# Revolutionizing Customer Segmentation: A Synergy of Software Engineering and SaaS Innovation

Jill Padalia
*Institute of Technology*
*Nirma University*
Ahmedabad, India
21bce173@nirmauni.ac.in

Nemi J. Patel
*Institute of Technology*
*Nirma University*
Ahmedabad, India
21bce210@nirmauni.ac.in

***Abstract*—This study looks into Software as a Service (SaaS) in the field of software engineering to help us learn more about how to divide customers into groups. We use advanced clustering methods, such as K-means clustering, Gaussian Mixture Model (GMM) clustering, and Agglomerative Hierarchical Clustering, to look into the complex patterns and tastes of SaaS users. Applying the elbow method helps figure out the best number of groups, which improves the accuracy of the segmentation process. In addition, the study presents the silhouette score as a way to measure and choose the best clustering algorithm for our dataset. Through careful comparison and analysis, this study not only finds separate groups of customers, but it also finds the clustering method that best captures the natural structures in the SaaS customer landscape. The results are very important for personalized service delivery, targeted marketing, and future progress in the fast-paced field of software engineering in the SaaS realm.**

## I. Introduction

In the rapidly changing environment of modern business, the efficient application of technology is of the utmost importance for obtaining and preserving a position of competitive advantage. The application of software engineering, which provides a methodical and disciplined approach to the process of developing, designing, testing, and maintaining software systems, plays an essential part in this endeavor. When companies make efforts to improve their operations, the incorporation of reliable software engineering processes becomes increasingly crucial. The proposed framework examines the application of software engineering principles to the field of customer segmentation. It demonstrates how a customer segmentation software solution that is well-architected can transform the way businesses understand and engage with their client base.

In recent years, Software as a Service, often known as SaaS, has emerged as a disruptive paradigm in the software industry. SaaS provides businesses with solutions that are both scalable and cost-effective to fulfill the different requirements of their operations. Software as a service (SaaS) is a delivery model for application software that uses the internet rather than traditional on-premises installations and maintenance. This paradigm shift not only simplifies the delivery of software but also offers up new opportunities for developing solutions that are both efficient and agile with a focus on the client. The proposed framework investigates the application of software as a service (SaaS) to the optimization of customer segmentation methods. It demonstrates how using a cloud-based approach can improve the responsiveness and agility of organizations operating in today's fast-paced marketplaces.

The process of breaking a client base into discrete groups based on shared qualities, behaviors, or preferences is known as customer segmentation. client segmentation is a key marketing technique. Customer segmentation takes on a greater significance when considered in the context of software as a service (SaaS). Businesses can acquire deeper insights into the behavior, preferences, and requirements of users of their SaaS platforms if they apply advanced algorithms and analytical tools within those platforms. This, in turn, makes it possible to cater services, messages, and features to distinct groups of customers. Not only does the deployment of customer segmentation in SaaS maximize the allocation of resources, but it also increases customer happiness by delivering solutions that are specifically tailored to their needs.

Customer segmentation can be applied by different clustering algorithms. Some of them are K-means clustering, Hierarchical Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Mean Shift [1] [2], Agglomerative Clustering, Gaussian Mixture Model (GMM), Self-Organizing Maps (SOM) [3] [4] [5], Agglomerative Nesting, Fuzzy C-Means (FCM) [6] [7] [8], and OPTICS (Ordering Points To Identify the Clustering Structure). Agustino [9] discusses customer segmentation based on RFM analysis and also compares DBSCAN and Affinity Propagation algorithms but the problem is DBSCAN algorithm cannot handle high-dimensional data, the affinity Propagation algorithm may produce a large number of clusters. Aerts [10] focuses on customer segments in alternative finance using DBSCAN and agglomerative clustering but some limitations are that DBSCAN is ineffective in detecting multiple segments and agglomerative clustering is ineffective in tracking clusters. Wilson [11] discusses customer segmentation based on temporal variation and an LDA-based

model for clustering short temporal behavior sequences, but limitations around scalability and accurately representing temporal behavior sequences (TBS). Yang [12] discusses customer segmentation using distance based clustering and parametric mixture models but these models are not suitable for behavioral pattern-based segmentation and lack of focus on natural clusters in data. Zhou [13] uses decision tree algorithm for bank customer classification and rough set theory to reduce overfitting but the limitations are there is overfitting in traditional decision tree algorithms and time-consuming pruning step in traditional decision tree algorithms.

