JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2025**, 22, 025-032

ỨNG DỤNG PHƯƠNG PHÁP PHÂN CỤM ĐỂ PHÂN LOẠI KHÁCH HÀNG DỰA TRÊN HÀNH VI MUA SẮM

Nguyễn Minh Đức, Trần Nguyễn Ngọc Minh Thiệu, Lê Quốc Dũng, Tào Hữu Đạt, Phan Nguyệt Minh*
Trường Đại học Sài Gòn, Thành phố Hồ Chí Minh, Việt Nam
* Tác giả liên hệ: *minhpn@sgu.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận: 14/4/2025
Ngày hoàn thiện: 19/5/2025
Ngày chấp nhận: 21/5/2025
Ngày đăng: 15/9/2025

TỪ KHÓA

Phân cụm khách hàng;
Hành vi mua sắm;
K-Means;
Hierarchical Clustering;
Gaussian Mixture Model (GMM).

TÓM TẮT

Phân cụm khách hàng là rất quan trọng cho việc tối ưu hóa chiến lược marketing. Nghiên cứu này ứng dụng và so sánh hiệu quả của ba thuật toán phân cụm phổ biến: K-Means, Hierarchical Clustering và Gaussian Mixture Models (GMM) để phân loại khách hàng dựa trên hành vi mua sắm và đặc điểm nhân khẩu học (tuổi, giới tính, tổng chi tiêu). Sử dụng ba bộ dữ liệu bán lẻ (hai từ Kaggle, một từ Sling Academy), nghiên cứu tiến hành tiền xử lý dữ liệu, áp dụng các thuật toán phân cụm, và đánh giá hiệu quả bằng các chỉ số Silhouette Score, Davies-Bouldin Index, và Calinski-Harabasz Index. Kết quả cho thấy GMM hoạt động hiệu quả nhất trong việc phân cụm dựa trên tổng chi tiêu và giới tính, tạo ra các nhóm rõ ràng. Hierarchical Clustering tỏ ra phù hợp khi cần phân tích chi tiết theo độ tuổi trên một số bộ dữ liệu, trong khi K-Means cung cấp một giải pháp cân bằng, đặc biệt hiệu quả khi cấu trúc cụm rõ ràng hoặc cần kết quả nhanh chóng. Nghiên cứu đề xuất lựa chọn thuật toán phù hợp dựa trên mục tiêu kinh doanh và đặc tính dữ liệu cụ thể, giúp doanh nghiệp xây dựng chiến lược marketing cá nhân hóa hiệu quả hơn.

APPLYING CLUSTERING METHODS TO CLASSIFY CUSTOMERS BASED ON SHOPPING BEHAVIOUR

Nguyen Minh Duc, Tran Nguyen Ngoc Minh Thieu, Le Quoc Dung, Tao Huu Dat, Phan Nguyet Minh*
Saigon University, Ho Chi Minh City, Vietnam
*Corresponding Author: *minhpn@sgu.edu.vn*

ARTICLE INFO

Received: Apr 14th, 2025
Revised: May 19th, 2025
Accepted: May 21st, 2025
Published: Sep 15th, 2025

KEYWORDS

Customer segmentation;
Shopping behaviour;
K-Means;
Hierarchical Clustering;
Gaussian Mixture Model (GMM).

ABSTRACT

Customer segmentation is crucial for optimizing marketing strategies. This study applies and compares the effectiveness of three common clustering algorithms: K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM) to classify customers based on shopping behavior and demographics (age, gender, total spending). Utilizing three retail datasets (two from Kaggle, one from Sling Academy), the research performs data preprocessing, applies the clustering algorithms, and evaluates their performance using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The results indicate that GMM performs most effectively for segmenting based on total spending and gender, creating distinct clusters. Hierarchical Clustering proves suitable for detailed age-based analysis on specific datasets, while K-Means offers a balanced solution, particularly effective when cluster structures are clear or rapid results are needed. The study recommends selecting appropriate algorithms based on specific business objectives and data characteristics, enabling businesses to develop more effective personalized marketing strategies.

