International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Machine Learning-Based Customer Segmentation for Retail Insights

Chintha Hari haran[a], Akkanapalli Hemanth[b], Nenavath Saanki[c], Mallipeddi Sahith[d] , M.Deenababu [e,*]

[a,b,c,d] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100
[e] Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

Abstract:

The dynamic retail environment demands data-driven approaches to decipher customer preferences, enhance marketing precision, and boost profitability. This project leverages advanced machine learning techniques to segment mall customers by analyzing demographic and transactional data. Through thorough data cleaning, visualization, and feature analysis, critical patterns in spending behavior, income distribution, and customer demographics were uncovered. Four clustering algorithms — K-Means++, Gaussian Mixture Model (GMM), Hierarchical Clustering, and DBSCAN — were applied to categorize customers, with performance assessed using Silhouette Score and Davies-Bouldin Index metrics. K-Means++ emerged as the top performer, yielding the highest Silhouette Score and lowest Davies-Bouldin Index, indicating superior cluster quality. Visualizations such as scatter plots, dendrograms, and correlation matrices enriched the analysis by revealing behavioral insights and inter-attribute relationships. This system surpasses conventional segmentation by providing automated, scalable, and accurate customer grouping, enabling retailers to craft targeted promotions, optimize resource allocation, and elevate customer satisfaction. Future improvements could include real-time data integration, deep learning methods like autoencoders, and an interactive platform for actionable insights.

I. Introduction

The retail sector is undergoing rapid transformation due to increasing competition, evolving consumer preferences, and the proliferation of data-driven decision-making. Malls and retailers must adapt to these changes by understanding customer behavior, optimizing marketing strategies, and enhancing operational efficiency. The demand for personalized customer experiences has surged, driven by factors such as e-commerce growth, shifting demographics, and economic fluctuations. Traditional customer segmentation methods, often manual and static, struggle to keep pace with these dynamic trends, limiting their effectiveness in modern retail planning. Economic shifts, such as inflation and changing spending patterns, further complicate customer dynamics, while the rise of digital shopping has reshaped purchasing habits. Retailers are now leveraging big data and machine learning to gain actionable insights into customer preferences and behaviors.

However, conventional segmentation tools lack the ability to uncover hidden patterns or adapt to real-time changes, hindering long-term business strategies. The need for automated, scalable, and precise customer segmentation is more critical than ever. By employing machine learning algorithms, retailers can accurately group customers, enabling tailored marketing and improved customer retention. This study proposes a machine learning-driven approach to mall customer segmentation. Utilizing a comprehensive dataset of customer demographics and transactional records, the system identifies spending trends, income distributions, and behavioral patterns. Clustering algorithms such as K-Means++, Gaussian Mixture Model (GMM), Hierarchical Clustering, and DBSCAN are implemented to segment customers with high precision. Among these, K-Means++ achieves the best performance, as evaluated by Silhouette Score and Davies-Bouldin Index metrics. The findings benefit multiple stakeholders: retailers can optimize marketing campaigns and store layouts, customers receive personalized experiences, and mall management can enhance resource allocation. Additionally, this research lays the groundwork for future real-time

segmentation and advanced analytics in retail.

II.Literature Survey

The The integration of machine learning into customer segmentation has transformed retail analytics and business intelligence. Researchers have explored clustering techniques and data-driven approaches to improve customer classification and marketing strategies. Aman Banduni and Prof. Ilavedhan A. (2022) emphasized the role of unsupervised learning in customer segmentation, demonstrating how K-Means and GMM automate the process and enhance marketing personalization [1]. Kamalpreet Bindra and Anuranjan Mishra (2021) conducted a comparative study of clustering algorithms, highlighting the strengths and limitations of K-Means, DBSCAN, and Hierarchical Clustering in handling real-world customer data [2]. Kai Peng et al. (2020) explored Mini-Batch K-Means for large-scale segmentation, showcasing its scalability for retail applications [3]. Fionn Murtagh and Pedro Contreras (2018) studied Hierarchical Clustering, noting its interpretability through dendrograms for understanding customer relationships [4]. D. P. Yash Kushwaha and Deepak Prajapati (2019) analyzed K-Means efficiency, identifying its effectiveness in spending-based segmentation despite sensitivity to initialization [5]. Manju Kaushik and Bhawana Mathur (2017) compared K-Means and Hierarchical Clustering, emphasizing trade-offs between computational efficiency and detailed customer insights [6]. Han et al. (2011) introduced foundational clustering concepts in data mining, widely applied in retail analytics [7]. Kotler and Keller (2016) discussed marketing management principles, underscoring the importance of segmentation for customer engagement [8]. McKinsey & Company (2021) highlighted the role of analytics in retail transformation, predicting a shift toward AI-driven strategies [9]. IBM (2020) showcased enterprise applications of machine learning in customer behavior analysis, improving business outcomes [10].

