (Post-hoc Analysis) với Tukey's Honest Significant Difference


Sau khi ANOVA chỉ ra rằng tồn tại sự khác biệt có ý nghĩa thống kê giữa các nhóm, kiểm định hậu định Tukey HSD được tiến hành để xác định cụ thể những cặp nhóm nào có sự khác biệt. Mỗi cặp so sánh trong phân tích hậu định ngầm định một cặp giả thuyết thống kê riêng, trong đó giả thuyết không cho rằng trung bình của hai nhóm là bằng nhau, và đối thuyết cho rằng chúng khác nhau. Quyết định bác bỏ H_0

dựa trên việc giá trị p nhỏ hơn mức ý nghĩa α hoặc chỉ số Q vượt quá ngưỡng lý thuyết.

Kiểm định hậu định Tukey HSD được thực hiện để xác định cụ thể cặp cụm nào khác biệt có ý nghĩa thống kê về các chỉ số Length, Recency, Frequency, Monetary sau khi ANOVA cho kết quả tổng thể có ý nghĩa.

Bảng kết quả cung cấp:


- + meandiff: chênh lệch trung bình giữa các cụm.
- + p-adj: giá trị p đã điều chỉnh để kiểm soát sai lầm loại I.
- + reject: kết luận bác bỏ giả thuyết H_0 tại mức ý nghĩa 5%.
- Length (L)

 Post Hoc (Tukey HSD) cho biến: Length
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
0 1 -901.1535 0.0 -906.52 -895.7869 True
0 2 -263.4836 0.0 -268.8329 -258.1342 True
0 3 -883.0837 0.0 -889.3258 -876.8416 True
1 2 637.6699 0.0 632.501 642.8388 True
1 3 18.0698 0.0 11.9816 24.1579 True
2 3 -619.6001 0.0 -625.6731 -613.5271 True
```

Hình 16: Kết quả sau khi thực hiện Tukey HSD cho biến Length

- Nhận xét:
 - + Tất cả các cặp cụm đều có p-adj = 0.0 và reject = True, cho thấy sự khác biệt rõ rệt có ý nghĩa thống kê giữa các cụm về Length.
- Recency (R)


 Post Hoc (Tukey HSD) cho biến: Recency
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
0 1 723.6979 0.0 716.7847 730.6112 True
0 2 158.084 0.0 151.193 164.9751 True
0 3 159.049 0.0 151.0079 167.0901 True
1 2 -565.6139 0.0 -572.2726 -558.9552 True
1 3 -564.649 0.0 -572.4918 -556.8061 True
2 3 0.9649 0.989 -6.8584 8.7882 False
```

Hình 17: Kết quả sau khi thực hiện Tukey HSD cho biến Recency

- Nhận xét:
 - + Chỉ có cụm 2 và cụm 3 ($p = 0.989 > 0.05$) do đó không có sự khác biệt đáng kể giữa 2 cụm này.

- + Còn lại hầu hết các cặp cụm đều có $p\text{-adj} = 0.0$ và $\text{reject} = \text{True}$, cho thấy sự khác biệt rõ rệt có ý nghĩa thống kê giữa các cụm về Recency..
- Frequency (F)

 Post Hoc (Tukey HSD) cho biến: Frequency
Multiple Comparison of Means - Tukey HSD, FWER=0.05


```
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject

0	1	-5.1384	0.0	-5.1984	-5.0784	True
0	2	-2.4234	0.0	-2.4832	-2.3636	True
0	3	-5.0485	0.0	-5.1183	-4.9788	True
1	2	2.715	0.0	2.6572	2.7727	True
1	3	0.0899	0.0038	0.0218	0.1579	True
2	3	-2.6251	0.0	-2.693	-2.5572	True

Hình 18: Kết quả sau khi thực hiện Tukey HSD cho biến Frequency

- Nhận xét:
 - + Tất cả các cặp cụm đều có $p\text{-adj} = 0.0$ và $\text{reject} = \text{True}$, cho thấy sự khác biệt rõ rệt có ý nghĩa thống kê giữa các cụm về Frequency.
- Monetary (M)

 Post Hoc (Tukey HSD) cho biến: Monetary
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
```

