



Comparative Analysis Of Clustering Algorithms For Customer Segmentation In Beauty And Skincare Retail

Annavinayagi V¹, Mrs. P. Jasmine Lois Ebenezer MCA., M.Phil.²

¹ Reg.No:25081205300112004, Department of Computer Application and Research Centre, Sarah Tucker College(Autonomous), Tirunelveli.

² Associate Professor & Head, Department of Computer Application and Research Centre, Sarah Tucker College(Autonomous), Tirunelveli.

ABSTRACT :

Customer segmentation helps businesses identify groups of customers with similar characteristics and purchasing behaviour. This study presents a comparative analysis of three unsupervised machine learning algorithms, namely K-Means, Agglomerative Clustering, and DBSCAN, for customer segmentation in the beauty and skincare retail domain. A dataset containing 1,000 customer records was used for analysis. The dataset was preprocessed through missing-value handling, duplicate removal, categorical encoding, feature selection, and feature scaling. K-Means, Agglomerative Clustering, and DBSCAN were applied to the standardized dataset and their performance was evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. K-Means achieved the highest Silhouette Score of 0.3548, the lowest Davies-Bouldin Index of 1.1272, and the highest Calinski-Harabasz Index of 503.23. Based on these results, K-Means provided the strongest clustering performance among the tested algorithms. The K-Means model identified three customer segments with different income, spending behaviour, age, shopping frequency, and purchase characteristics. The resulting segmentation can support targeted customer strategies in the beauty and skincare retail sector.

Keywords: Customer Segmentation, K-Means, Agglomerative Clustering, DBSCAN, Machine Learning, Beauty and Skincare Retail, Clustering

1. INTRODUCTION

The increasing availability of customer data has created opportunities for businesses to understand customer characteristics and purchasing behaviour through data-driven techniques. Customer segmentation is an important approach that divides customers into groups based on similarities in their demographic and purchasing characteristics.

The beauty and skincare retail industry contains customers with different income levels, spending behaviour, shopping frequency, product preferences, and purchasing patterns. Treating all customers as a single group may not effectively address these differences. Identifying customer segments can help businesses understand their customer base and develop suitable marketing and customer engagement strategies.

Clustering is an unsupervised machine learning technique used to group similar data points without requiring predefined class labels. Different clustering algorithms use different approaches to identify groups within a dataset. Therefore, comparing multiple algorithms can provide a better understanding of their suitability for a particular customer dataset.

The proposed study compares three clustering algorithms: K-Means, Agglomerative Clustering, and DBSCAN. The algorithms are applied to a beauty and skincare customer dataset containing 1,000 records. The clustering results are evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Based on the evaluation results, the most suitable clustering approach is identified and the resulting customer segments are analyzed.

2. PROPOSED METHODOLOGY

The proposed customer segmentation framework consists of dataset collection, data preprocessing, missing value handling, duplicate removal, categorical encoding, feature selection, and feature scaling. The preprocessed data is then used to apply K-Means, Agglomerative Clustering, and DBSCAN algorithms. The resulting clusters are visualized and evaluated using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Finally, customer segment profiling is performed to identify distinct customer groups and understand their characteristics and purchasing behaviour.

3. DATASET DESCRIPTION

The dataset used in this study contains 1,000 beauty and skincare customer records with nine attributes: Customer ID, age, gender, annual income, spending score, shopping frequency, favorite product, preferred shopping mode, and average purchase amount. The Customer ID attribute is used only for customer identification and is excluded from the clustering process. The remaining eight attributes—age, gender, annual income, spending score, shopping frequency, favourite product, preferred shopping mode, and average purchase amount—are selected as clustering features to analyze customer characteristics and purchasing behaviour.

3.2 DATA PREPROCESSING

Data preprocessing was performed to prepare the customer dataset for clustering. Missing values in numerical attributes, including annual income, spending score, and average purchase amount, were replaced with their respective mean values, while missing values in categorical attributes, including gender, shopping frequency, favourite product, and preferred shopping mode, were replaced with their respective mode values. Duplicate records were identified and removed to prevent repeated observations from affecting the clustering results. Since clustering algorithms require numerical input, the categorical attributes were converted into numerical representations using Label Encoding. Finally, the selected features were standardized using Standard Scaler because they have different numerical ranges. This scaling process helps prevent features with larger values from disproportionately influencing the clustering results.

