2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS) | 979-8-3315-0546-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CSITSS64042.2024.10816952

Optimizing Customer Segmentation: A Comparative Analysis of Clustering Algorithms Using Evaluation Metrics

Noor Sakina
dept. of Computer Science and Engineering
RV College of Engineering
Bengaluru, India
noorsakina.cd22@rvce.edu.in

Ajith P Arun
dept. of Chemical Engineering
RV College of Engineering
Bengaluru, India
ajitharun.ch22@rvce.edu.in

Pavithra R
dept. of Electronics and Communication Engineering
RV College of Engineering
Bengaluru, India
pavithrar.ec22@rvce.edu.in

Vishalakshi Prabhu H
dept. of Computer Science and Engineering
RV College of Engineering
Bengaluru, India
vishalaprabhu@rvce.edu.in

Praveen Kumar Gupta
dept. of Biotechnology
RV College of Engineering
Bengaluru, India
praveenkgupta@rvce.edu.in

***Abstract*—Customer segmentation is the route to data-driven decisions within organizations serving a wide variety of customer needs. This research explores customer segmentation with the utilization of K-means, hierarchical clustering, fuzzy C-means, along with DBSCAN to behavioural and demographic data. These algorithms are measured against each other using their silhouette score, Davies-Bouldin index, and cluster cohesion value. These results highlight that clustering can reveal meaningful patterns in customer data for improving personalized marketing strategies, customer retention, and profitability.**

Keywords—Clustering algorithm, Cluster cohesion, Customer segmentation, Davies-Bouldin Index, DBSCAN, Fuzzy C-Means Clustering, Hierarchical Clustering, K-Means clustering, Silhouette Score

I. INTRODUCTION

Customer segmentation is based on grouping customers with some common characteristics, thus helping the business devise measures that will facilitate in getting an review of what the customer wants and needs [1]**.** Understanding the needs of consumers and behaviours is a key facet in highly competitive modern markets toward product and service development. Customer segmentation helps an organization to subgroup their customers into unique clusters with similar traits, enabling focused marketing efforts, enhanced and personalized services, and efficient allocation of resources. Given today's explosion of data and the advances in machine learning, clustering algorithms have grown as important tools in segmenting customers based on behaviour, demographics, and transactions.

In this paper, four major clustering algorithms will be considered: Hierarchical Clustering, K-means clustering, DBSCAN Clustering, and Fuzzy Clustering-Means. All of these algorithms have several strengths and are well-matched to certain types of data. We will compare them using performance indicators including the Silhouette Score, Davies-Bouldin Index, and Cluster Cohesion, each of which has the potential to be used to evaluate how meaningful the clusters formed are Hierarchical Clustering represents data in detail but has problems related to its scalability. K-Means is very popular because of its efficiency but the count of clusters needed must be specified in advance and can be highly sensitive to initial choices and DBSCAN resists noise and continues to find clusters of different shapes. In contrast, Fuzzy C-Means allow a point to belong to multiple clusters.

In the following research, we will review clustering techniques for customer segmentation, their advantages and limitations, and real applications in the world. We intend to run all of these on real datasets to compare performances in improving customer relationship management and guiding strategic business decisions.

II. LITERATURE REVIEW

The combination of K-Means with Affinity Propagation very often contributes to a highly effective form of customer segmentation for marketing. Analytical techniques like the Silhouette score helps to establish the ideal number of clusters.[2]. The given paper compares various clustering methods like K-Means and DBSCAN. It comments on the efficiency and accuracy of K-Means while recommending further research in using more algorithms to overcome the challenges posed by large datasets [3]. RFM modelling combined with K-Means enhances CRM customer segmentation. Customer segmentation is important in CRM for targeted advertising and customer loyalty improvement [4]. A revised clustering algorithm facilitates improved telecom customer segmentation by circumventing the pitfalls of K-Means and DBSCAN and supports very focused marketing and enhancement of resources [5]**.** Text mining with K-Means efficiently analyses social media feedback for segmentation and heavily benefits the insurance sector [6]. Integrating clustering with machine learning, like Spectral clustering and GMMs, refines customer segmentation in support of personalized marketing [7].

Techniques like K-Means and BIRCH are used in classifying customers by psychographic and demographic factors to allow targeted marketing [8]. Recent research refines these traditional clustering methods like K-Means to overcome the challenges specific to the industry by incorporating text mining to further enhance knowledge about the customer [9]. A comparative study focuses on the better performance of agglomerative clustering while giving insights into the proper techniques of market segmentation [10]. RFM analysis using K-Means refines customer segmentation for e-commerce to facilitate correct resource management and customer satisfaction through personalized marketing [11].