Our study focuses on advanced clustering algorithms, such as K-means, agglomerative, and Gaussian Mixture Model (GMM) clustering, to fix the problems with traditional customer segmentation methods. Traditional methods, like demographic or behavioral research, don't always work well when trying to understand how people act today. We use these complex clustering methods in our study and look into the benefits of each one. K-means is a centroid-based method that works well for segmenting data, especially when the data set is big. With its hierarchical merging, agglomerative clustering shows how different customer groups are connected and depend on each other. As a probabilistic model, Gaussian Mixture Model (GMM) clustering can handle uncertain data and give a more nuanced picture of customer groups. We found that this comparison helps us understand the pros and cons of these algorithms, which gives businesses a better way to divide customers in a way that works with the complicated way markets work today.

## II. Methodology

### A. Data Source

| | CustomerID | Gender | Age | Annual Income ($) | Spending Score (1-100) | Profession | Work Experience | Family Size |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Male | 19 | 15000 | 39 | Healthcare | 1 | 4 |
| **1** | 2 | Male | 21 | 35000 | 81 | Engineer | 3 | 3 |
| **2** | 3 | Female | 20 | 86000 | 6 | Engineer | 1 | 1 |
| **3** | 4 | Female | 23 | 59000 | 77 | Lawyer | 0 | 2 |
| **4** | 5 | Female | 31 | 38000 | 40 | Entertainment | 2 | 6 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **1995** | 1996 | Female | 71 | 184387 | 40 | Artist | 8 | 7 |
| **1996** | 1997 | Female | 91 | 73158 | 32 | Doctor | 7 | 7 |
| **1997** | 1998 | Male | 87 | 90961 | 14 | Healthcare | 9 | 2 |
| **1998** | 1999 | Male | 77 | 182109 | 4 | Executive | 7 | 2 |
| **1999** | 2000 | Male | 90 | 110610 | 52 | Entertainment | 5 | 2 |

Fig. 1: Dataset structure

The dataset of the proposed approach is taken from Kaggle. In this proposed approach, we give a complete analysis of business Customer Data, which provides insights into the characteristics of clients who are suitable for an inventive business. The dataset consists of 2000 records and contains eight essential fields, including Customer ID, Gender, Age, Annual Income, Spending Score (a statistic issued by the shop based on customer behavior and the type of spending), Profession, Work Experience (in years), and Family Size as shown in Fig.1. The use of membership cards makes it easier to collect consumer information, which in turn enables the proprietor of a store to obtain a profound understanding of the people who frequent their establishment.

### B. Data preprocessing

The first step of data preprocessing includes loading the dataset for data exploration. To ensure the quality of data is not compromised, a search for missing values is carried out. The null entries present in the dataset are removed by removing the rows with missing values. After cleaning the dataset a pie chart is utilized in order to investigate the demographic distribution of gender as shown in Fig.2. This chart shows the graphical depiction of the percentage of male(40.7%) and female(59.3%) clients. In addition, a boxplot is used to explore the distribution of ages across gender groups as shown in Fig.3. This helped shed light on any possible age-related patterns that may exist among the client base. Further Preprocessing is done by plotting a pie chart that provides a visual representation of the diverse range of occupations held by individuals, offering valuable insight into the distribution of employment kinds as shown in Fig.4. Fig.5 shows the box plot of the distribution of age across professions. This lets us look at the ages of people in different jobs, which shows us possible connections between age and certain job types. Also, another boxplot(Fig.6) is taken of annual income Distribution across different professions for further analysis.

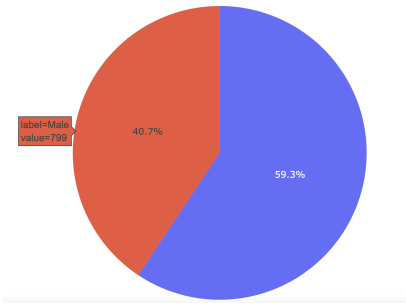


Fig. 2: Gender distribution

### C. Proposed approach

Once the data is known and preprocessed, the clustering algorithms are applied. The focus is reduced to two essential features, notably 'Annual Income' and 'Spending Score (1-100),' which results in the formation of a subset from the dataset for a more targeted investigation. This allows for a more thorough analysis. In order to promote meaningful and consistent comparisons, a min-max scaling is implemented in order to normalize these particular properties. The subsequent grouping analyses will have a higher degree of success if the values are first subjected to this standardization process, which converts them to a similar scale.