group1	group2	meandiff	p-adj	lower	upper	reject

0	1	-3176.1533	0.0	-3214.2982	-3138.0084	True
0	2	-1541.7845	0.0	-1579.807	-1503.762	True
0	3	-3103.5845	0.0	-3147.9524	-3059.2166	True
1	2	1634.3688	0.0	1597.6287	1671.1089	True
1	3	72.5688	0.0001	29.2948	115.8427	True
2	3	-1561.8	0.0	-1604.9662	-1518.6339	True

Hình 19: Kết quả sau khi thực hiện Tukey HSD cho biến Monetary

- Nhận xét:
 - + Tất cả các cặp cụm đều có $p\text{-adj} = 0.0$ và $\text{reject} = \text{True}$, cho thấy sự khác biệt rõ rệt có ý nghĩa thống kê giữa các cụm về Monetary.

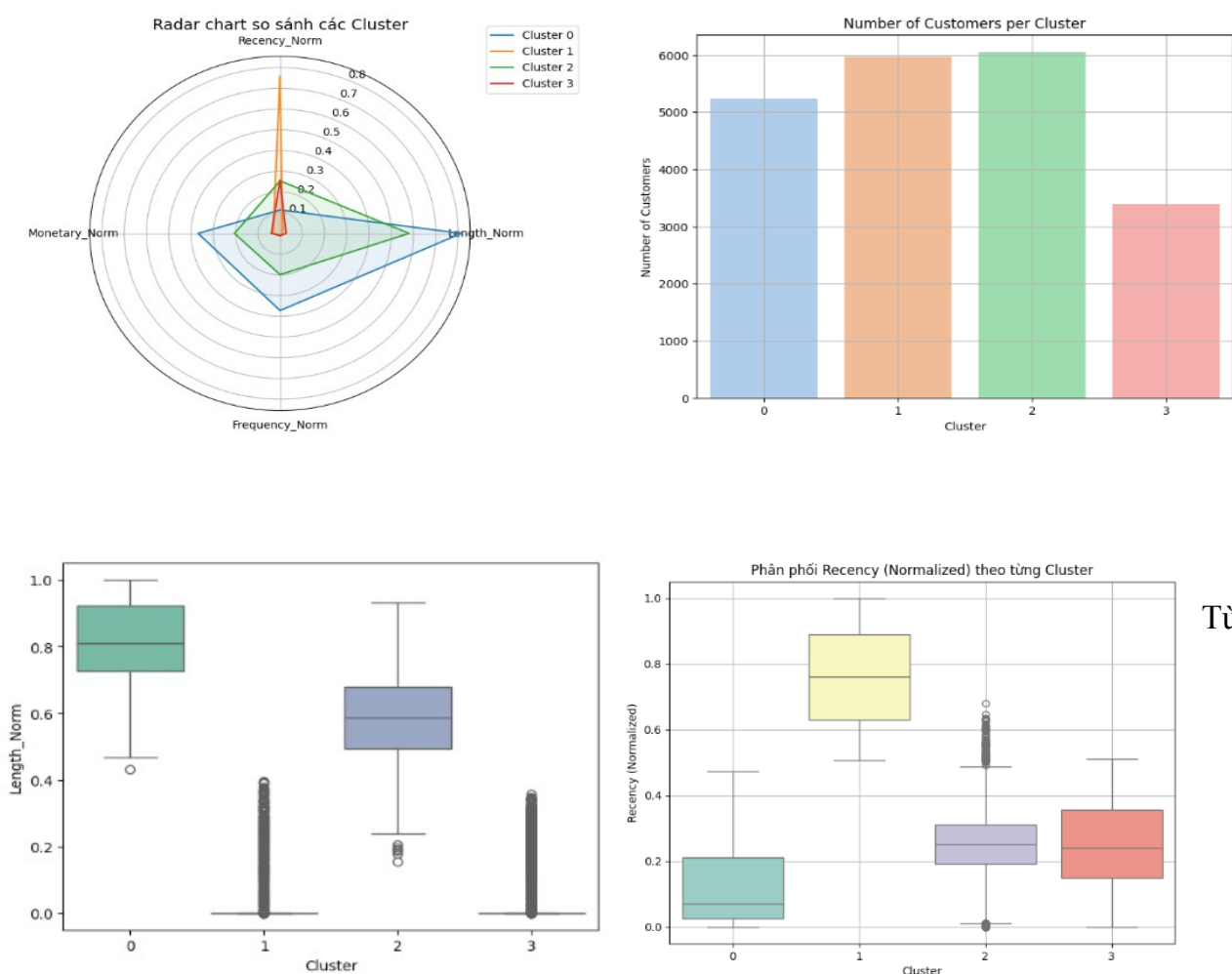
IV. KẾT QUẢ

Các đặc trưng hành vi mua sắm theo từng cụm được thể hiện rõ ràng qua các biểu đồ boxplot, biểu đồ radar và bảng tổng hợp trung bình các chỉ số LRFM.

Bảng tổng hợp LRFM + số lượng khách:

	Cluster	Num_Customers	Mean_Recency	Mean_Length	Mean_Frequency	Mean_Monetary
0	0	5231	128.560	913.752	6.210	3481.299
1	1	5969	852.257	12.599	1.072	305.145
2	2	6052	286.644	650.269	3.787	1939.514
3	3	3400	287.609	30.669	1.162	377.714

Hình 20: Kết quả sau khi chạy thuật toán Kmeans



Hình 21: Hình ảnh trực quan về kết quả của từng cụm