III. Dataset Overview

The dataset chosen gives granular insight into most aspects of the sales transactions on Amazon. This includes the product category and size, which give insights into market position and inventory preference. The sale date would be useful for trend and seasonality analyses, while the sale status and fulfilment method are indicators of order completion and delivery efficiency. Style information provides data about product preference, while unique identifiers like SKU and ASIN identify the exact tracking inventories and products. The courier status gives an understanding of delivery performance. The quantity and sale amount are key to the sales volume and revenue analysis. The B2B indicator tells business from individual transactions, while the currency used is relevant for financial analysis and currency conversion. This dataset can be suited to perform in-depth analysis of sales trends, inventory management, and profitability.

IV. Methodology

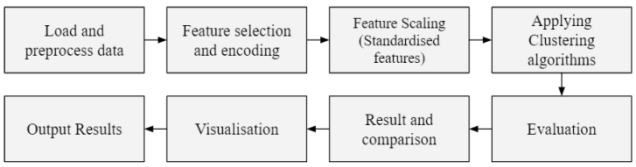

Fig. 1. Workflow diagram

Fig. 1 illustrates the workflow diagram in a stepwise manner. The data preparation is carried out first, relevant features are chosen and then the features are normalized to a common scale. Further clustering algorithms are implemented on the prepared dataset. Evaluation of the clustering model of each algorithm is performed using various evaluation metrics. The acquired results are correlated using comparison metrics and then visualized through graphs to find out the best algorithm.

V. Clustering Algorithms and Graphs Obtained

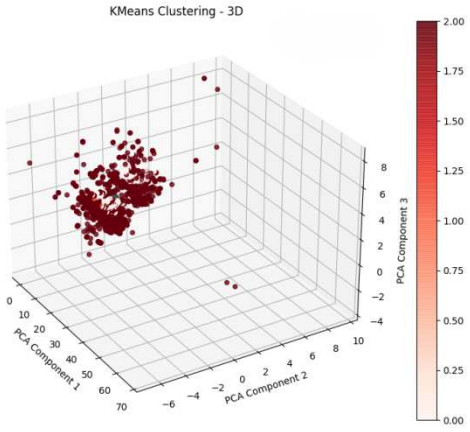

Fig. 2. Graph representing K-Means Clustering

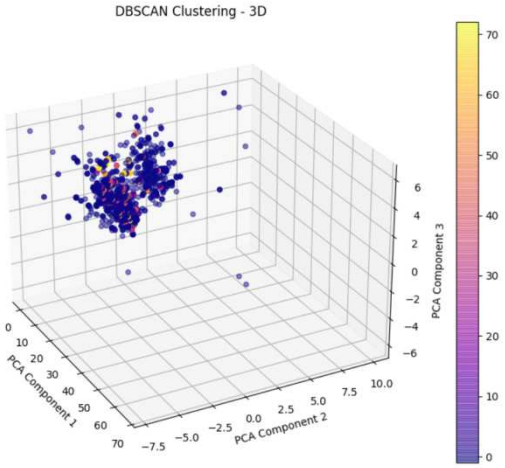

Fig.3 Graph representing DBSCAN Clustering

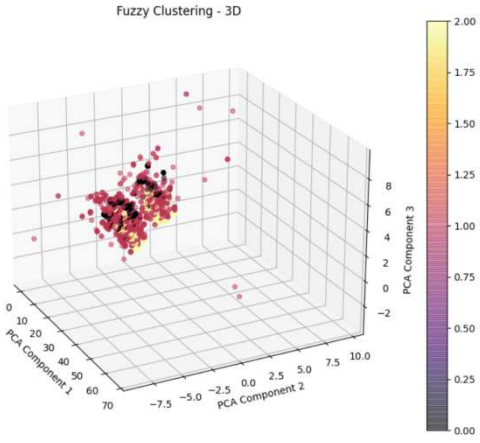

Fig. 4. Graph representing Fuzzy C-Means Clustering

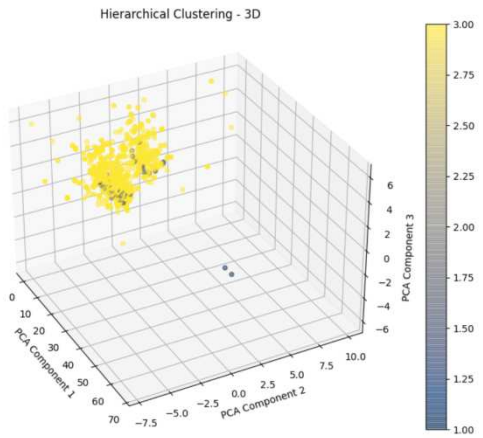

Fig.5. Graph representing Hierarchical Clustering

A. K-Means Clustering

K-Means is a clustering technique that breaks up data into K- distinct clusters. This clustering technique is a very popular method, as it groups data points in a way that, the data points of the same cluster resemble one another, more than those of the different cluster [12].