The proposed approach consists of different clustering algorithms which are Kmeans clustering, Agglomerative hierarchical clustering, and Gaussian mixture model clustering. The clustering method is applied to identify separate consumer

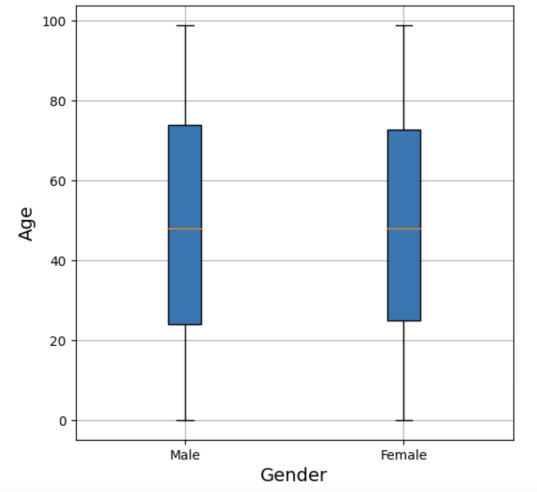


Fig. 3: Box plot

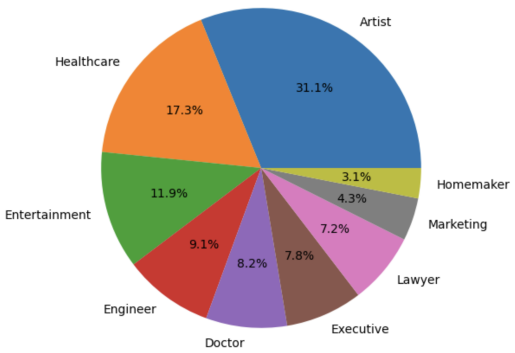


Fig. 4: Distribution of Profession Data Values

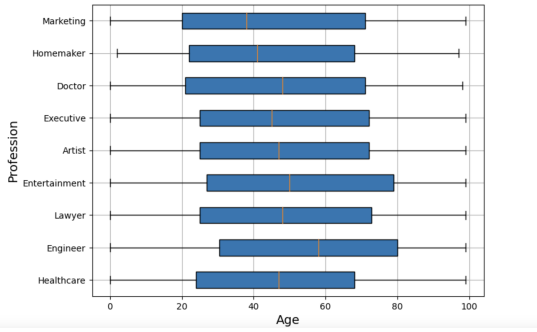


Fig. 5: Age Distribution across Professions

groupings based on the 'Annual Income' and 'Spending Score' attributes. An iterative application of the technique is performed for a variety of cluster numbers (K) to zero in on the ideal grouping that results in the lowest possible sum of squared errors. The graph shown in Fig.7 illustrates the sum of squared errors for a variety of cluster counts and is extremely helpful in determining an appropriate value for K. From the Fig.7, the number of clusters is known to be 4. Once the number of clusters is determined K means algorithm is

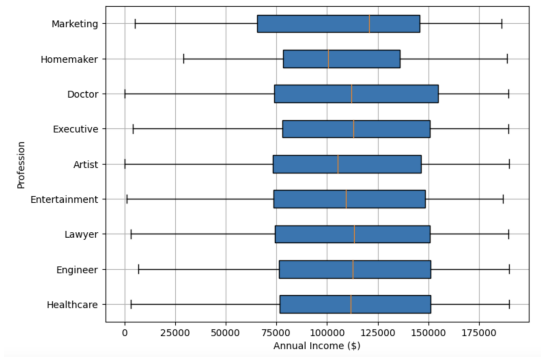


Fig. 6: Age Distribution across Professions

applied to place each data point into the appropriate cluster. After that, the findings of the clustering were merged back into the dataset in a completely seamless way as shown in Fig.8, which provided a comprehensive picture of how consumers are grouped according to their spending patterns and annual income. All three clustering algorithms are applied after knowing the grouping of customers according to clusters.