đó, ta có thể rút ra đặc điểm và đặt tên cho từng nhóm như sau:

1. Cụm 0 – Khách hàng trung thành và có giá trị cao (VIP Loyal Customers)

	Variable	Count (number of customers)	Mean	Standard Deviation	Min	25%	50%	75%	Max
0	Length	5231.0	913.752	131.765	483.00	815.000	909.00	1035.000	1122.00
1	Recency	5231.0	128.560	112.283	1.00	29.000	81.00	240.000	534.00
2	Frequency	5231.0	6.210	1.889	1.00	5.000	6.00	7.000	15.00
3	Monetary	5231.0	3481.299	1208.742	445.73	2656.145	3408.38	4205.765	9436.61

Hình 22: Kết quả của Cụm 0

Khách hàng trong cụm này có Recency trung bình thấp (128 ngày), cho thấy họ mua hàng khá gần đây. Đồng thời, Length rất cao (trung bình 913 ngày), chứng minh rằng đây là những khách hàng đã gắn bó với doanh nghiệp trong thời gian dài. Tần suất mua hàng (Frequency) cao nhất (6 lần) và giá trị chi tiêu (Monetary) cũng cao nhất (trung bình 3.481), cho thấy họ không những mua nhiều mà còn mang lại doanh thu lớn. Đây chính là nhóm khách hàng trung thành, có giá trị cao và cần được ưu tiên giữ chân thông qua các chương trình khách hàng thân thiết, ưu đãi cá nhân hóa hoặc chiến lược chăm sóc đặc biệt.