4. K-MEANS CLUSTERING

K-Means is an unsupervised clustering algorithm that groups data points into a predefined number of clusters based on their similarity to cluster centroids. In this study, the Elbow Method was used to determine the optimal number of clusters by calculating the Within-Cluster Sum of Squares (WCSS) for different values of K. The resulting Elbow curve indicated that K = 3 was suitable for the dataset. Therefore, the K-Means algorithm was implemented with three clusters, and the resulting cluster labels were assigned to the customer dataset. The clusters were visualized using annual income and spending score to illustrate the distribution of customers across the identified segments.

5. AGGLOMERATIVE CLUSTERING

Agglomerative Clustering is a hierarchical clustering algorithm that initially treats each observation as an individual cluster and progressively merges similar clusters until the desired number of clusters is obtained. In this study, the algorithm was applied with three clusters to enable a consistent comparison with K-Means. The resulting cluster assignments were visualized using annual income and spending score to examine the distribution of customers across the identified segments.

6. DBSCAN CLUSTERING

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that identifies groups of closely located data points and classifies isolated observations as noise. In this study, DBSCAN was implemented with an epsilon (eps) value of 1.8 and a minimum sample value of 5. The resulting cluster assignments were visualized using annual income and spending score to examine the distribution of customers across the identified clusters and noise points.

7. CLUSTER VISUALIZATION

The clustering results were visualized using scatter plots, with annual income on the X-axis and spending score on the Y-axis. These variables provide an interpretable two-dimensional representation of the identified customer groups. However, the clustering process was performed using all eight selected features. Separate scatter plots were generated for K-Means, Agglomerative Clustering, and DBSCAN to compare the distribution of customers across the resulting clusters.

8. Clustering Evaluation

Three internal clustering evaluation metrics were used to compare the performance of the clustering algorithms. The Silhouette Score measures the similarity of each observation to its own cluster compared with other clusters, where a higher value indicates better-defined and well-separated clusters. The Davies–Bouldin Index evaluates cluster compactness and separation, with a lower value indicating better clustering performance. The Calinski–Harabasz Index measures the ratio of between-cluster dispersion to within-cluster dispersion, where a higher value indicates better-defined clusters. These metrics were used to compare K-Means, Agglomerative Clustering, and DBSCAN and identify the best-performing algorithm among the tested configurations. The evaluation was performed using the selected clustering features after feature scaling. For DBSCAN, noise observations were excluded from the metric calculations. The combined results provide a quantitative basis for assessing cluster quality and comparing the effectiveness of the three algorithms.

9. EXPERIMENTAL RESULTS

The three clustering algorithms were applied to the standardized customer dataset.

The obtained evaluation results are shown below.

Algorithm	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Means	0.3548	1.1272	503.23
Agglomerative	0.3347	1.1666	459.93
DBSCAN	0.1491	1.6188	154.76

K-Means obtained the highest Silhouette Score of 0.3548.

It also obtained the lowest Davies-Bouldin Index of 1.1272.

The Calinski-Harabasz Index was also highest for K-Means at 503.23.

Therefore, K-Means demonstrated the strongest clustering performance among the three algorithms under the tested configurations.

10. CUSTOMER SEGMENTATION RESULTS

Based on the evaluation results, K-Means was selected for detailed customer segment analysis.

The K-Means algorithm generated three customer segments.

Cluster	Number of Customers	Percentage
Cluster 0	424	42.4%
Cluster 1	250	25.0%
Cluster 2	326	32.6%
Total	1000	100%

10.1 Cluster 0

Cluster 0 consists of 424 customers, representing 42.4% of the dataset. The average annual income of this group is approximately 15.56 LPA, with an average spending score of 72.91 and an average purchase amount of ₹3,308. These characteristics indicate that customers in this segment have relatively high incomes and spending levels. Therefore, this group can be considered a high-value customer segment and targeted with premium skincare products, exclusive offers, and personalized marketing strategies.

10.2 Cluster 1

Cluster 1 consists of 250 customers, representing 25% of the dataset. The average annual income of this group is approximately 4.59 LPA, with an average spending score of 33.09 and an average purchase amount of ₹665. These characteristics indicate relatively low income and spending levels. Therefore, this segment can be considered a budget-conscious customer group and targeted with affordable skincare products, discounts, and value-based promotional offers.

