

Optimizing E-commerce Through Advanced Customer Segmentation: A Comparative Study of Clustering Algorithms

Teja Chaudhari¹, Vivek Gamit², Tirth Ganvit³, Lad rohan⁴, Shankar Parmar⁵ and Deep Upadhyaya⁶

¹⁻²Government Engineering College, Computer Department, Bharuch, India

Email: {tejashchaudhari1807, vgamit637}@gmail.com

³⁻⁵Government Engineering College, Computer Department, Bharuch, India

Email: {ganvittirth46@gmail.com, rohan31054@gmail.com, shankar.parmar@gtu.edu.in }

⁶Government Engineering College, Computer Department, Godhra, India

Email: deep.upadhyaya@gtu.edu.in

Abstract— The rising consumer awareness of online shopping may be driving e-commerce platforms to develop hence a broader insight into the purchasing behavior around their customers. Retail businesses are adopting highly advanced ways of customer segmentation to deliver more customer-focused services that magnify their profits. The objective of this research is to build a customer segmentation model, demonstrating the insights that can be derived from an UK online retailer's dataset comprises 541,909 consumer related attributes available on UCI machine learning repository. This study evaluates customer value based on Recency, Frequency and Monetary (RFM) framework and employs three different clustering algorithms: K-Means cluster algorithm, Gaussian Mixture Model Cluster (GMM), DBSCAN. The results show that GMM has the best performance as it attains a maximum Silhouette Score of 0.98 among others. Customer segmentation success impacts better marketing, customer retention and revenue growth. With the importance of data-driven decision-making growing day by day, clustering algorithms like K-Means and DBSCAN or even GMM have a significant role in understanding consumer buying behavior which can further improve your marketing efforts. The research also evaluates these models by using the Silhouette Score and Davies-Bouldin Index metrics to classify different levels of data segmentation.

Index Terms— Customer Segmentation, Behavioral segmentation, Demographic Analysis, Unsupervised Machine Learning, RFM Analysis, K-Means Clustering, Gaussian Mixture Model (GMM), DBSCAN Algorithm.

I. INTRODUCTION

In the retail business, customer segmentation has become mandatory as businesses try to know their customers better, design services better and improve profitability. customer segmentation takes place so that profitability can be optimized through dividing customers into segments, based on their behavior, demographics, buying patterns, and preferences.

Thus, businesses are able to target favorable segments, and improve their marketing efforts as well as delivering personalized customer experience.

With the advent of Big Data Analytics (BDA) and Business Intelligence (BI) tools, segmentation has become simpler for firms to manage huge data, to extract customer patterns and to build data driven marketing strategies. And both new and existing retail organizations use BDA and BI to guide innovations and resource allocation to enable effective strategies, which require a deep understanding of consumer behavior. Companies are performing this analysis with online customer data to generate personas (representing different groups of individuals with particular preferences and behaviours)[1], [2].

As a foundation, these personas make the work of targeted marketing that addresses distinct customer segments easier. Despite these advancements in customer segmentation techniques, however, there is still no fully integrated contextual framework of these data driven insights into strategic decision-making process. Recent research has identified this gap in practice development, especially with regards to the impact on strategic marketing issues.

After 2017, several studies have tried different clustering approaches for customer segmentation and provided helpful knowledge in the field. For example, Hicham [1] demonstrated the Silhouette score of 0.72, and Turkmen [2] slightly lower the score of 0.6 in the same study.

Correct segmentation and effectiveness are highly dependent on the choice of clustering algorithm. Traditional methods are common but our research uses a variety of methods including Principal Component Analysis (PCA) for dimensionality reduction and Gaussian Mixture Model (GMM) clustering of the resulting features in order to create an enhanced, interpretable segmentation. Retail has seen some great transformations of the decision-making process with Big Data Analytics (BDA) where they can now identify customer patterns, optimize marketing strategies, and target the best customers with the right products. While the benefits of BDA are clear, challenges of translating data insight into actionable strategies still act as a barrier to BDA's optimal utilization in segmentation.. While BDA brings so much concerning customer behavior, a lot of these organizations struggle to incorporate this intelligence into their marketing decision making process[3], [4]. In recent research, we have seen the focus on improving the segmentation model's interpretability and accuracy and with new advances in methods, such as Gaussian Mixture Model (GMM) and Principal Component Analysis (PCA), gaining popularity. GMM's probabilistic approach of customer segmentation and PCA's reduction of dataset dimensions are the two key components of what we compare.