Doi: https://doi.org/10.61591/jslhu.22.713
Available online at: https://js.lhu.edu.vn/index.php/lachong

1. INTRODUCTION

In the context of fierce competition, understanding customers is a key factor enabling businesses to devise effective business strategies. Customer clustering, the process of grouping customers with similar characteristics and behaviors, allows businesses to personalize marketing, optimize resources, and increase return on investment (ROI). Traditional segmentation methods often rely on subjective assumptions, whereas unsupervised machine learning-based clustering techniques can objectively detect hidden patterns in data. The RFM (Recency, Frequency, Monetary) model, although popular, has limitations in fully describing customer behavior, often overlooking important demographic and psychological factors [1], [2], for example, not directly considering how age or personal preferences can influence purchasing decisions. This study aims to delve deeper into analyzing behavioral and demographic characteristics (age, gender, total spending) by applying and comparing three clustering algorithms: K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM) across multiple retail datasets. Specific objectives include: (1) Selecting the most suitable clustering method for each type of characteristic (age, gender, total spending); (2) Identifying and interpreting customer segments based on clustering results; (3) Proposing recommendations for selecting and applying clustering algorithms in real-world business practice.

The novelty of this research lies in conducting a quantitative, simultaneous, and systematic comparison of the effectiveness of these three common clustering algorithms across multiple diverse retail datasets, concurrently considering aspects of demographics and shopping behavior, thereby providing a clearer empirical basis for selecting algorithms in specific scenarios.

This research contributes to filling the gap in previous studies by providing a quantitative and diverse comparison of the effectiveness of clustering algorithms across different aspects of customer data.

2. RESEARCH METHODOLOGY

2.1 Data Sources

The study utilizes three publicly available datasets:

- **Dataset 1:** Retail Sales Dataset (Kaggle): Synthetic data simulating a retail environment, including transaction and demographic information (Customer ID, Age, Gender, Annual Income, Total Spend, Years as Customer, etc.). *(Sample size is 1000)*.
- **Dataset 2:** Online Retail Customer Churn Dataset (Kaggle): Data on online retail customer interactions, including demographics, spending behavior, and satisfaction levels (Transaction ID, Date, Customer ID, Gender, Age, Product Category, Total Amount, etc.). *(Sample size is 1000)*.
- **Dataset 3:** Customers Sample Data (Sling Academy): Sample data of 1000 customers including personal information, contact details, and purchase history (first_name, last_name, email, gender, age, spent, job, etc.). *(Sample size is 1000)*.

Using multiple datasets with varying characteristics enhances the generalizability of the comparison results.

2.2 Data Preprocessing

The data preprocessing steps included:

- *Data Cleaning:* Removing rows with missing values in crucial columns (e.g., Customer ID, Total Amount). Handling missing values in other columns by imputation with the mean (numerical columns) or mode (categorical columns) or dropping columns with excessively high missing rates (>50%).
- *Outlier Handling:* Using the Interquartile Range (IQR) method to identify and remove outliers in numerical columns like 'Total Amount' and 'Age'. Specifically, values outside the range [Q1 - 1.5*IQR, Q3* + *1.5*IQR] were excluded.
- *Data Standardization:* Using StandardScaler from the scikit-learn library to standardize numerical variables to the same scale (mean 0, standard deviation 1), ensuring variables with different scales do not disproportionately influence distance-based clustering algorithms (like K-Means).
- *Categorical Variable Encoding:* Converting categorical variables (like 'Gender') into numerical format using one-hot encoding or label encoding as required by the algorithms.

2.3 Clustering Algorithms

These three algorithms were selected because they represent popular and diverse clustering approaches: centroid-based (K-Means), hierarchical (Hierarchical Clustering), and probability model-based (GMM), and are also commonly applied effectively in customer segmentation problems:

- **K-Means:** Partitions data into *k* clusters by minimizing the sum of squared distances from each data point to its nearest cluster center. Advantages: simple, fast. Disadvantages require pre-specifying *k*, sensitive to outliers and non-spherical cluster shapes [1], [4].
- **Hierarchical Clustering Agglomerative:** Builds a hierarchy of clusters by progressively merging the closest clusters. Advantages: it does not require pre-specifying *k*, provides hierarchical structure. Disadvantages: high computational complexity for large datasets [2, pg. 484-497]. Ward's or Average linkage methods are commonly used.
  - The linkage method consistently used for all three datasets was 'ward' with the 'Euclidean' distance metric.
- **Gaussian Mixture Model (GMM):** Assumes data is generated from a mixture of Gaussian distributions. Clusters based on the probability of a data point belonging to each Gaussian component (soft clustering). Advantages: flexible with elliptical and overlapping clusters. Disadvantages: more complex, computationally intensive [2].
  - The covariance type used was 'full', which is the default in the scikit-learn library when not otherwise specified, allowing for the most flexible model.