This technique is referred to as the centroid—where the value for the cluster is the mean of the objects contained within the cluster. K-Means initializes the centroid for each cluster randomly and then performs iterative adjustments to optimally place the centroids based on the training data [13]. Assignment of each data point to the closest centroid is iteratively done, followed by recalculation of centroids based on current members in a cluster. The repetition of this process continues to converge when the assignments no longer change the algorithm. K-Means is a basic clustering algorithm and probably one foremost used. The sum of Squared Errors, Silhouette Score, and Davies-Bouldin Index are commonly applied techniques for identifying the ideal count of clusters [14].

Fig. 2 shows the scatter plot that we collected for our data set on Amazon for K-Means clustering which shows well-separated and tightly grouped clusters. Therefore, the boundaries between the different groups of data points are clear. Each cluster is distinctly coloured, and the data points are in close relation with their respective centroids, which again goes on to denote a well-run algorithm for forming compact, well-defined clusters. There is a good separation, with tight groupings indicating that K-Means does a pretty good job in segmenting data. The first(X), second(Y) and third(Z) components of the PCA, shows the directions of maximum variance in the data for K-Means clustering visualization.

B. DBSCAN Clustering

DBSCAN is short for Density-Based Spatial Clustering which is an unmonitored machine learning algorithm which is frequently applied within the clustering tasks. It clusters together the closely packed points, marking the high-density regions, and classifies those of the low-density regions as outliers or simply noise. Another added advantage of DBSCAN is its identification of unusual customers with unique spending patterns. This is an excellent tool to enhance customer satisfaction and maximize profitability [15].

One among the major features of DBSCAN clustering is that it generates a scatter plot as seen in Fig.3, where clusters are of variable shape and size depending on density; some of the points will be classified as noise. Basically, the algorithm works by defining a neighbourhood around each point based on a certain specified radius or by any other user-defined distance metric [16]. Otherwise, it is an excellent algorithm in finding clusters of any shape, although the plot indicated that some clusters overlap or are less far apart as compared to K-Means. These outlier points reveal that these are low-density regions and are not considered as any cluster by DBSCAN. The X, Y and Z axes represents the three PCA components. PCA stands for principal Component analysis which shows the directions of maximum variance in data for DBSCAN clustering visualization.

C. Fuzzy C-Means Clustering

Fuzzy Clustering is a technique related to segmentation analysis in which every data point goes with all available clusters, although to different degrees of membership. Another algorithm, Fuzzy C-means, was developed in 1984 by Bezdek et al. The algorithm is an improvement over the C-means algorithm [17]. In contrast to traditional clustering methods, which place every point into a single cluster, fuzzy clustering provides more graded classification. This method employs membership values, ranging from 0 to 1, denoting the extent to which every single data entry belongs to different clusters.

The Fuzzy C-Means clustering technique yields a scatter plot represented in Fig.4 featuring intersecting clusters, where every data entry can be associated with numerous clusters, each including varying degrees of affiliation. Thus, some areas will flow into others, showing characteristic features when dealing with such uncertainty for the assignment to clusters. Overlapping clusters say something about the nature: some points just do not belong to one single cluster, and hence it captures fuzziness in cluster boundaries. Herein, X, Y and Z components stand for the three principal components of PCA which represents Principal Component Analysis. These describe the directions of maximum variance within data and are utilised for visualization purposes of fuzzy c-means clustering.

D. Hierarchical Clustering

This algorithm constructs a multilevel hierarchy of clusters by either merging or splitting existing clusters by incorporation of analogous attributes or vice versa. The underlined principle of the hierarchical clustering algorithm relies on grouping similar features together, or separating them depending on their dissimilarities [18]. While traditional clustering techniques imply that one data point is designated for a unique cluster, the hierarchical clustering technique allows the formation of interconnected clusters at levels of similarity. It starts by treating every data point as an independent cluster and gradually It then does the successive merging of the closest pairs of clusters until the whole dataset is combined into one cluster. In the process, a dendrogram can be utilised for visualizing the hierarchical relationships

amongst the clusters and the distances at which the clusters are combined.

