Comparative Analysis of Distance Metrics in DBSCAN for Customer Segmentation

Pragna Sree Koripuri[1], Preetika Kadiri[1], Sreeja Geddha[1] , Dr.S.Rasheed Uddin[1]

[1]*Department of CSE-AIML,CMR Engineering College*

---***---

Abstract - There are various strategies accessible for client division to tailor items to person inclinations. Be that as it may, numerous existing approaches battle to handle circumstances where the information is unpredictable, and conventional calculations may fall flat to distinguish such unpredictable fragments. In these cases, it gets to be challenging to bunch clients in a way that reflects their interesting item inclinations. To address this issue, we propose the utilize of an elective calculation "DBSCAN" that can viably oversee sporadic client portions, guaranteeing that indeed one of a kind or exception clients are legitimately distinguished and catered to with customized item offerings.

Key words: Client Division, Subjective Shape of Clusters, Exception Clients, Unpredictable information.

1.INTRODUCTION

Client segmentation is the division of a population into specific groups having a unifying particularity, or preference in order to allow a business to understand their guests more and offer applicable products, services, and marketing [1]. The ways applied in the segmentation of a population include demographic, behavioral, and psychographic types of segmentation and have the eventuality to deliver particular guests , accordingly leading to elevated satisfaction, fidelity, and, in general, a better business result [2].

In certain cases, noisy and nebulous data can hamper the effectiveness of traditional segmentation styles. utmost the models are preset on clear datasets; they fail to pick trends in complex information and produce deceiving or vague parts, and thus, end up customizing products less effectively [3].

This would mean client data, for case, becomes incomprehensible and therefore hard to apply meaningful segmentation. Standard ways may miss some veritably unique client actions about them, leaving businesses unfit to deliver individualized products and services[4]. therefore, making it hard for the guests who don't fit into these predefined molds to get their requirements addressed .

2. RELATED WORKS

Ester et al. (1996) introduced the DBSCAN algorithm, which is pivotal for density-based clustering, especially for identifying clusters of arbitrary shapes in large datasets [10]. This foundational work was critical to this research, as DBSCAN's ability to handle clusters of varying shapes made it suitable for the project's dataset, where clusters did not conform to specific geometric shapes .

Carnein and Trautmann demonstrated the use of stream clustering to dynamically group customers, which was relevant to this research in terms of adapting clustering techniques for client data [5]. Although this research focused on static datasets rather than real-time clustering, the approach inspired the application of DBSCAN to capture variations in customer segments more effectively, even in a non-streaming context .

Yu et al. developed a three-way clustering modification to DBSCAN, which was insightful for understanding adaptations of DBSCAN to particular data structures [6]. Their work on tailoring DBSCAN's functionality guided my consideration of alternative distance metrics, such as the Euclidean, Manhattan, and Minkowski measures, to enhance the evaluation of cluster quality and fit DBSCAN to the unique shape-based characteristics of my data .

Ozan (2018) conducted a case study on customer segmentation using machine learning methods, highlighting the practical applications of various clustering techniques in understanding consumer behavior. This study emphasizes the importance of data-driven approaches in segmentation and complements the findings of this research by showcasing how machine learning can enhance segmentation strategies. Ozan's work also aligns with the current project’s objectives of leveraging advanced clustering algorithms to derive insights from complex customer datasets [7].

3. PROPOSED APPROACH

The proposed approach is aimed at handling problems encountered in the traditional customer segmentation by application of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifying irregular patterns of behavior from customers. Two approaches follow based on features such as age, annual income, and spending score: one considering all customers together, and another by dividing the data into male and female subsets. Each of these techniques utilizes Euclidean, Manhattan, and Minkowski distance functions to compare the differences among customer behaviors. As there are diverse kinds of behavioral patterns, employing different values of MinPts and epsilon increases robustness for cluster formation using these distance functions [8].

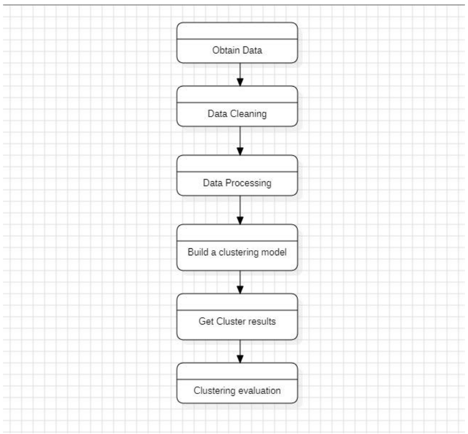


Fig 1:Data Flow

For the first one of these techniques, not accounting for gender, separate applications of DBSCAN can form clusters based on age, income, and spending score . This approach aims to find common and outlying groups in the overall population. In the second approach, the dataset is stratified by gender, and therefore, targeted segmentation can be achieved. Different clusters are created for male and female customers with scaled data yield different cluster patterns. This dual approach works towards a comprehensive coverage of customer behaviors and preferences that cluster methods in their more traditional settings were likely to miss [9].