These are the basis of our work that in addition to being able to obtain a segmentation model which is more accurate and interpretable, is able to combine PCA and GMMs to segment[5], [6]. It also enhances segmentation to give us actionable insights for real world marketing strategy. Finally, metrics such as the Silhouette Score that tells us how much our clusters define the dataset are tremendously important for assessing algorithm's effectiveness. Hicham, Karim, and Turkmen's studies are benchmarks in retail segmentation with scores of 0.72 and 0.6, respectively. To build upon these benchmarks, we explore applying PCA and GMM to form such segmentation model with a more accurate and interpretable degree of model[7], [8]. Big Data Analytics and BI tools help customer segmentation in order for us to implement personalized marketing and profitability. These models are refined by our PCA and GMM based research for a data driven retail market.

II. LITERATURE REVIEW

This literature review examines the literature of past studies conducted in customer segmentation as well as unsupervised clustering approach based on K means, GMM, DBSCAN, and hierarchical clustering for studying the customer behavior without labelled data. Each method is evaluated based on its strengths and weaknesses of segmenting customers and impact on marketing performance, customer loyalty and practical application challenges such as scalability and interpretability.

A. Customer Segmentation using Machine Learning Techniques

Fig. 1 shows the importance given by this analysis of multiple machine learning algorithms including K-Means, Gaussian Mixture Models (GMM), DBSCAN, and hierarchical clustering in understanding consumer behaviour. Kmeans is very popular for its simplicity of implementation and it works well with large datasets [2],[9]. DBSCAN finds clusters of shapes varying, hierarchical clustering relate customers group, and GMM gives a probabilistic view of the data distribution. Utilizing these algorithms lets businesses build targeted marketing campaigns, improve customer experiences and ultimately create more loyal customers[4],[10].

B. From Analysis of Clustering Algorithms: Effectiveness and Limitations

The research finds that GMM often has high Silhouette scores indicating well defined clusters and strong ability to differentiate between dissimilar customer segments[3],[11]. Each algorithm has trade-offs: GMM is good, but

slow; K-Means is fast, but too prone to oversimplification; and DBSCAN is good with noise, but struggles with sparse data. Thus, each algorithm should be carefully considered as a method for segmentation as their strengths and weaknesses are highlighted[7], [12].

C. Clustering Methods and their Scalability and Complexity

Hierarchical clustering and DBSCAN to scaled enhance the scalability to get the high-level view of customer groups, showing relation and pattern for different granularity. However, as its computational complexity can be high, it becomes impractical for very large datasets where computation can take too long and resources will be over used[8],[13]. However, this limitation requires a tradeoff between computational resources constraints and detailed analysis potential since DBSCAN is very effective in detecting dense clusters while its performance for datasets of varying densities is lacking. As a result, practitioners must find the balance between model complexity, speed and the characteristics of the data[6],[14].

D. Research Directions in Customer Segmentation

The research on customer segmentation continues to explore the algorithms that serve flexibility with the ability to interpret. However, for many problems, advanced techniques such as GMM and DBSCAN[10], [12] tend to perform better than traditional methods, with regard to flexibility and detecting outliers[15].

Their complexity frustrates meaningful insights and makes these models hard to interpret. Future work should improve the interpretability as long as it does not impair efficiency or can leverage hybrid models of algorithm strengths[16]. If addressed, these challenges present an opportunity for advancement of customer segmentation methods to improve the effectiveness and informed Ness of marketing strategy choices[17].

III. METHODOLOGY

Unsupervised machine learning clustering methods such as K means and Gaussian Mixture Model along with DBSCAN are used in this study to segment customers into groups based on their purchasing behavior. Key customer features are identified with the Recency, Frequency, and Monetary (RFM) model, before applying these algorithms. Fig. 1 shows the customer segmentation process flow.

A. Data Cleaning

Preprocessing includes handling missing values, duplicates, which is a key ingredient to perform accurate customer segmentation, as removing the records with CustomerIDs missing will let us focus on the data most relevant to clustering customers [18]. To ensure data integrity, duplicate rows were removed from the dataset, and a new feature TotalPrice was created by multiplying Quantity and UnitPrice to gain a more insightful RFM dataset.