2.4 Determining the Optimal Number of Clusters (k)

The optimal number of clusters (*k*) was determined using two main methods before running algorithms (especially for K-Means) and for evaluation afterward (for all three):

- **Elbow Method**: Plotting the sum of squared errors within clusters (inertia) against the number of

clusters k. The "elbow point," where the rate of decrease sharply slows, is chosen as the optimal k.

- **Silhouette Score**: Measuring how similar a data point is to its own cluster compared to other clusters. Values range from -1 to 1, with higher values indicating better clustering. The k yielding the highest Silhouette Score is often selected.

2.5 Evaluating Clustering Performance

The effectiveness of the algorithms on each dataset and criterion (age, gender, total spending) was quantitatively assessed using three common indices:

- **Silhouette Score:** As described above. A value near +1 indicates that the object is well-matched to its own cluster, a value near -1 indicates that the object may have been misassigned, and a value near 0 indicates that the object is close to the boundary between two clusters.
- **Davies-Bouldin Index:** Measures the average ratio of within-cluster similarity to between-cluster separation. Lower values indicate better clustering (compact and well-separated clusters), with a value of 0 being ideal.
- **Calinski-Harabasz Index (Variance Ratio Criterion):** Measures the ratio of between-cluster variance to within-cluster variance. Higher values indicate dense and well-separated clusters. That is, points in the same cluster are close together and points in different clusters are far apart.

3. RESULTS AND DISCUSSION

Figure 1 illustrates the hierarchical structure of clusters based on the distances between data points. This diagram visualizes the process of merging clusters at different levels and helps you identify the potential number of clusters by observing the lengths of the branches. You can choose a representative dendrogram, such as the one generated when clustering by 'age' and 'spent', to illustrate how Hierarchical Clustering works and the cluster structure that is formed.

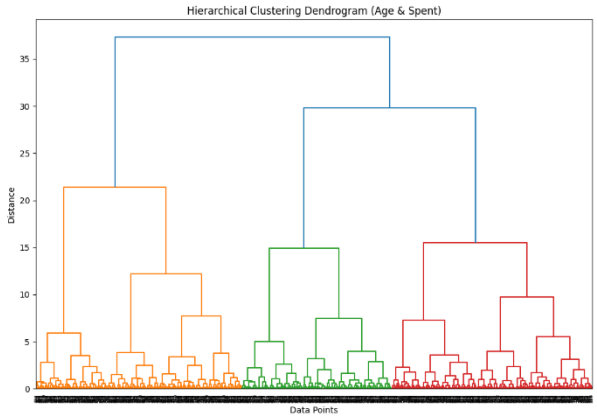

Figure 1: *Dendrogram resulting from Hierarchical Clustering of dataset 3*

Figure 1 illustrates the hierarchical structure of clusters based on the distances between data points. This diagram visualizes the process of merging clusters at different levels and helps you identify the potential number of clusters by observing the lengths of the branches. You can choose a representative dendrogram, such as the one generated when clustering by 'age' and 'spent', to illustrate how Hierarchical Clustering works and the cluster structure that is formed.

The evaluation results of the effectiveness of three clustering algorithms (K-Means, Hierarchical, GMM) on three datasets based on the features Age, Gender, and Total Spending are summarized and compared. The determination of the optimal number of clusters, *k*, reveals differences between the methods and datasets (for example, Elbow and Silhouette might suggest different *k* values, and GMM sometimes requires a larger *k*). Table 1 summarizes the optimal algorithm selection based on the evaluation metrics for each case.

Table 1. *Recommended Clustering Algorithm for Each Dataset and Feature Set.*

Dataset	Feature	Optimal Algorithm	Main Reason (based on Silhouette score)
Dataset 1	Age	K-Means	Silhouette (0.476), Calinski-Harabasz (1094.8) are high; Davies-Bouldin (0.728) is low.
Dataset 1	Gender	K-Means	Silhouette (0.684), Calinski-Harabasz (1703.5) are the highest. Davies-Bouldin (0.644) is good.
Dataset 1	Spent	GMM	Clear separation of 3 spending levels (low, medium, high) although the index is not always the highest.
Dataset 2	Age	K-Means	Silhouette (0.373) is the highest. Other indices are reasonable.
Dataset 2	Gender	GMM/ K-Means	Equivalent indices (Silhouette ~0.434, C-H ~838). GMM might be better if soft clustering is needed.
Dataset 2	Spent	GMM/ Hierarchical	GMM clearly separates 3 groups. Hierarchical (Silhouette 0.547, C-H 1122.5) performs best in terms of indices.
Dataset 3	Age	K-Means	Silhouette (0.382), Calinski-Harabasz (776.6) are the highest.