It's also different by using distinct clustering parameters and distance metrics, so more can be learned about customer groups with unique characteristics. Clusters generated are scored based on a specific scoring system that weighs age, income, and spending score; thus, the results of segmentation are not only statistically significant but also practically useful. It improves a business's ability to cater to unique customer preferences; hence, businesses can offer more personalized products and services to customers.

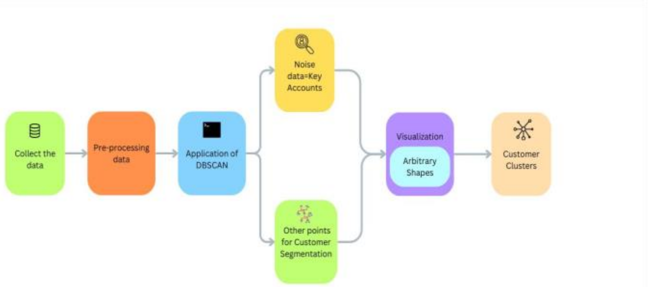
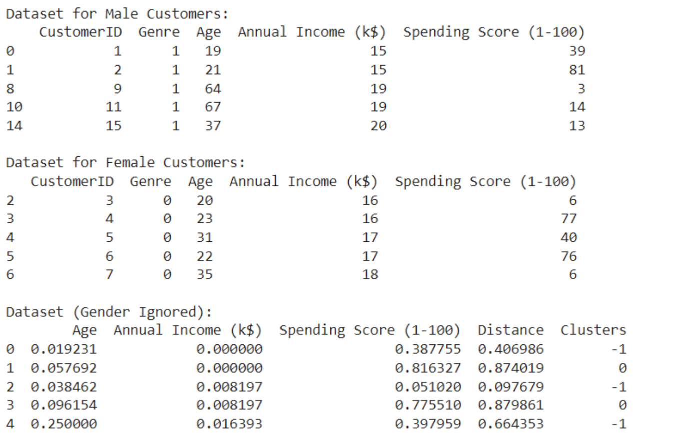


Fig 2: System Architecture

4.RESULT ANALYSIS

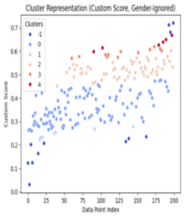
In this section, applying the DBSCAN clustering calculation to the client information set brings forward major discoveries into client division beneath two essential approaches: one that disregards gender and another that isolates information into male and female subsets.

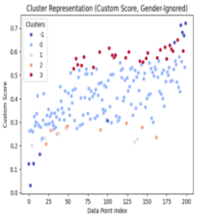


```
Dataset for Male Customers:
    CustomerID  Genre  Age  Annual Income (k$)  Spending Score (1-100)
0            1      1   19                  15                      39
1            2      1   21                  15                      81
8            9      1   64                  19                       3
10          11      1   67                  19                      14
14          15      1   37                  20                      13

Dataset for Female Customers:
   CustomerID  Genre  Age  Annual Income (k$)  Spending Score (1-100)
2           3      0   20                  16                       6
3           4      0   23                  16                      77
4           5      0   31                  17                      40
5           6      0   22                  17                      76
6           7      0   35                  18                       6

Dataset (Gender Ignored):
        Age  Annual Income (k$)  Spending Score (1-100)  Distance  Clusters
0  0.019231            0.000000                0.387755  0.406986        -1
1  0.057692            0.000000                0.816327  0.874019         0
2  0.038462            0.008197                0.051020  0.097679        -1
3  0.096154            0.008197                0.775510  0.879861         0
4  0.250000            0.016393                0.397959  0.664353        -1
```

Fig 3: Dataset

In the first approach, DBSCAN was applied using three distance measures: Manhattan, Euclidean, and Minkowski. The choice of distance metric significantly influenced cluster shapes and sizes. Manhattan distance produced elongated clusters, while Euclidean distance resulted in compact, circular clusters. Minkowski distance yielded irregularly shaped clusters, reflecting the complex relationships between 'Age', 'Annual Income', and 'Spending Score'. A customized score, calculated as a weighted sum of these features, allowed ranking customers by their overall value based on income and spending behavior.





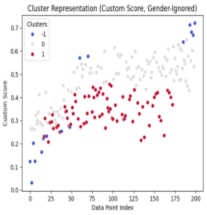


Fig 4: Manhattan **Fig 5**:Euclidean **Fig 6**:Minkowski

In the second approach, the data were divided into male and female subsets, and DBSCAN was applied separately to each. The results showed significant differences in clusters for each gender, indicating gender-dependent spending and income behaviors. For males, clusters were less dense with Manhattan distance and more compact with Euclidean distance, while Minkowski distance produced complex shapes similar to the gender-ignored dataset. Female customers exhibited similar trends but with differently shaped clusters, suggesting distinct spending and income patterns compared to males.