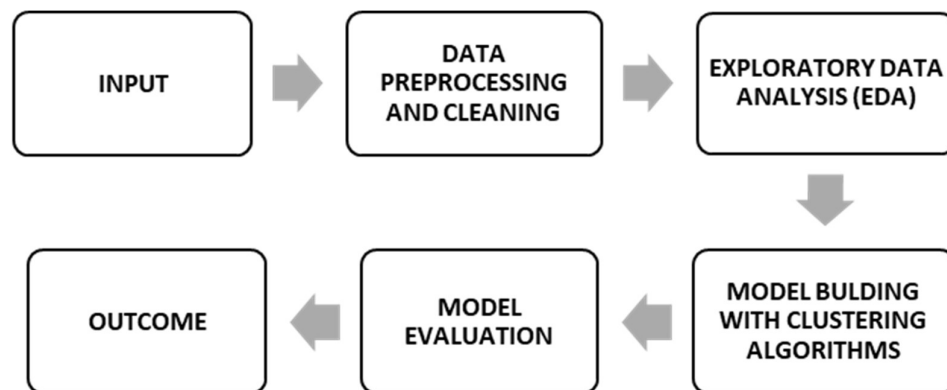


Figure 1. Customer Segmentation Process Flow

B. RFM

RFM analysis uses Recency, Frequency and Monetary scores to quantify customer value (shown in Table I), and segment customers based on RFM score weightage as Fig. 1 depicts the Recency, Frequency, Monetary values.

$$\text{RFM Score} = 0.15 \times \text{R Rank} + 0.28 \times \text{F Rank} + 0.57 \times \text{M Rank} \quad (1)$$

Prioritizing monetary value segmentation separates Top, High Value, and Medium Value groups in order to allocate resources and differentiate (bespoke) customer experience by level.

C. Clustering algorithm

K means

K-Means is a commonly used algorithm to cluster data which partitions data into clusters such that variance is minimized in each cluster minimizing the variation within each cluster[19]. It repeatedly picks a data point closest to mean of any cluster and assigns it to the cluster, recalculates these means as the centroid of their assigned points. The goal of KMeans clustering is to minimize intra cluster variance, expressed as (2).

$$\sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

Where:

μ_i is the mean of points in cluster C_i .

GMM

GMM (Gaussian Mixture Modelling): Unlike K-Means, GMM assumes clusters follow a Gaussian (normal) distribution, and thus is able to model clusters whose shapes and sizes differ. The assignment of probabilities for each data point to belong to each cluster is represented as fit each cluster to a mean and covariance, represented using a Gaussian distribution[20]. Generally speaking, this is called probabilistic clustering method where points have more than single cluster assigned with some probability.

DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a clustering method that works based on density, and therefore does particularly well with data with non-spherical shapes and varying densities [21].

Unlike Kmeans, DBSCAN doesn't require to specify no. of clustering before and it finds the core points by means of neighborhood radius eps and min sample points for density and marks other points as noise.

IV. MODEL EVALUATION

Silhouette Score: The Silhouette Score is to assess how close together or how separated other clusters are. This value can range from -1 to 1, where values closer to 1 implies dense, well separated clusters and 0 implies overlap, and negative values indicate incorrect assignments. (3) presents the mathematical explanation of the Silhouette Score.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

Where:

$a(i)$ = the mean intra-cluster distance,

$b(i)$ = smallest mean distance to any other cluster.

Davies-Bouldin Index (DBI): Davies-Bouldin Index (DBI) is a clustering assessment measure that determines the intended mean belief between clusters through by distinguishing un-interference and intercellular separation. Formula is described in (4).

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \quad (4)$$

Where:

σ_i = the dispersion of cluster c_i ,

$d(c_i, c_j)$ = the distance between cluster centers.

DBI calculation shows lower values as it shows more defined clusters and it indicates that clusters are compact and are spaced apart from each other [22].

IV. RESULT

A. Data cleaning

The first phase consisted of cleaning of the dataset to have the data of high quality and reliability by removing blank entries, duplicates etc.

B. RFM Analysis

Next RFM analysis was carried out to evaluate customers based on view of Recency, Frequency, and Monetary dimension, in which Recency reflected active customers and Frequency indicated loyalty. Table I shows the resulting Calculated metrics.