Dataset	Feature	Optimal Algorithm	Main Reason (based on Silhouette score)
Dataset 3	Gender	Hierarchical	Silhouette (0.545) is the highest, Davies-Bouldin (0.595) is the lowest, C-H (1122.5) is the highest.
Dataset 3	Spent	KMeans/ GMM	K-Means (Silhouette 0.434, C-H 838.8) has good, stable indices. GMM separates groups by gender better.

Discussion:

The results from Table 1 indicate that no single algorithm is universally optimal for all cases.

- **K-Means** is often effective when the data has a relatively clear cluster structure, particularly when clustering by **age** (Dataset 1, 2, 3) or **gender** when the separation is not too complex (Dataset 1). It is also a good choice when simplicity and speed are required. However, it struggles to separate closely related age groups (e.g., distinguishing Gen Z and Millennials if there isn't a significant difference in spending) and may not accurately reflect gender structure if spending behaviour between males and females is similar across certain spending levels.
- **Hierarchical Clustering** proves superior when clustering by **gender** on Dataset 3, indicating its ability to discover hierarchical structures or more complexly shaped groups that K-Means misses. It is also useful when separating detailed **spending** groups (Dataset 2) or when needing to divide age groups into different life stages (e.g., young, middle-aged, elderly - as in the manual analysis of Dataset 3). However, its performance can decrease with large datasets.
- **GMM** demonstrates a clear strength in clustering based on **total spending** (Dataset 1, 2, 3), often producing very distinct and stable spending segments (low, medium, high), even when quantitative metrics are not always the absolute highest. It also performs well in separating by **gender** when spending behaviour between males and females differs significantly at various spending levels (Dataset 3), thanks to its ability to model different distributions. However, it is often the least effective when clustering solely based on **age**, suggesting that age alone does not follow a clear Gaussian distribution in these datasets.

Further analysis of the generated clusters reveals typical customer groups such as: Low Spenders/Budget Shoppers, Moderate Spenders/Loyal Customers, High Spenders/VIPs, Younger Age Group, Middle-aged/Older Age Group, Male/Female Groups with Distinct Spending Behaviours. For example, GMM often separates male customers who tend to spend at two extremes (low or very high), while female spending is more dispersed. Hierarchical Clustering can separate the younger age group (18-45) and the older age group (50-80) more effectively than K-Means.

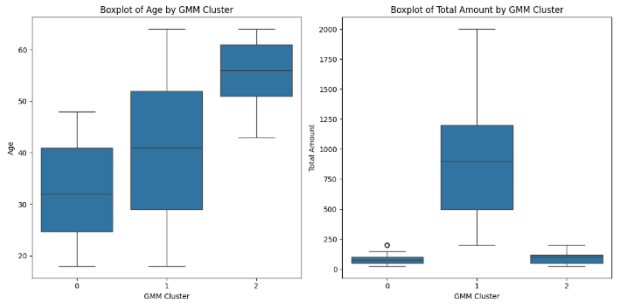

Figure 2. *Boxplot of Age and Total Spending by Cluster – Clustering using Gaussian Mixture Model (GMM) – Dataset 1*

Figure 2: Boxplot of Age (left) and Total Spending (right) by Cluster – Clustering using Gaussian Mixture Model (GMM) – Dataset 1. The chart shows GMM effectively separates customers into three groups based on total spending (Cluster 0: low, Cluster 1: medium, Cluster 2: high), while the separation by age is less clear.

This chart clearly demonstrates the separation of customers into three clusters based on total spending when using the GMM algorithm on Dataset 1.

- **Cluster 0** represents the group of customers with very low spending levels.
- **Cluster 1** indicates the group of customers with a higher average spending level.
- **Cluster 2** identifies the group of customers with very high and consistent spending levels.

⇒ The distinct differences in the range and position of the boxplots across these three clusters make it easy for the reader to observe the effectiveness of GMM in classifying customers by spending level.