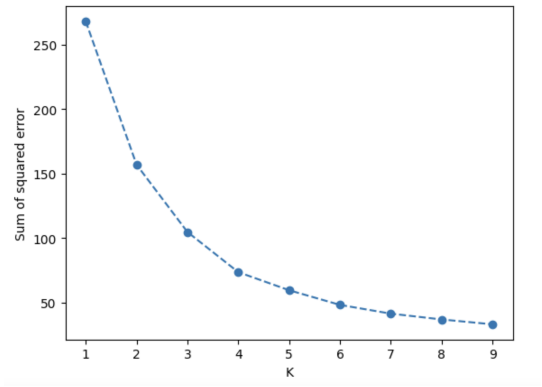


Fig. 7: Identifying number of clusters

| | Annual Income ($) | Spending Score (1-100) | cluster |
|---|---|---|---|
| **0** | 0.078958 | 0.39 | 3 |
| **1** | 0.184236 | 0.81 | 0 |
| **2** | 0.452694 | 0.06 | 3 |
| **3** | 0.310569 | 0.77 | 0 |
| **4** | 0.200027 | 0.40 | 3 |

Fig. 8: Predicted clusters

The Silhouette Score is applied to determine how effective each technique for clustering is by comparing how well it clustered the data. The Silhouette score of K means, Gaussian Mixture Model, and Agglomerative Clustering are 0.3779, 0.3732, and 0.3256 respectively. After the computations are completed, it is found that the K-Means clustering methodology has the greatest Silhouette Score compared to all of

the other approaches that are evaluated. Given the dataset and the parameters that are selected, this finding shows that the K-Means approach provides the most distinct and well-defined clusters, which establishes it as the preferred clustering technique for this particular investigation. The Flowchart of the proposed approach is Fig.9

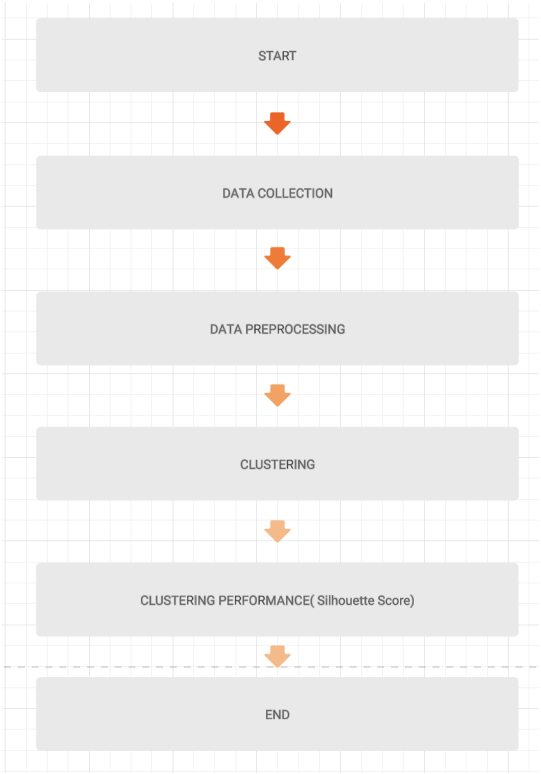


Fig. 9: Flow Chart

## III. RESULTS AND DISCUSSIONS

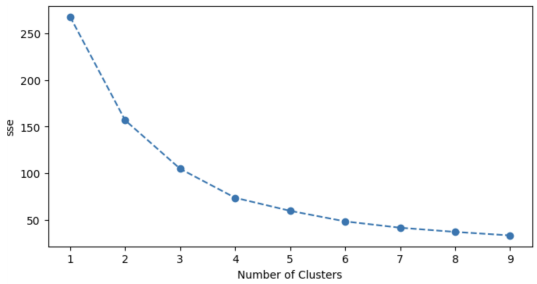


Fig. 10: Elbow method

Fig.10 shows the graph of the Elbow method for determining the number of clusters. The elbow approach is a method that is used in clustering analysis as a strategy to estimate the optimal amount of clusters that should be present in a dataset. It requires executing a clustering method, such as K-means, for a range of different cluster numbers and then plotting the within-cluster sum of squares (WCSS) against the number of clusters. The WCSS is the value that is calculated by adding up the squared distances that separate each data point from the cluster centroid to which it is assigned. When there are more clusters, the weighted cluster sum score (WCSS) goes down, which results in a plot that often has a sloping downward direction. The elbow point on the curve, on the other hand, indicates the ideal cluster number which is 4 in this approach as shown in Fig.10, which is the point at which the decrease in WCSS starts to plateau. This technique provides a visually intuitive approach to establishing a balance between the complexity of the model and the ability to capture relevant patterns in the data. As a result, it directs the selection of an appropriate number of clusters for analysis.