2. Cụm 1 – Khách hàng không còn hoạt động (Lost Customers)

	Variable	Count (number of customers)	Mean	Standard Deviation	Min	25%	50%	75%	Max
0	Length	5969.0	12.599	57.426	0.00	0.0	0.00	0.00	445.00
1	Recency	5969.0	852.257	173.012	572.00	711.0	857.00	1002.00	1127.00
2	Frequency	5969.0	1.072	0.378	1.00	1.0	1.00	1.00	6.00
3	Monetary	5969.0	305.145	358.812	8.47	51.4	221.55	378.41	2919.79

Hình 23: Kết quả của cụm 1

Đây là nhóm có Recency rất cao (trung bình 852 ngày) – tức đã rất lâu không còn mua hàng. Đồng thời, Length rất thấp (12 ngày), thể hiện rằng họ chỉ xuất hiện trong thời gian ngắn. Tần suất và chi tiêu cũng rất thấp, cho thấy đây là những khách hàng từng mua 1–2 lần và đã “rời bỏ” doanh nghiệp. Do đó, nhóm này được xếp vào dạng khách hàng không còn hoạt động. Do tỷ lệ chi tiêu thấp, doanh nghiệp có thể cân nhắc không đầu tư nhiều ngân sách cho nhóm này hoặc chỉ triển khai một số chiến dịch hồi sinh khách hàng với chi phí tối ưu.

3. Cụm 2 – Khách hàng tiềm năng (Potential Loyalists)

	Variable	Count (number of customers)	Mean	Standard Deviation	Min	25%	50%	75%	Max
0	Length	6052.0	650.269	138.892	174.00	555.000	659.00	762.000	1046.00
1	Recency	6052.0	286.644	131.555	1.00	217.000	284.00	352.000	766.00
2	Frequency	6052.0	3.787	1.346	1.00	3.000	4.00	5.000	10.00
3	Monetary	6052.0	1939.514	782.433	60.06	1367.875	1905.11	2462.953	5038.94

Hình 24: Kết quả của cụm 2

Khách hàng thuộc cụm 2 có Length khá cao (~650 ngày), cho thấy mối quan hệ mua hàng kéo dài. Recency ở mức trung bình (~286 ngày), có nghĩa họ chưa mua hàng gần đây nhưng không quá lâu. Tần suất mua (3.8 lần) và giá trị chi tiêu (~1.939) khá ổn định, chứng minh đây là những khách hàng có tiềm năng phát triển thành nhóm trung thành nếu được chăm sóc đúng cách. Doanh nghiệp có thể nhắm đến nhóm này trong các chiến dịch kích thích mua lại, nâng hạng hoặc chương trình gắn kết dài hạn.

4. Cụm 3 – Khách hàng mới & chưa hoạt động nhiều (Newcomers)

	Variable	Count (number of customers)	Mean	Standard Deviation	Min	25%	50%	75%	Max
0	Length	3400.0	30.669	83.912	0.00	0.00	0.00	0.00	401.00
1	Recency	3400.0	287.609	140.896	1.00	168.00	272.00	403.00	578.00
2	Frequency	3400.0	1.162	0.511	1.00	1.00	1.00	1.00	5.00
3	Monetary	3400.0	377.714	411.060	8.47	83.56	246.31	442.04	3392.37

Hình 25: Kết quả của cụm 3

Nhóm này có Recency và Length ở mức trung bình đến thấp (~287 và 30 ngày), nghĩa là họ mới bắt đầu giao dịch gần đây nhưng chưa duy trì lâu. Tần suất mua và

chi tiêu thấp nhất trong tất cả các nhóm. Đây là nhóm khách hàng mới đang trong giai đoạn đầu trải nghiệm dịch vụ/sản phẩm. Doanh nghiệp nên có chiến lược onboarding rõ ràng như email hướng dẫn, chương trình chào mừng, mã giảm giá thử nghiệm... để thúc đẩy hành vi mua lại và kéo dài mối quan hệ với nhóm này.