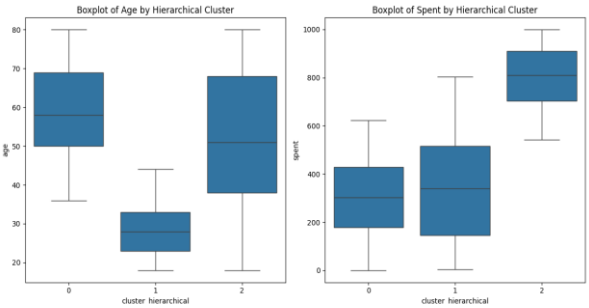

Figure 3. *Boxplot of Age (left) and Spending (right) by Hierarchical Cluster – Dataset 3*

Figure 3: Boxplot of Age (left) and Spending (right) by Hierarchical Cluster - Dataset 3. The chart illustrates a relatively clear separation of customers into three clusters based on age (Cluster 1: young, Cluster 0: middle-aged to older, Cluster 2: dispersed), demonstrating the capability of Hierarchical Clustering in age-based analysis.

The left-hand chart in this figure illustrates a relatively clear separation of customers into three clusters based on age when using the Hierarchical Clustering algorithm on Dataset 3.

- **Cluster 1** focuses on the younger customer group.
- **Cluster 0** represents the middle-aged to older

customer group.

- **Cluster 2** shows a more dispersed group in terms of age.

⇒ Although there might be some overlap, the trend of age separation between the clusters, especially between the younger and older groups, is quite evident, helping to visualize the capability of Hierarchical Clustering in age-based analysis.

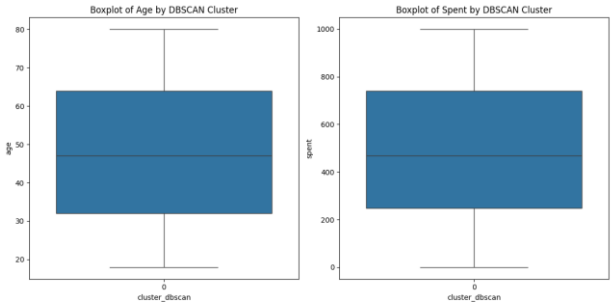

Figure 4. *Age and Total Spending Clustering using DBSCAN – Dataset 3*

Figure 4: Age and Total Spending Clustering using DBSCAN – Dataset 3. This chart illustrates the results when applying DBSCAN with the tested parameters, showing that most data is grouped into one cluster (cluster_id = 0) or marked as noise (-1), failing to create meaningful customer segments based on age and spending.

During the method selection process, the density-based algorithm DBSCAN [DBSCAN Reference] was considered and tested as an alternative, particularly due to its theoretical ability to handle noise and detect clusters of arbitrary shapes. A process to find the optimal parameters for DBSCAN was carried out, where the *epsilon* (eps) value was tested in the range of 0.1 to nearly 2 (with a step of 0.1), and the minimum number of samples (min_samples) was tested in the range of 2 to 9. In addition, specific values such as *eps*=0.5, *min_samples*=5, and the optimal values found from the parameter grid (e.g., *eps*=0.1, *min_samples*=2 in one run) were also applied to evaluate the clustering results on the feature space comprising age and total spending.

However, the results obtained from these experiments showed that DBSCAN was not suitable for the research's segmentation objectives. **Figure 4** illustrates a typical example of the distribution of age and spending according to the "clusters" identified by DBSCAN. The main observation from the experiments is that, with the parameter sets used, **DBSCAN *tended to group all or most of the data into a single cluster*** (cluster_id = 0), only identifying a very few points as noise (-1) or failing to create meaningful statistical or business-relevant separate clusters. This demonstrates that the algorithm, under the experimental data and parameters, was unable to distinguish meaningful customer groups based on the two variables 'age' and 'spent'.

The direct consequence of this is a ***loss of detailed analysis capability*.** When the data is not separated into distinct groups, identifying and describing target segments such as high/low spending groups or young/old age groups becomes impossible, contradicting the core objective of customer segmentation.

The reasons for not continuing further analysis with DBSCAN in this study include:

- **Difficulty in parameter selection:** Determining the appropriate *epsilon* and *MinPts* is crucial and sensitive. Suboptimal choices can lead to erroneous results, as observed (grouping into a single cluster).
- **Difficult to interpret for business objectives:** DBSCAN clusters are based on density and lack clear centroids like K-Means or distribution parameters like GMM, making it harder to create customer profiles and interpret purchasing behaviour**.**
- **Not suitable for uneven density:** Customer data often has groups with varying densities (e.g., many low-volume buyers, few high-volume buyers), which makes it difficult for DBSCAN to apply a single set of parameters effectively.
- **Does not meet clear segmentation objectives:** In marketing, it is often necessary to divide customers into a *defined* number of clusters (e.g., 3 or 5 strategic groups). K-Means and GMM directly meet this requirement, whereas the number of clusters in DBSCAN depends on the data and parameters.