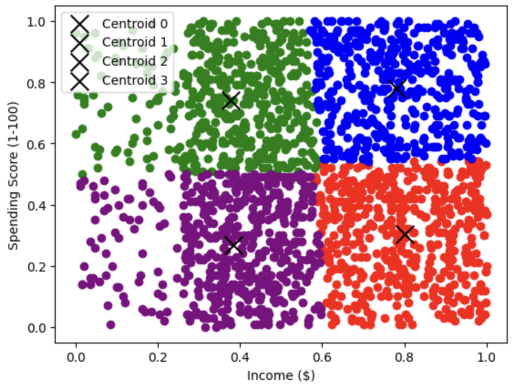


Fig. 11: K means clustering

Fig.11 shows that each datapoint belongs to which cluster after applying K means clustering. Since the elbow method showed that 4 is the best number for K, the next step in using the K-means clustering algorithm is to carefully place four initial centers in the feature space. Then, each data point is put into the cluster whose center is closest to it. This starts the process of making preliminary clusters. The algorithm improves these groups over and over again by finding the new centroids as the average of the data points that belong to each one. The iterative process keeps going until there are no more small changes in the places of the centers. This is called convergence. The main goal is to lower the total within-cluster sum of squares (WCSS), which is the square of the distances between data points and the centers of their clusters.

Fig.12 shows that each datapoint belongs to which cluster after applying Agglomerative Hierarchical clustering. Agglomerative Hierarchical Clustering works from the bottom up, with each data point starting as its cluster. The algorithm then figures out how different two groups are from each other using certain metrics, like Euclidean distance or linkage methods like Ward, complete, or average linkage. The number of groups is decreased by merging the two most similar ones first, and then the dissimilarity matrix is updated to reflect this. This process is done again and again until all the data points are in one group. The process of clustering can be seen in a dendrogram, where the height of the united branches shows

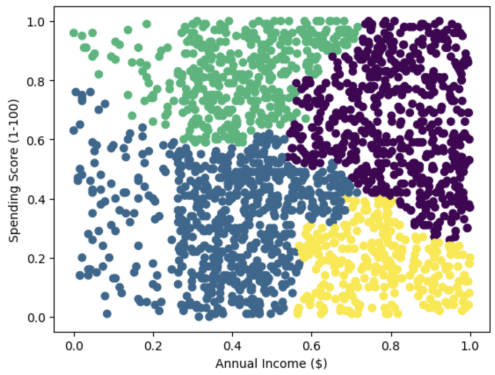


Fig. 12: Agglomerative clustering

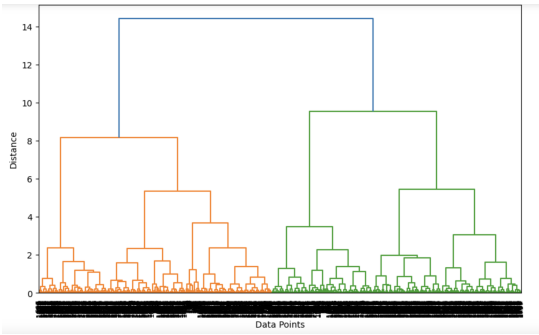


Fig. 13: Agglomerative Hierarchical clustering Dentogram

how different they are. With the dendrogram visualization of Agglomerative Hierarchical Clustering(Fig.13), you can easily look at different cluster numbers by cutting the dendrogram at different heights. Hierarchical clustering is easy to understand, but it can be hard to run on a computer, especially for big datasets. The results of the clustering depend on the linkage method and distance metric that are used.

Fig.14 shows that each datapoint belongs to which cluster

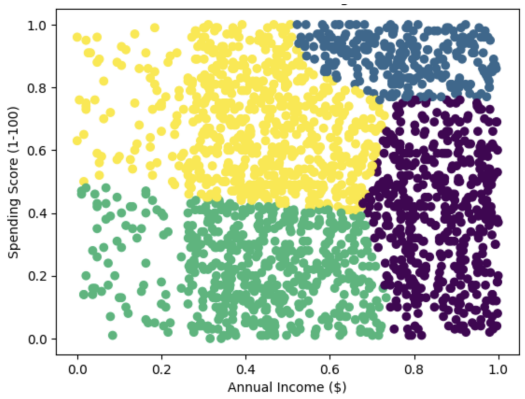


Fig. 14: GMM clustering

after applying Gaussian mixture model clustering. GMM clustering is a probabilistic method that works with the idea that data comes from a mix of Gaussian distributions. This lets clusters be of different sizes and forms. The Expectation-Maximization (EM) algorithm is run over and over by the algorithm. It starts by setting up the parameters, such as the mean, covariance, and weights for each Gaussian component. This is done by finding the chance that each data point belongs to each Gaussian component (E-step) and then changing the parameters to make the observed data more likely (M-step). It keeps going until convergence, at which point the parameters become stable. GMM uses a soft assignment to give odds to each data point in each cluster. This gives us a more complete picture of uncertainty. The method works well with datasets that have complicated structures and clusters that overlap. It can find elliptical or non-spherical clusters and give information about covariance structures. Like K-means, GMM is sensitive to how the parameters are set up, and it often takes more than one setup to find the global best solution.