10.3 Cluster 2

Cluster 2 consists of 326 customers, representing 32.6% of the dataset. The average age of this group is approximately 23.79 years, with an average spending score of 83.84 and an encoded shopping frequency value of 2.00. This cluster has the highest average spending score and encoded shopping frequency value among the three groups. These characteristics suggest a young customer segment with high spending engagement. Therefore, this group can be targeted with trending skincare products, loyalty rewards, personalized recommendations, and digital marketing campaigns.

11. BUSINESS IMPLICATIONS

The identified customer segments can support differentiated marketing strategies in the beauty and skincare retail sector. The high-income and high-value segment can be targeted with premium products, personalized recommendations, and loyalty benefits. The low-income and low-spending segment can be approached with affordable products, discounts, and promotional offers. The young, high-engagement segment can be targeted through trending skincare products, reward programs, and personalized recommendations. Thus, customer segmentation provides a foundation for developing targeted marketing strategies based on customer characteristics and purchasing behaviour.

12. DISCUSSION

The experimental results demonstrate that the performance of clustering algorithms varies when applied to the same customer dataset. K-Means achieved the highest Silhouette Score and Calinski-Harabasz Index, along with the lowest Davies-Bouldin Index, indicating the strongest

clustering performance among the three tested algorithms. Agglomerative Clustering produced comparatively similar results, whereas DBSCAN achieved lower evaluation performance under the selected parameter configuration. The K-Means model identified three customer segments with distinct demographic and purchasing characteristics, including differences in income, spending score, age, shopping frequency, and average purchase amount. These findings highlight the usefulness of clustering techniques for identifying customer groups and supporting data-driven marketing strategies in the beauty and skincare retail sector.

13. LIMITATIONS

The study has the following limitations:

- The analysis uses a single beauty and skincare customer dataset.
- The clustering results depend on the selected features and preprocessing techniques.
- Different algorithm parameters may produce different clustering results.
- The clustering algorithms were evaluated using internal validation metrics.
- The identified customer segments require validation using real-world business data.

14. FUTURE ENHANCEMENT

Future work can improve the proposed segmentation framework by evaluating additional clustering algorithms and testing different parameter configurations. Dimensionality-reduction techniques can be explored to enhance the visualization of high-dimensional customer data. The framework can also be evaluated using larger, real-world beauty and skincare datasets to assess its generalizability. Furthermore, an interactive customer segmentation dashboard can be developed to help businesses explore customer groups, analyze their characteristics, and make informed marketing decisions.

15. CONCLUSION

This study presented a comparative analysis of K-Means, Agglomerative Clustering, and DBSCAN for customer segmentation in the beauty and skincare retail domain. A dataset containing 1,000 customer records was preprocessed through missing-value handling, duplicate removal, categorical encoding, feature selection, and feature scaling before applying the three clustering algorithms. Their performance was evaluated using the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. K-Means achieved a Silhouette Score of 0.3548, a Davies–Bouldin Index of 1.1272, and a Calinski–Harabasz Index of 503.23, demonstrating the strongest performance among the tested algorithms. The K-Means model identified three customer segments containing 424, 250, and 326 customers, respectively. These segments exhibited different income levels, spending behaviour, age groups, and average purchase amounts. The findings demonstrate the potential of clustering techniques to support customer understanding, targeted marketing, and data-driven decision-making in the beauty and skincare retail sector.

11. REFERENCE

1. A. Sri Nugroho, A. Saputro, F. A. Suhardi, and J. Heikal, "FMCG Industry Customer Segmentation for Cosmetic Product at PT. Paragon Technology and Innovation," *Syntax Idea*, vol. 6, no. 2, pp. 900–914, 2024.
2. A. Afzal, L. Khan, M. Z. Hussain, M. Z. Hasan, M. Mustafa, A. Khalid, R. Awan, F. Ashraf, Z. A. Khan, and A. Javaid, "Customer Segmentation Using Hierarchical Clustering," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 2024.
3. S. A. Laga, D. Hermansyah, C. L. Rithmaya, M. Zainuddin, G. A. I. Purnama Aji, and I. R. Mukhlis, "Analyzing Customer Loyalty Levels through Segmentation in Aesthetic Clinics Using K-Means and RFAM," *International Journal on Informatics for Development*, vol. 13, no. 2, pp. 473–484, 2024.
4. J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.
5. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231, 1996.
6. A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
7. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.
8. P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
9. D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
10. F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
11. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008.