In summary, based on the observed experimental results (DBSCAN not separating effectively, K-Means and GMM producing clear, meaningful clusters) and the suitability for the application objectives (requiring clear segmentation, easy to interpret for marketing), K-Means and GMM are evaluated as **more stable and suitable** than DBSCAN within the scope of this study.

Table 2. *Comparison of Main Clustering Algorithms*

Feature	K-Means	Hierarchical Clustering	Gaussian Mixture Model (GMM)
Principle	Partitioning based on distance to cluster centroids	Building a hierarchical tree (agglomerative/divisive)	Gaussian mixture distribution modeling
Advantages	Simple, fast, easy to implement	Flexible, doesn't require *k* beforehand, structural	Suitable for complex data, soft clustering
Disadvantages	Needs *k* beforehand, sensitive to outliers	Slow with large datasets	Complex, resource-intensive
Number of Clusters Required	Needs to be determined beforehand (*k*)	Doesn't need to be determined beforehand	Needs to be determined beforehand (number of components)
Cluster	Best with	Flexible with	Best with

Feature	K-Means	Hierarchical Clustering	Gaussian Mixture Model (GMM)
Shape	spherical clusters	various shapes	elliptical clusters
Typical Applications	Basic, fast grouping	Structure exploration, detailed analysis	Soft clustering, diverse behavior

Table 2 highlights the core differences between the three selected algorithms, thereby explaining the necessity of comparing them within the specific context of this customer clustering.

- **K-Means,** with its advantages of speed and simplicity, is a good starting point, especially useful when dealing with very large datasets or when customer segments are expected to be relatively separate and spherical (e.g., clearly distinguishing between very low and very high spending groups). However, the requirement to pre-specify the number of clusters (k) is a significant limitation because the natural number of customer segments is often unknown. At the same time, its sensitivity to outliers (e.g., a few extremely high-spending customers) and the assumption of spherical cluster shapes may not perfectly align with the actual distribution of demographic and purchasing behaviour data, which is often complex and non-symmetrical.
- **Hierarchical Clustering** overcomes the drawback of pre-determining *k* in K-Means, providing a visual insight into the hierarchical structure of the data through a dendrogram. This is particularly valuable in the exploratory phase, helping to identify the potential number of clusters or understand the relationships between different customer groups (e.g., are young, low-spending customers closer to older, low-spending customers than to young, medium-spending customers?). The ability to handle clusters with flexible shapes is also an advantage. However, the main disadvantage is the computational complexity that increases rapidly with data size, making it less feasible for very large-scale retail datasets.
- **Gaussian Mixture Model (GMM)** offers a probability-based approach, allowing for "soft clustering," meaning each customer can belong to multiple clusters with different probabilities. This is very suitable for the reality of customer behaviour, where a person can exhibit characteristics of multiple segments (e.g., primarily a budget shopper but occasionally spends like a VIP). The ability to model elliptical clusters is also more flexible than K-Means. However, GMM is theoretically and computationally more complex, requiring more resources, and the interpretation of results may not be as intuitive as.

The choice of the optimal algorithm clearly involves trade-offs between speed, flexibility, interpretability, and suitability for the underlying data structure. Therefore, the experimental evaluation of the effectiveness of all three methods on different datasets and criteria, as presented in the results section, is necessary to provide the most appropriate recommendation for the customer segmentation problem based on purchasing behaviour.

This study emphasizes the importance of selecting the appropriate algorithm for the objectives and data. If the goal is segmentation by spending level, GMM is often the top choice. If in-depth analysis by age or exploration of unknown structures is needed, Hierarchical Clustering might be suitable. K-Means is a powerful, balanced tool when the cluster structure is relatively clear.

Potential contributions of the research:

A key highlight of this research is the simultaneous and systematic quantitative comparison of the effectiveness of three popular clustering algorithms (K-Means, Hierarchical Clustering, GMM) across multiple diverse retail datasets and considering all three important customer aspects: age, gender, and total spending. While previous studies often focused on one or two algorithms (e.g., Chen et al. [1] primarily used K-Means with RFM), or were applied in a narrow context lacking direct comparative benchmarking on the same data (e.g., Wedel & Kamakura [3] used GMM), this research provides a more comprehensive comparative perspective, offering clear empirical bases for selecting the appropriate algorithm depending on specific data characteristics and segmentation goals within the retail domain.