A hierarchical clustering scatter plot as shown Fig.5 is always divided into distinct clusters that differ in their resemblance with one another, with the X, Y and Z axes representing the three principal components PCA derived from the composite set of points. These components allow the viewers to see the greatest variation directions vividly, making the cluster structure more understandable. Hierarchical clustering, in contrast to techniques like fuzzy clustering, is a clear-cut of clusters, no overlapping is observed as each point fits to exactly one cluster at each level of the hierarchy. The structure of the chart exposes the inherent connections and separations within the data, which in turn can be looked at as the arrangements and groupings pattern of the list.

The graph of K-Means represents well-separated, clearly identifiable clusters and hence a perfect result of the segmentation. DBSCAN shows several shapes and densities of clusters consisting of some noise points, hence it can handle irregular and noisy clusters. Fuzzy C-Means shows the allowed clusters to allow overlap, flexibility in terms of fuzzy boundaries. Hierarchical clustering represents well-delineated, nested clusters and hence reflects its strength in the uncovering of multi-level relationships. Graphs from the various clustering algorithms have closeness in appearance particularly because this lower-dimensional representation employs the approach of PCA is to transform the high-dimensional data to a lower dimensional capturing most of the variances in data. PCA helps clustering by reducing dimensionality, maintaining important variance, enhances the performance of algorithms, and makes visualization of data easier. Same principal components are utilised throughout all the clustering methods, hence the scatter plots are showing similar variance-based projections of the data. Although the underlying algorithms may vary in the way they form clusters, the underlying PCA transformation focuses more on the variance rather than the actual structure of the clustering, hence making resultant scatter plots look so similar.

VI. COMPARATIVE ANALYSIS OF CLUSTERING ALGORITHMS USING SILHOUETTE SCORE, DAVIES-BOULDIN INDEX, AND CLUSTER COHESION

TABLE 1. Comparative Performance Metrics of Various Clustering Algorithms

Algorithm	Silhouette Score	Davies-Bouldin Index	Cluster Cohesion
K-Means	0.077241	6.706627	1.35E+01
DBSCAN	0.172461	3.553307	1.17E-15
Fuzzy Clustering	0.188819	12.310082	1.79E+01
Hierarchical Clustering	0.113975	5.428147	1.26E+01