5. Phân tích Customer Lifetime Value (CLV)

1. Xây dựng các biến phục vụ tính CLV

Dữ liệu đầu vào được sử dụng để xây dựng các biến trung gian phục vụ tính toán CLV bao gồm: giá trị chi tiêu tổng (Monetary), số lần mua hàng (Frequency) và số ngày gắn bó (Length). Từ các biến này, hai biến trung gian được tạo ra như sau:

+ Giá trị đơn hàng trung bình (Average Order Value):

$$\text{AvgOrderValue} =$$

+ Vòng đời khách hàng theo năm (LifetimeYear)

$$\text{LifetimeYear} =$$

2. Công thức và cách tính Customer Lifetime Value (CLV)

Công thức CLV được áp dụng trong nghiên cứu này là phiên bản cơ bản, được trích từ các tài liệu học thuật như Gupta et al. (2006) và Farris et al. (2010), cụ thể như sau:

$$\text{CLV} = \text{AvgOrderValue} \times \text{Frequency} \times \text{LifetimeYear}$$

- Trong đó:

- + AvgOrderValue: Giá trị trung bình của mỗi đơn hàng, được tính bằng tổng chi tiêu chia cho số lần mua hàng.
- + Frequency: Số lần khách hàng thực hiện giao dịch mua hàng
- + LifetimeYear: Khoảng thời gian khách hàng gắn bó với doanh nghiệp, tính bằng số năm giữa lần mua đầu tiên và lần mua gần nhất.

Sau khi các biến trung gian được xác lập đầy đủ, giá trị CLV của từng khách hàng được tính toán theo công thức sau:

$$= \times \times$$

3. Phân tích CLV theo cụm khách hàng

Giá trị CLV sau khi tính toán được tổng hợp theo từng cụm khách hàng (Cluster) đã được phân loại bằng thuật toán K-means. Mục tiêu của phân tích nhằm đánh giá mức độ chênh lệch giá trị giữa các nhóm, từ đó hỗ trợ ra quyết định trong hoạt động chăm sóc và giữ chân khách hàng.

Phương pháp tính được thực hiện bằng cách lấy trung bình CLV của từng cụm theo công thức:

$$=$$

- Trong đó:

- : là CLV trung bình của cụm k
- : là số khách hàng trong cụm đó
- là giá trị CLV của từng khách hàng i

Kết quả phân tích:

Sau khi thực hiện phép tổng hợp, nhóm thu được giá trị CLV trung bình tương ứng với từng cụm khách hàng như trình bày trong bảng dưới đây:

Cluster	Average CLV
0	8622.116420
1	36.972006
2	3466.455400
3	82.791523

Kết luận:

Kết quả phân tích cho thấy có sự phân hóa rõ rệt về giá trị vòng đời khách hàng giữa các cụm:

- + Cluster 0 là nhóm có giá trị CLV trung bình cao nhất (8,622.12), cho thấy đây là những khách hàng thường xuyên mua hàng, đơn hàng có giá trị cao và duy trì mối quan hệ lâu dài với doanh nghiệp. Đây là nhóm cần được ưu tiên giữ chân và chăm sóc đặc biệt.
- + Cluster 2 có CLV trung bình khá cao (3,466.46), thể hiện tiềm năng phát triển thành nhóm khách hàng trung thành nếu được chăm sóc tốt hơn.
- + Ngược lại, Cluster 1 (36.97) và Cluster 3 (82.79) có CLV rất thấp, là nhóm khách hàng mới mua hàng 1 lần, không thường xuyên hoặc đã ngừng tương tác. Do đó, cần có các chiến dịch marketing kích hoạt lại hoặc đánh giá lại mức độ phù hợp của sản phẩm/dịch vụ với nhóm này.

4. Phân tích CLV theo phân khúc khách hàng

Bên cạnh việc phân tích theo cụm, CLV cũng được phân tích theo các phân khúc sử dụng (Segment) bao gồm Consumer, Corporate và Home Office. Mục tiêu nhằm so sánh giá trị vòng đời giữa các nhóm khách hàng theo đặc điểm định sẵn trong dữ liệu, từ đó đưa ra định hướng chiến lược phù hợp.