While the research has provided valuable comparisons, several limitations need to be acknowledged. Firstly, the datasets used, although diverse, are primarily from public sources (Kaggle, Sling Academy) and may not fully represent the specific characteristics of every particular retail industry or geographical region. Some of the data is aggregated, which may not fully reflect the complexity of real-world data. Secondly, the research mainly focused on three features: age, gender, and total spending. Other important factors such as detailed purchase history (product type, purchase frequency of each type), psychographic characteristics (lifestyle, attitudes), preferred shopping channels, or geographic location were not integrated into the analysis, which could further enrich the customer segments. Thirdly, the selection of algorithms was limited to K-Means, Hierarchical, and GMM, without exploring other methods such as DBSCAN, neural network-based clustering, or ensemble techniques. Finally, the analysis was primarily static at a single point in time, without delving into the changes in customer behaviour over time. These limitations open up avenues for further research in the future.

Practical Implications for Marketing:

The identification of customer segments as discussed provides significant insights into developing personalized marketing strategies:

- **Low Spenders/Budget Shoppers:** This is often the largest group or a group of new customers. Suitable strategies include promotional offers, volume discounts, introducing value-added products (upselling) at reasonable prices, and easily achievable loyalty programs to encourage repeat purchases.
- **Moderate Spenders/Loyal Customers:** This group has stable value. Focus should be on maintaining loyalty through customer loyalty programs, related product suggestions based on purchase history (cross-selling), early

notifications about discounts, and personalized communication via email or messages.

- **High Spenders/VIPs:** This group generates the most revenue. Special care strategies are needed: exclusive offers, priority customer service, gifts on special occasions, premium shopping experiences, and personal account management.
- **Segmentation by Age (e.g., Young vs. Older):** Younger groups (Gen Z, Millennials) can be effectively reached through digital channels, social media, trending content, and influencer collaborations. Older groups may prefer stability, value, detailed product information, and more traditional channels like email or physical stores.
- **Segmentation by Gender:** When GMM or Hierarchical Clustering clearly separates behavior by gender, businesses can design advertising campaigns, messaging, and product recommendations tailored to the specific preferences of males or females within each spending segment.

Applying the appropriate algorithm (such as GMM to capture gender-based spending differences, or Hierarchical Clustering for life stage analysis) will help make these strategies more precise and effective.

Limitations of the Research:

While the research has provided valuable comparisons, several limitations need to be acknowledged. Firstly, the datasets used, although diverse, are primarily from public sources (Kaggle, Sling Academy) and may not fully represent the specific characteristics of every particular retail industry or geographical region. Some of the data is aggregated, which may not fully reflect the complexity of real-world data. Secondly, the research mainly focused on three features: age, gender, and total spending. Other important factors such as detailed purchase history (product type, purchase frequency of each type), psychographic characteristics (lifestyle, attitudes), preferred shopping channels, or geographic location were not integrated into the analysis, which could further enrich the customer segments. Thirdly, the selection of algorithms was limited to K-Means, Hierarchical, and GMM, without exploring other methods such as DBSCAN, neural network-based clustering, or ensemble techniques. Finally, the analysis was primarily static at a single point in time, without delving into the changes in customer behaviour over time. These limitations open up avenues for further research in the future.

4. CONCLUSION

This research successfully compared the effectiveness of three clustering algorithms – K-Means, Hierarchical Clustering, and GMM – in segmenting customers based on age, gender, and total spending using three retail datasets. The quantitative results and analysis showed that: GMM is often most effective for segmenting by total spending and gender combined with behaviour; Hierarchical Clustering is suitable for detailed age-based analysis or exploring hierarchical structures; and K-Means is a balanced and efficient choice when the cluster structure is clear, or speed is required. This research contributes by providing an empirical comparative assessment, helping researchers and practitioners better understand the strengths and weaknesses of each algorithm in specific customer segmentation scenarios, especially when combining demographic and shopping behavior data, thereby supporting data-driven decision-making.