TABLE I. RFM COMPONENTS

CustomerID	Recency	Frequency	Monetary
12346	325	2	0.00
12347	38	182	4310.00
12348	74	31	1797.24
12349	18	73	1757.55
12350	309	17	334.40

TABLE II. CUSTOMER SEGMENTATION BASED ON RFM_SCORE

CustomerID	RFM Score	Customer segment
12346	0.15	Lost customers
12347	4.35	High value customer
12348	2.25	Low value customer
12349	3.49	Medium value customer
12350	1.14	Lost customer

Finally, the weight normalized metrics were used to generate a composite RFM score in Table II. As shown in Fig.2.

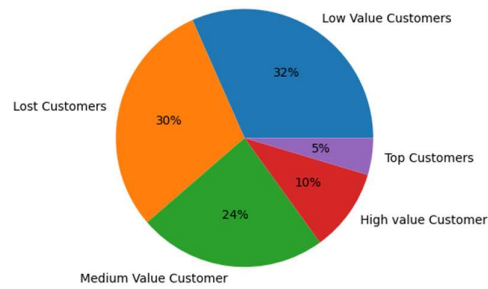


Figure 2. Visualization of customer segments

C. Clustering Algorithm

K-Means

After standardizing features using MinMaxScaler, we applied the KMeans algorithm in the customer segmentation project. In Fig. 3 I plotted inertia to determine the optimal number of clusters using elbow method. Silhouette scores analysis showed that 2 clusters have high score 0.7168 but a Davies Bouldin index high 0.5624, meaning the clusters are not compact, while 5 clusters got a silhouette score 0.6108 and a Davies Bouldin index low 0.5106, which is better balance (Figure 4). The silhouette and Davies-Bouldin index scores for varying cluster numbers are presented in Table III.

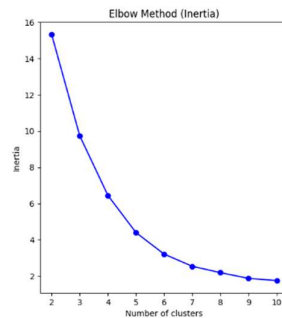


Figure 3. Elbow method

TABLE III. SILHOUETTE AND DBI SCORE FOR K- MEANS

Number of clusters	silhouette score	Davies-Bouldin index
2 clusters	0.71	0.56
3 clusters	0.62	0.58
4 clusters	0.60	0.54
5 clusters	0.61	0.51

TABLE IV. SILHOUETTE AND DBI SCORE FOR GMM

Number of components	silhouette score	Davies-Bouldin index
2 clusters	0.987	0.0092
3 clusters	0.757	0.4881
4 clusters	0.758	0.4031
5 clusters	0.7628	0.3764

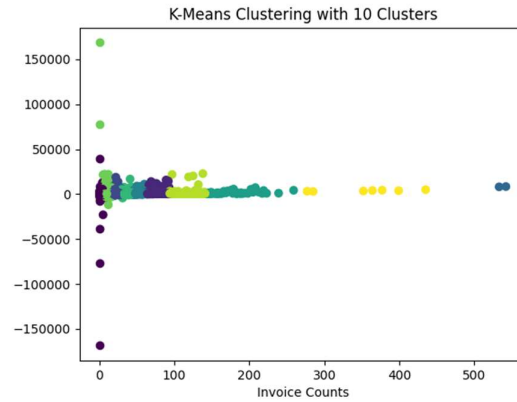


Figure 4. K-means algorithm visualisation

Gaussian Mixture Model (GMM)

After doing PCA for the dataset and reducing the dataset to two components, we applied the Gaussian Mixture Model (GMM). Table IV shows the silhouette score that GMM achieves 0.7628 when runs 2 to 5 components, while the Davies Bouldin index for K Means at a value of 0.3764. Fig. 5 shows that GMM's softer cluster boundaries allowed a more nuanced view of customer groups than K-Means.