## Conclusion

In summary, the research presented here has successfully implemented consumer segmentation within the context of the Software as a Service (SaaS) industry, more specifically within the field of software engineering. We set out to identify unique client subgroups by utilizing the potent methods of K-means clustering, Gaussian Mixture Model (GMM) clustering, and Agglomerative Hierarchical Clustering. The criteria for these subgroups were based on significant characteristics. The incorporation of the elbow approach into the process of segmentation offered an essential mechanism for identifying the optimum number of clusters, which increased the accuracy of our studies. We were able to designate important client segments through the detailed use of these clustering approaches, which shed light on patterns and preferences within the software engineering SaaS market. These discoveries have important repercussions for specialized service delivery, marketing tactics, and product development, which helps nurture an approach that is more informed and targeted in the ever-changing ecosystem of Software as a Service. In the area of software engineering, the combination of more complex clustering approaches with the elbow method has not only helped us gain a deeper understanding of consumer behavior, but it has also prepared the road for future refinements and breakthroughs in client segmentation.

## Future Work

In the context of software engineering, the current research constitutes a significant step toward understanding client segmentation inside the Software as a Service (SaaS) landscape. Nevertheless, there are interesting directions that could be pursued in the course of future research and development. To begin, a more in-depth study might include the incorporation of more clustering algorithms, as well as a comparison with the approaches that are already in place, and possibly the incorporation of newly developed techniques in the fields of machine learning and data analytics. In addition, the implementation of sophisticated predictive modeling techniques, such as machine learning classifiers, has the potential to improve both the precision and the granularity of consumer

segmentation. The introduction of time-series analysis into the segmentation framework is necessary because investigating temporal patterns and trends in customer behavior could also yield useful insights and is therefore required. In addition, a more in-depth examination into the elements that influence customer happiness and loyalty within each identified group could be of great assistance in efficiently designing SaaS products. Last but not least, as the SaaS landscape continues to change, it will be vital to make continuous updates and improvements to the segmentation strategy in order to accommodate evolving trends and technologies. This will ensure that the findings continue to be relevant and actionable in the dynamic area of software engineering that falls under the umbrella of the SaaS domain.

## REFERENCES


[1] "Mean shift algorithm to determine customer segmentation in online store sales," 2021.

[2] "Optimization of segmentation algorithms through mean-shift filtering preprocessing," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 3, pp. 622–626, 2014.

[3] "Using self organizing maps and k means clustering based on rfm model for customer segmentation in the online retail business," 2020.

[4] "Multi-behavior rfm model based on improved som neural network algorithm for customer segmentation," *IEEE Access*, vol. 10, pp. 122501–122512, 2022.

[5] "Customer segmentation based on self-organizing maps: A case study on airline passengers," *Journal of Aeronautics and Space Technologies*, vol. 13, no. 2, pp. 227–233, 2020.

[6] "Customer segmentation with rfm model using fuzzy c-means and genetic programming," *Matrik: jurnal manajemen, teknik informatika, dan rekayasa komputer*, vol. 22, no. 2, pp. 239–248, 2023.

[7] "Customer segmentation using fuzzy c-means algorithm in telco industry," 2022.

[8] "Clustering customer data using fuzzy c-means algorithm," vol. 9, no. 1, pp. 1–14, 2021.

[9] "Comparison of the dbscan algorithm and affinity propagation on business incubator tenant customer segmentation," *Jurnal Sistem Informasi dan Komputer*, vol. 12, no. 2, pp. 315–321, 2023.

[10] "Tracking customer segments in alternative finance using time-evolving cluster analysis," 2020.

[11] "Clustering short temporal behaviour sequences for customer segmentation using lda," *Expert Systems*, vol. 35, no. 3, 2018.

[12] "Behavioral pattern-based customer segmentation," 2009.

[13] "Bank customer classification algorithm based on improved decision tree," pp. 30–33, 2022.