Giá trị CLV trung bình theo từng phân khúc được tính theo công thức:

=

- Trong đó:

- + : là CLV trung bình của phân khúc j
- + : là số khách hàng trong cụm đó
- + là giá trị CLV của từng khách hàng i

Kết quả phân tích:

Segment	Average CLV
Consumer	3385.633358
Corporate	3090.229713

Home Office	2983.906455
-------------	-------------

- Kết luận:

Phân tích CLV theo phân khúc khách hàng cho thấy sự khác biệt nhất định giữa các nhóm sử dụng:

- + Nhóm Consumer có giá trị CLV trung bình cao nhất (3,385.63), cho thấy đây là nhóm khách hàng cá nhân có mức độ tiêu dùng cao và tiềm năng giữ chân tốt. Đây có thể là nhóm mục tiêu chính trong các chiến dịch gia tăng giá trị đơn hàng hoặc xây dựng lòng trung thành dài hạn.
- + Corporate đứng thứ hai (3,090.23), là nhóm khách hàng doanh nghiệp. Mặc dù CLV không vượt trội nhưng vẫn đủ ổn định để duy trì, đặc biệt nếu doanh nghiệp có chính sách bán hàng theo hợp đồng hoặc đơn hàng lớn.
- + Home Office có CLV thấp nhất (2,983.91), có thể là nhóm khách hàng nhỏ lẻ hoặc không mua hàng thường xuyên. Đây là nhóm cần được xem xét kỹ hơn để xác định nguyên nhân và đưa ra chiến lược kích thích nhu cầu phù hợp.

V. KẾT LUẬN

Phân cụm khách hàng là một trong những kỹ thuật cốt lõi trong phân tích dữ liệu định hướng marketing, nhằm mục tiêu nhóm các khách hàng có đặc điểm hành vi tương đồng thành những cụm riêng biệt, từ đó hỗ trợ doanh nghiệp hiểu rõ cấu trúc tập khách hàng của mình và đưa ra các chiến lược tiếp thị phù hợp. Trong nghiên cứu này, nhóm đã tiến hành phân cụm khách hàng dựa trên mô hình mở rộng LRFM – bao gồm bốn chỉ số phản ánh hành vi tiêu dùng: thời điểm mua gần nhất (Recency), độ dài quan hệ (Length), tần suất mua (Frequency), và tổng chi tiêu (Monetary). Dữ liệu thực tế được xử lý, chuẩn hóa và phân cụm bằng thuật toán học máy K-Means, một phương pháp phổ biến trong phân tích không giám sát nhờ khả năng tối ưu hóa khoảng cách nội cụm và ngoại cụm.

Bên cạnh quy trình phân cụm, nhóm cũng triển khai kiểm định giả thuyết thống kê nhằm đánh giá mức độ khác biệt có ý nghĩa giữa các cụm khách hàng. Phân tích phương sai đa biến (MANOVA) được sử dụng để xác định sự khác biệt tổng thể giữa các cụm trên tổ hợp các chỉ số LRFM, sau đó là phân tích phương sai một chiều (ANOVA) trên từng biến riêng lẻ, và cuối cùng là phân tích hậu định với kiểm định Tukey HSD để chỉ ra cụ thể những cặp cụm có khác biệt rõ rệt. Chuỗi kiểm định này không chỉ giúp xác nhận hiệu lực thống kê của quá trình phân cụm, mà còn đóng vai trò như một lớp giải thích bổ sung, làm rõ các chiều cạnh hành vi tạo nên sự phân hóa giữa các nhóm khách hàng.

Thông qua việc áp dụng mô hình LRFM kết hợp thuật toán K-Means và các phương pháp kiểm định thống kê, nhóm không chỉ hiểu rõ hơn về cách tổ chức và triển khai một quy trình phân tích khách hàng hoàn chỉnh, mà còn rút ra được những bài học thực tiễn về việc sử dụng kỹ thuật phân cụm làm nền tảng cho các chiến lược quản

trị quan hệ khách hàng. Nghiên cứu này là minh chứng cho giá trị ứng dụng của phân tích dữ liệu trong việc ra quyết định kinh doanh, đặc biệt trong bối cảnh doanh nghiệp cần xây dựng chiến lược tiếp thị cá nhân hóa và định hướng giữ chân khách hàng có giá trị cao.