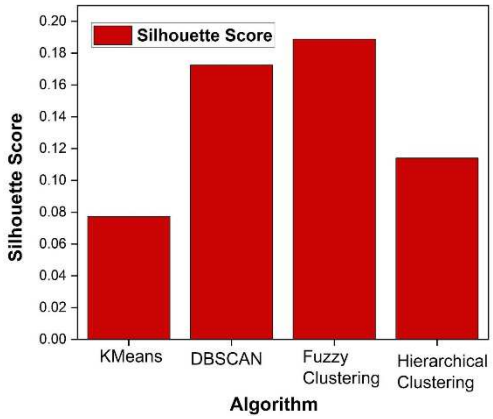

Fig. 6. Graphical representation of silhouette score of algorithms

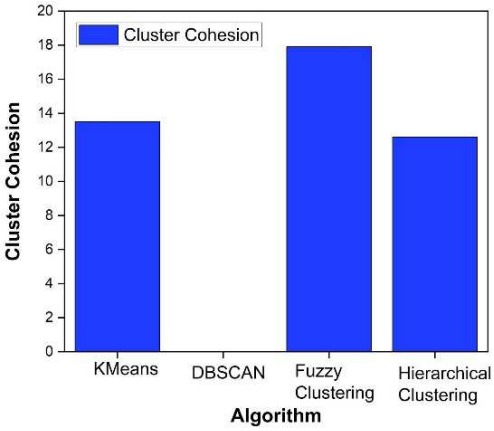

Fig. 7. Graphical representation of cluster cohesion of algorithms

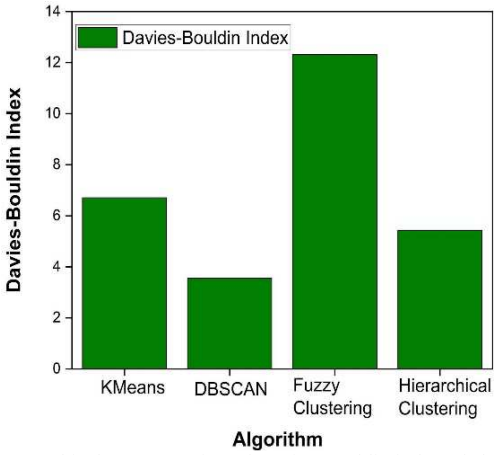

Fig. 8. Graphical representation of Davies Bouldin index of algorithms

A. *Silhouette Score*

Silhouette method computes the position of each object about its segment by assessing the mean separation among the entities within the same segment and among segments [19]. It measures the quality of clustering by comparing intra-cluster cohesion and inter-cluster separation. The measure of every point would be the difference of these values, between the separation and compactness, and is then divided by the maximum of both these values [20].

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i)} \quad (1)$$

The silhouette score of the ith data point is represented by $s(i)$,the mean separation among the$i$ th data point and other points of the same segment are represented by $a(i)$ and $b(i)$is the mean separation between the ith sample and samples in the nearest segment.

Pertaining to the silhouette score obtained Fuzzy Clustering has the best-defined clusters, followed by DBSCAN and hierarchical while K-Means has the weakest score indicating that clusters formed by K-Means have the maximum overlap due to its assumption of spherical clusters whereas the fuzzy algorithm allows data points to belong to multiple clusters.

B. Davies-Bouldin Index

DBI evaluates the effectiveness of clustering assessment by measuring the average resemblance ratio between each cluster with its closest counterpart.

$$DBI = \frac{1}{N}\sum_{i=1}^{n} max_{j\neq i}\frac{d_i+d_j}{d_{ij}} \quad (2)$$

In which Ndenotes the overall count of clusters,d_i is the mean distance amidst the entire set of points within the ith cluster,$d_j$is the mean distance amidst all points within the jth cluster, $d_{ij}$is the separation within the centroidal points associated with the ith and jth cluster.

For the DBI score obtained DBSCAN shows the best cluster separation, followed by hierarchical and K-Means while Fuzzy Clustering has the least DBI value indicating that Fuzzy clustering facilitates the overlap of cluster boundaries whereas DBSCAN creates clusters with maximum separation along with noise handling.

C. Cluster Cohesion

Measures the compactness of clusters, typically by measuring the total distances between the points inside the cluster.

$$Cohesion\ (C_k) = \sum_{i=1}^{n_k} II(x_i - \mu_k)II^2 \quad (3)$$

$$Total\ Cohesion = \sum_{k=1}^{K}\sum_{i=1}^{n_k} II(x_i - \mu_k)II^2 \quad (4)$$

Where k represents a specific cluster,$n_k$indicates the count of data points in the cluster k, x_i is a data point within a cluster k,$x_i$is the centroidal point associated with the cluster k and$K$ represents the entire count of clusters.

For the value obtained for cohesion, we can see that DBSCAN achieves the tightest clusters, indicating higher cohesion, while Fuzzy Clustering has the least cohesive clusters henceforth resulting in these values. Since the total cohesion value of DBSCAN is infinitesimally small hence it coincides with the x-axis.

VII. Conclusion

By taking into consideration the performance of algorithms using the Davies-Bouldin Index, silhouette score, and Cluster Cohesion, DBSCAN comes out to be the most reliable for Customer segmentation as it returns the smallest value of the Davies-Bouldin Index, which is an excellent separation between clusters. DBSCAN showed the highest cohesion, and thus the clusters were very tight and well-defined. On the other hand, Fuzzy Clustering returned the highest Silhouette Score with well-defined clusters, which shows how similar data points are within the same cluster compared to those of different clusters. However, the results for Fuzzy Clustering were very bad concerning the separation of the clusters, as reflected by the prominent value of Davies-Bouldin Index, and cohesion. K-Means has the lowest Silhouette Score and thus has the worst clustering performance of all. Hierarchical clustering has performed moderately out of all but struggles with maintaining the cluster distinctiveness as well as its scalability. Of all algorithms on this dataset, DBSCAN is the most robust and reliable having strikingly balanced cohesion and separation metrics of the clusters.

These findings underline the critical importance of choosing an adequate clustering algorithm, taking into consideration the particularities of dataset at hand, as the efficiency of customer relationship management is governed by that and it is from that level where strategic business decisions originate.

Acknowledgement

We intend to appreciate the individuals and organization that contributed towards the dataset utilised in this research. The dataset, based on sales transactions data on amazon, provided a very insightful basis for the analysis of customer segmentation. This dataset played a pivotal role in allowing us to compare various clustering algorithms such as K-Means, DBSCAN, Fuzzy C Means and hierarchical clustering, using key evaluation metrics including the silhouette score, Davies-Bouldin index and cluster cohesion.