The research recommends that businesses should: (1) Select clustering algorithms based on specific business objectives (e.g., targeting by spending level or life stage) and data characteristics. (2) Integrate multiple data sources (behavioral, demographic, psychographic) for a comprehensive view. (3) Evaluate cluster quality using both quantitative metrics and practical business significance. (4) Apply clustering results to personalize marketing strategies, optimize customer experiences, and allocate resources effectively. Automating the analysis process also helps save time and ensure consistency.

5. REFERENCES

[1] 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*(3), 197–208.

DOI:10.1057/dbm.2012.17

[2] Vohra, R., Pahareeya, J., Hussain, A., Ghali, F., & Lui, A. (**2020**). Using self organizing maps and K means clustering based on RFM model for customer segmentation in the online retail business. *Lecture Notes in Computer Science*, 484-497. Springer.

[3] Wedel, M., & Kamakura, W. A. (**2000**). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Kluwer Academic Publishers.

[4] Steinbach, M., Karypis, G., & Kumar, V. (**2000**). A comparison of document clustering techniques. *KDD workshop on text mining*.

[5] Karl, T. (**2024**, February 12). DBSCAN vs. K-Means: A guide in python. *New Horizons*. [URL...]

[6] **(Hierarchical Clustering):** Narayanamma, P. L., & Govindan, N. K. (**2021**). An efficient customer segmentation approach using K-Means and Hierarchical clustering techniques for online E-commerce data. *Annals of the Romanian Society for Cell Biology*, *25*(6), 13000-13010.

[7] **(GMM):** Al-Shboul, M., Al-Sayyed, R., & Cristea, A. I. (2023). A New Probabilistic Customer Segmentation Approach Based on Gaussian Mixture Models and Weighted RFM Features for E-commerce Websites. *Information*, *14*(2), 97.

DOI: https://doi.org/10.3390/info14020097

[8] **(*So sánh K-Means, Hierarchical, DBSCAN*)** Alshboul, O., Sheta, A., & Al-Sayyed, R. (**2022**). Customer Segmentation Using Machine Learning Approaches: K-Means, Hierarchical Clustering, and DBSCAN. *In International Conference on Information Technology and Applications* (pp. 187-199). Springer, Singapore.

DOI: https://doi.org/10.1007/978-981-19-3371-2_15

[9] **(Factors & Segmentation)** Putri, D. M., & Nugroho, L. E. (2023). Customer Segmentation Based on Online Purchasing Behavior Using K-Means Clustering. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *7*(1), 161-168.

DOI: https://doi.org/10.29207/resti.v7i1.4698 *(Phân khúc dựa trên hành vi mua online)*

[10] Cui, G., Cheng, Q., & Kwok, K. (**2023**). Artificial

intelligence in marketing research and practice: a review and research agenda. *Review of Marketing Research*, *20*, 73-98.

DOI: https://doi.org/10.1108/S1548-643520230000020004

[11] Kumar, V., & Pansari, A. (**2021**). Artificial Intelligence and Machine Learning in Marketing. *Journal of the Academy of Marketing Science*, *49*(2), 189-191.

DOI: https://doi.org/10.1007/s11747-020-00765-2

[12] Hennig, C. (**2019**). Cluster validation by measurement of clustering characteristics relevant to the user. In *Data Analysis and Applications 1* (pp. 1–24). John Wiley & Sons, Inc.

[13] Runzhao, Y., & Qianni, C. (**2019**). Time-satisfaction of data series based on computer original genetic algorithm gradually covers the location and algorithm of electric vehicle charging station. *Journal of Intelligent & Fuzzy Systems*, *37*(5), 5993–6001.

DOI: https://doi.org/10.3233/jifs-179181

[14] Elayaraja, M., Maheshwari, D., Manikandan, R., & Ramkumar, M. (**2023**). Customer Segmentation using Machine Learning Algorithms: A Review. In *2023 International Conference on Networking and Communications (ICNWC)* (pp. 1–6).

[15] Sinaga, K. P., & Yang, M.-S. (**2020**). Unsupervised K-means clustering algorithm. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 80716–80727.

DOI: https://doi.org/10.1109/access.2020.2988796

[16] Dursun, A., & Caber, M. (**2016**). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, *18*, 153–160.

DOI: https://doi.org/10.1016/j.tmp.2016.03.001

[17] Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63.

DOI: https://doi.org/10.1016/j.procs.2010.12.011

[18] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (**2017**). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1–21.

[19] Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (**2021**). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, *59*(102168), 102168.

DOI: https://doi.org/10.1016/j.ijinfomgt.2020.102168