VI. TÀI LIỆU THAM KHẢO

- Nguyen, A. T. V., McClelland, R., & Thuan, N. H. (2020). Omni-channel customer segmentation: A personalized customer journey perspective. *Journal of Consumer Behaviour*.
- Tsitsis, K., & Chorianopoulos, A. (2009). *Data Mining Techniques in CRM: Inside Customer Segmentation*. Wiley.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, 68(4), 106–125.
- Moore, D. S., Notz, W. I., & Fligner, M. A. (2018). *The Basic Practice of Statistics* (8th ed.). Macmillan Learning.
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage.
- Sharma, S. (1996). *Applied Multivariate Techniques*. John Wiley & Sons.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Pearson Education.
- Stone, M., & Jacobs, R. (2007). *Successful direct marketing methods*. McGraw-Hill.
- Hughes, A. M. (2005). *Strategic database marketing*. McGraw-Hill.
- Sorkun, M. F., Nabatchi, E., & Delen, D. (2022). Customer segmentation and behavior modeling using RFM, demographics and online shopping preferences. *Expert Systems with Applications*, 196, 116578.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Pearson Education.
- Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences* (5th ed.). Routledge.
- Wei, C. P., Chiu, I. T., & Lin, Y. H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, 39(5), 5529–5533.
- Ghasemzadeh, M., Rezaei, F., & Ghasemi, R. (2021). Analyzing customer value with RFM-T Model. *Journal of Retailing and Consumer Services*, 61, 102580.
- Gupta, S., Lehmann, D. R., & Stuart, J. A. (2006). *Customer lifetime value*. Harvard Business School Working Paper, No. 07-039.

- Farris, P. W., Bendle, N. T., Pfeifer, P. E., & Reibstein, D. J. (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance* (2nd ed.). Pearson Education.
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage.
- Sharma, S. (1996). *Applied Multivariate Techniques*. John Wiley & Sons.
- Christy, A. J., Umamakeswari, A., & Sivakumar, K. (2021). Customer segmentation using RFM and clustering techniques: A study in e-retail domain. *Journal of Retailing and Consumer Services*, 61, 102551.
- Arthur, D., & Vassilvitskii, S. (2007). *k-means++: The advantages of careful seeding*. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027–1035.
- Inaba, M., Katoh, N., & Imai, H. (1994). *Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering*. Proceedings of the 10th Annual Symposium on Computational Geometry, 332–339.
<https://doi.org/10.1145/177424.178065>
- Jain, A. K. (2010). *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Lloyd, S. (1982). *Least squares quantization in PCM*. IEEE Transactions on Information Theory, 28(2), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- 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, 281–297.
- Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). *An overview of clustering methods*. Engineering Applications of Artificial Intelligence, 20(2), 115–126. <https://doi.org/10.1016/j.engappai.2006.10.015>
- Punhani, R., & others. (2021). *Customer Segmentation and Targeted Marketing Strategies Using Machine Learning*. International Research Journal of Modernization in Engineering Technology and Science, 3(6).
https://www.irjmets.com/uploadedfiles/paper/volume3/issue_6_june_2021/11750/final/fin_irjmets1624195550.pdf
- Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.
- Xu, R., & Wunsch, D. (2005). *Survey of clustering algorithms*. IEEE Transactions on Neural Networks, 16(3), 645–678. <https://doi.org/10.1109/TNN.2005.845141>
- Magento. (2014). *Customer segmentation strategies*. Retrieved from <https://business.adobe.com/blog/basics/what-is-customer-segmentation>
- Hughes, A. M. (1994). *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program*. Probus Publishing.