References

[1] A. Afzal et al., “Customer Segmentation Using Hierarchical Clustering,” in 2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/I2CT61223.2024.10543349.

[2] S. S. Mim and D. Logofatu, “A Cluster-based Analysis for Targeting Potential Customers in a Real-world Marketing System,” in Proceedings - 2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing Conference, ICCP 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 159–166. doi: 10.1109/ICCP56966.2022.10053985.

[3] V. Mehta, R. Mehra, and S. S. Verma, “A Survey on Customer Segmentation using Machine Learning Algorithms to Find Prospective Clients,” in 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICRITO51393.2021.9596118.

[4] 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 14th IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 550–555. doi: 10.1109/ISCAIE61308.2024.10576353.

[5] T. Pitchayaviwat, "A study on clustering customer suggestion on online social media about insurance services by using text mining techniques," *2016 Management and Innovation Technology International Conference (MITicon)*, Bang-San, Thailand, 2016, pp. MIT-148-MIT-151, doi: 10.1109/MITICON.2016.8025228.

[6] 2016 Management and Innovation Technology International Conference (MITicon). IEEE, 2016.

[7] S. R. Regmi, J. Meena, U. Kanojia, and V. Kant, “Customer Market Segmentation using Machine Learning Algorithm,” in 2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1348–1354. doi: 10.1109/ICOEI53556.2022.9777146.

[8] V. Jabade, S. Ghadge, M. Jamadar, and P. Girase, “Customer Segmentation for Smooth Shopping Experience,” in 2023 4th International Conference for Emerging Technology, INCET 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/INCET57972.2023.10170126.

[9] N. U. Maisurah Mohd Faiz et al., “Comparative Analysis for Customer Profiling and Segmentation in Food and Beverages Using Data Mining Techniques,” in 2022 IEEE 10th Conference on Systems, Process and Control, ICSPC 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 286–291. doi: 10.1109/ICSPC55597.2022.10001767.

[10] D. Teslenko, A. Sorokina, K. Smelyakov, and O. Filipov, "Comparative Analysis of the Applicability of Five Clustering Algorithms for Market Segmentation," in 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences, eStream 2023 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/eStream59056.2023.10134796.

[11] N. Tressa, V. Asha, P. Kumar, O. Shree, M. Uday Kiran, and V. V. S. Reddy, "Customer-Based Market Segmentation Using Clustering in Data Mining," in 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things, IDCIoT 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 687–691. doi: 10.1109/IDCIoT59759.2024.10467258.

[12] ICCCBDA 2019: 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics: April 12-15, 2019, Chengdu, China. IEEE, 2019.

[13] Z. Sadreddini, I. Donmez, and H. Yanikomeroglu, "Cancel-for-Any-Reason Insurance Recommendation Using Customer Transaction-Based Clustering," IEEE Access, vol. 9, pp. 39363–39374, 2021, doi: 10.1109/ACCESS.2021.3064929.

[14] A. Razia Sulthana, A. Jaiswal, P. Supraja, and L. Sairamesh, "Customer Segmentation using Machine Learning," in 2023 3rd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICAECT57570.2023.10117924.

[15] EICT 2017: 3rd International Conference on Electrical Information and Communication Technology: 7-9 December 2017. IEEE, 2017.

[16] N. Godcares, A. Sirsath, A. Bongale, P. Kadam, R. Jayawal, and S. Patil, "Exploring Customer Segmentation in the Context of Market Analysis," in IEEE Region 10 Humanitarian Technology Conference, R10-HTC, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 444–449. doi: 10.1109/R10-HTC57504.2023.10461815.

[17] 2018 IEEE Latin American Conference on Computational Intelligence: 7-9 November 2018, Guadalajara, Mexico. Institute of Electrical and Electronics Engineers, 2018.

[18] N. S. Ayyildiz, A. Akcay, B. Yalcuva, A. Sayar, S. Ertugrul, and T. Cakar, "Segmentation for Factoring Customers: Using Unsupervised Machine Learning Algorithms," in 2023 Innovations in Intelligent Systems and Applications Conference, ASYU 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ASYU58738.2023.10296639.

[19] "2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)." IEEE, 20230214.

[20] "Customer Segmentation Using Fuzzy C-Means Algorithm in Telco Industry".