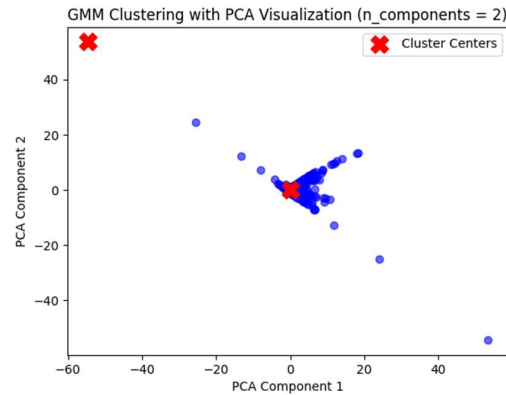


Figure 5. GMM result visualisation

DBSCAN

To find irregular clusters shaped irregularly, the data were standardized and grid searched for the optimal parameters to use for a DBSCAN, a density-based clustering algorithm. Using those values an eps of 0.3 and min_samples of 15 lead to a silhouette score of 0.678. Fig. 6 shows the visualization of the DBSCAN results. Customer segmentation using silhouette scores showed that GMM was better at handling overlapping clusters as

compared to K-Means and DBSCAN did not have a spherical assumption. Also, as shown in Table V, combining the regular methods with a more advanced machine learning provides the best initial segmentation, as shown in Table V.

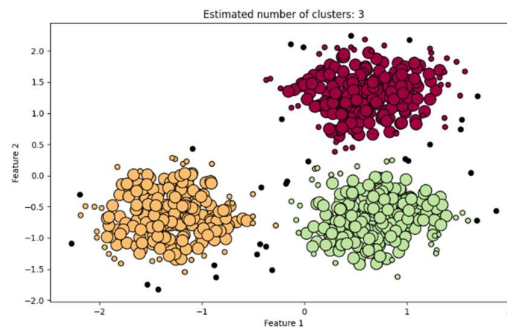


Figure 6. DBSCAN visualisation.

V. DISCUSSION

Collect data from an online retail market, pre-processing the same to remove unreliable entries; that is, missing values and duplicates. Preliminary data exploration reveals patterns of customer buy frequency and seasonal sales-the latter is depicted through several charts. Customer behaviours are also measured in terms of RFM scoring. The data will then go through the clustering algorithms K-Means and DBSCAN, but the optimization of the parameters will enhance the efficiency of the segmentation. The silhouette scores in addition to elbow methods for analysing the quality of the clusters will ensure that excellent segregation is achieved with a wide intra-cluster and inter-cluster distinction. Finally, visuals steer attention to very high-spend, high-frequency buying customer groups.

TABLE V. COMPARISON OF CLUSTERING ALGORITHMS FOR CUSTOMER SEGMENTATION

RESEARCH STUDIES	CLUSTERING ALGORITHM	SILHOUETTE SCORE
EXISTING STUDIES	K-means [5]	0.70
	K-means [12]	0.68
	K-means [17]	0.65
	GMM [14]	0.78
	GMM [19]	0.76
	DBSCAN [11]	0.64
	DBSCAN [20]	0.66
OUR STUDY	DBSCAN [21]	0.62
	K-means	0.71
	GMM	0.98
	DBSCAN	0.67

VI. CONCLUSION

This study analysed UK retail customer segmentation using RFM analysis and three clustering methods: K-Means, GMM, and DBSCAN. The work with the data cleaning made a solid dataset, and the Exploratory Data Analysis (EDA) showed us the shopping behaviours, in order to segment customers into high, mid and low value groups. With a silhouette score of 0.98 GMM outperformed the others, well modelling customer behaviour. DBSCAN did well in complex structure, K-Means did well on simpler cluster. Targeted marketing can be easily done by combining GMM with RFM analysis. Future research will focus on real time application and further techniques like deep learning to identify better pattern. For the cluster model to adapt to changing customer needs it is critical to refine the model continuously.

REFERENCES

- [1] N. Hicham and S. Karim, "Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022, Doi: 10.14569/IJACSA.2022.0131016.