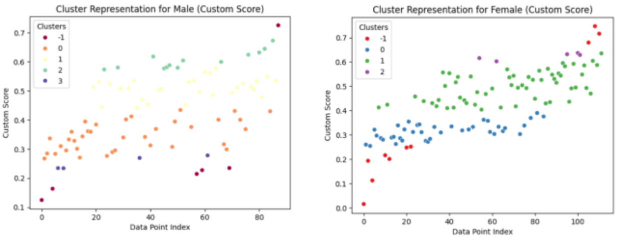


Fig 7: Manhattan

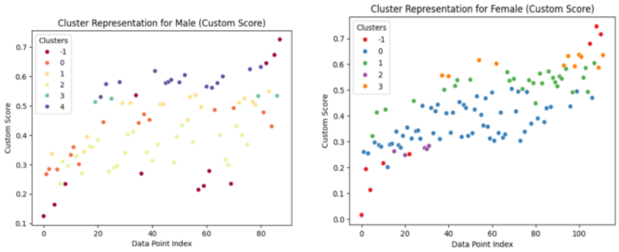


Fig 8: Euclidean

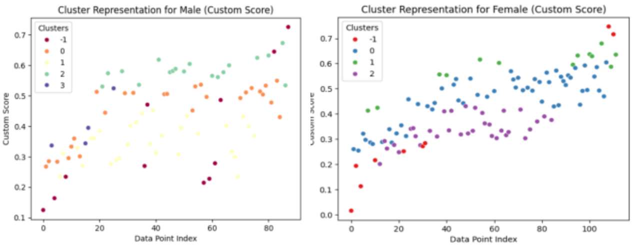
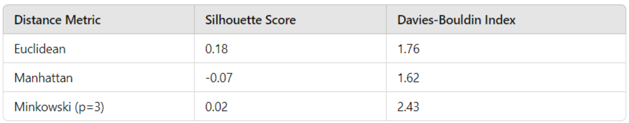


Fig 9:Minkowski

In both approaches, a specific score integrating 'Age', 'Annual Income', and 'Spending Score' highlighted the distribution of high-value customers across clusters. For the gender-ignored dataset, clusters with higher custom scores indicated segments with greater spending capacities. When dividing customers by gender, the technique revealed distinct average contributions to custom scores for male and female clusters, providing a clearer view of spending patterns and capacities within each gender group.



Distance Metric	Silhouette Score	Davies-Bouldin Index
Euclidean	0.18	1.76
Manhattan	-0.07	1.62
Minkowski (p=3)	0.02	2.43

Table 1:Distance Metric Comparison

The comparison of attained results easily defines the distinctions between distance criteria that were compared, as well as the distinctions in their separate analysis. Euclidean distance achieved the stylish score with the loftiest figure Score(0.18) and moderate Davies-Bouldin Index- 1.76, showing well- set and compact clusters. In discrepancy, Manhattan distance showed a negative figure Score(- 0.07), and well- defined clusters couldn't be determined, though the Davies- Bouldin Index value was fairly lower(1.62). Minkowski distance with p = 3 was set up with the worst clustering because its figure Score is too low at 0.02 and the Davies- Bouldin Index is veritably high at 2.43, which implies that the clusters are irregular. Hence, it's important to elect the correct distance metric so as to get the optimal quality of the clusters.

5.CONCLUSION

The DBSCAN clustering algorithm was successfully applied in this client segmentation design, identifying distinct client groups based on purchasing behaviors, income levels, and demographics. Unlike traditional methods like k-means, DBSCAN's strength lies in detecting clusters of arbitrary shapes and handling outliers, offering deeper insights into client segments. These results enable businesses to design targeted marketing strategies, improve client satisfaction, and make data-driven decisions. Future advancements could include using more complex datasets or enriching the feature set for even finer segmentation.

REFERENCES

1. John, A., Shobayo, A., & Ogunleye, T. (2023). Investigating Clustering Calculations for Client Division in the UK Retail Advertise. Diary of Retail Analytics, 12(4), 57-68.
2. Somya, R., Winarko, E., & Privanta, S. (2021). A Novel Approach to Collect and Analyze Market Customer Behavior Data on Online Shops.
3. Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Calculation. Universal Diary of Data Administration, 53, 102106.
4. Kaur, B., & Sharma, P. K. (2019). Implementation of Customer Segmentation Using an Integrated Approach.
5. Carnein, M., & Trautmann, H. (2019). Client Division Based on Value-based Information Utilising Stream Clustering. Data Frameworks, 83, 1-14.
6. Yu, H., Chen, L., & Wang, X. (2019). A Three-Way Clustering Strategy Based on a Made strides DBSCAN Algorithm.
7. Ozan, S. (2018). A Case Study on Customer Segmentation by Using Machine Learning Methods. IEEE.
8. Glory H. Shah (2013) An improved DBSCAN, a density based clustering algorithm with parameter selection for high dimensional data sets.
9. Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data.
10. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). Density-Based Calculations for Mining in Huge Databases.