- [2] B. Turkmen, "Customer Segmentation with Machine Learning for Online Retail Industry," *The European Journal of Social and Behavioural Sciences*, vol. 31, no. 2, pp. 111–136, Apr. 2022, Doi: 10.15405/ejsbs.316.
- [3] S. Miao, X. Chen, X. Chao, J. Liu, and Y. Zhang, "Context-based dynamic pricing with online clustering," *Prod Oper Manag*, vol. 31, no. 9, pp. 3559–3575, Sep. 2022, Doi: 10.1111/poms.13783.
- [4] R. Hadhoud and W. A. Salameh, "How Business Intelligence Can Help You to Better Understand Your Customers," *International Journal of Business Intelligence Research*, vol. 11, no. 1, pp. 50–58, Jan. 2020, Doi: 10.4018/IJBIR.2020010104.
- [5] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, Apr. 2016, Doi: 10.1098/rsta.2015.0202.
- [6] S. Ren, Y. Zhang, Y. Liu, T. Sakao, D. Huisin, and C. M. V. B. Almeida, "A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research directions," *J Clean Prod*, vol. 210, pp. 1343–1365, Feb. 2019, Doi: 10.1016/j.jclepro.2018.11.025.
- [7] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, Oct. 2023, Doi: 10.3390/analytics2040042.
- [8] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, no. 3, pp. 527–570, Sep. 2023, Doi: 10.1007/s10257-023-00640-4.
- [9] A. Seetharaman, I. Niranjana, V. Tandon, and A. S. Saravanan, "Impact of big data on the retail industry," *Corporate Ownership and Control*, vol. 14, no. 1, pp. 506–518, Nov. 2016, Doi: 10.22495/cocv14i1c3p11.
- [10] Z. Wang, "Customer Segmentation Based on Machine Learning Methods," *Highlights in Science, Engineering and Technology*, vol. 92, pp. 126–132, Apr. 2024, Doi: 10.54097/g70xqb16.
- [11] C. Vidden, M. Vriens, and S. Chen, "Comparing clustering methods for market segmentation: A simulation study," *Applied Marketing Analytics: The Peer-Reviewed Journal*, vol. 2, no. 3, p. 225, Sep. 2016, Doi: 10.69554/BUKQ9565.
- [12] X. Lin, W. Guan, and Y. Zhang, "Application of Data Mining Technology with Improved Clustering Algorithm in Library Personalized Book Recommendation System," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023, doi: 10.14569/IJACSA.2023.0141151.
- [13] H. Ibrahim Hayatu, A. Mohammed, and A. Barroon Isma'eel, "Big Data Clustering Techniques: Recent Advances and Survey," in *Machine Learning and Data Mining for Emerging Trend in Cyber Dynamics*, Cham: Springer International Publishing, 2021, pp. 57–79. doi: 10.1007/978-3-030-66288-2_3.
- [14] C. X. Gao et al., "An overview of clustering methods with guidelines for application in mental health research," *Psychiatry Res*, vol. 327, p. 115265, Sep. 2023, doi: 10.1016/j.psychres.2023.115265.
- [15] U. Sharma, G. Aditi, N. R. Roy, and S. N. Singh, "Analysis of Customer Segmentation Clustering Techniques," in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Jan. 2022, pp. 374–379. doi: 10.1109/Confluence52989.2022.9734147.
- [16] S. Chandra, S. Verma, W. M. Lim, S. Kumar, and N. Donthu, "Personalization in personalized marketing: Trends and ways forward," *Psychol Mark*, vol. 39, no. 8, pp. 1529–1562, Aug. 2022, doi: 10.1002/mar.21670.
- [17] S. Park and H. M. Kim, "Data-Driven Customer Segmentation Based On Online Review Analysis and Customer Network Construction," in *Volume 3A: 47th Design Automation Conference (DAC)*, American Society of Mechanical Engineers, Aug. 2021. doi: 10.1115/DETC2021-70036.
- [18] W. Xia et al., "A Comprehensive Study of the Past, Present, and Future of Data Deduplication," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1681–1710, Sep. 2016, doi: 10.1109/JPROC.2016.2571298.
- [19] A. Solichin and G. Wibowo, "Customer Segmentation Based on Recency Frequency Monetary (RFM) and User Event Tracking (UET) Using K-Means Algorithm," in *2022 IEEE 8th Information Technology International Seminar (ITIS)*, IEEE, Oct. 2022, pp. 257–262. doi: 10.1109/ITIS57155.2022.10009981.
- [20] "Issue Information," *J Biogeogr*, vol. 48, no. 2, Feb. 2021, doi: 10.1111/jbi.13887.
- [21] S. S. Ling, C. W. Too, W. Y. Wong, and M. H. Hoo, "Customer Relationship Management System for Retail Stores Using Unsupervised Clustering Algorithms with RFM Modeling for Customer Segmentation," in *2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, IEEE, May 2024, pp. 1–6. doi: 10.1109/ISCAIE61308.2024.10576353.
- [22] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Mar. 2020, pp. 306–310. doi: 10.1109/ICCMC48092.2020.ICCMC-00057.