Table .1. Literature Survey

Study	Key Contribution	Year
Banduni & Ilavedhan	Demonstrated K-Means and GMM for automated customer segmentation.	2022
Bindra & Mishra	Compared K-Means, DBSCAN, and Hierarchical Clustering for segmentation.	2021
Peng et al.	Introduced Mini-Batch K-Means for scalable customer segmentation.	2020
Murtagh & Contreras	Explored Hierarchical Clustering with dendrograms for customer relationships.	2018
Kushwaha & Prajapati	Analyzed K-Means efficiency in spending-based segmentation.	2019
Kaushik & Mathur	Compared K-Means and Hierarchical Clustering for retail applications.	2017
Han et al.	Provided foundational clustering techniques for data mining.	2011
Kotler & Keller	Emphasized segmentation's role in marketing management.	2016
Glassdoor Economic Research	Analyzed retail market trends and the demand for customer segmentation techniques and skills.	2018
McKinsey & Company	Predicted retail shifts with AI-driven analytics.	2021
Smith & Johnson (2020)	Investigated the effectiveness of deep learning models, such as autoencoders, in enhancing customer segmentation accuracy.	2020
Deloitte Insights (2022)	Explored the impact of real-time data analytics and dynamic segmentation on customer behavior prediction and personalized marketing strategies.	2022
IBM	Showcased machine learning in customer behaviour analysis.	2020

III. Methodology

The methodology for this research centers on applying machine learning clustering techniques and statistical analysis to segment mall customers effectively. The process involves data collection, preprocessing, feature engineering, model selection, and evaluation.

3.1. Data Collection

This study utilizes a Kaggle dataset comprising 200 customer records with 5 attributes: Customer ID, Gender, Age, Annual Income, and Spending Score. This dataset provides real-world insights into mall customer behavior across demographics and purchasing patterns. To ensure robustness, supplementary validation was conducted using retail industry reports from sources like McKinsey and IBM. The combination of structured attributes and behavioral data enables a comprehensive analysis of customer segmentation trends.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Customer	Gender	Age	Annual In	Spending Score (1-100)							
2	1	Male	19	15	39							
3	2	Male	21	15	81							
4	3	Female	20	16	6							
5	4	Female	23	16	77							
6	5	Female	31	17	40							
7	6	Female	22	17	76							
8	7	Female	35	18	6							
9	8	Female	23	18	94							
10	9	Male	64	19	3							
11	10	Female	30	19	72							
12	11	Male	67	19	14							
13	12	Female	35	19	99							
14	13	Female	58	20	15							
15	14	Female	24	20	77							
16	15	Male	37	20	13							
17	16	Male	22	20	79							
18	17	Female	35	21	35							
19	18	Male	20	21	66							

Fig 1. Sample Dataset

3.2. Data Preprocessing

The dataset underwent preprocessing to enhance quality and model performance. Missing values, duplicates, and inconsistencies were addressed through data cleaning. Numerical features (e.g., Annual Income, Spending Score) were normalized using Min-Max Scaling, while categorical variables (e.g., Gender) were encoded. Outlier detection removed anomalous entries, ensuring a high-quality dataset for clustering.

3.3. Feature Engineering

Feature engineering enhanced clustering accuracy by emphasizing key customer attributes. Spending Score and Annual Income were standardized to capture purchasing power and behavior. Age was categorized into groups (e.g., young adults, middle-aged) to reflect demographic trends. Gender was analyzed to identify spending differences, enriching the segmentation process.

3.4. Model Training

Four clustering algorithms were trained to segment customers, The dataset was split into 80% training and 20% testing to assess model performance. The following models were implemented

3.4.1 K-Means++

K-Means++ is an improved version of K-Means that enhances centroid initialization, reducing clustering errors and improving convergence speed. Instead of random selection, it strategically chooses initial centroids by maximizing the minimum squared distance from existing centroids. This approach minimizes intra-cluster variance, leading to more stable and accurate clustering results

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

3.4.2 Gaussian Mixture Model (GMM)

A probabilistic clustering model, such as the Gaussian Mixture Model (GMM), assigns data points to clusters based on probability distributions rather than fixed assignments. Unlike hard clustering methods, where each data point belongs to only one cluster, GMM provides soft clustering, meaning each point has a probability of belonging to multiple clusters. The model assumes that the data is generated from a mixture of multiple Gaussian distributions, each with its own mean (μk\mu_kμk) and covariance matrix (Σk\Sigma_kΣk). The probability of a data point xxx belonging to a particular cluster is given by:

$$P(x) = \sum_{\{k=1\}}^{K} \pi_k N(x|\mu_k, \Sigma_k) \qquad (2)$$

GMM is particularly useful for capturing complex, elliptical-shaped clusters and is widely applied in customer segmentation, anomaly detection, and speech recognition due to its flexible nature.

3.4.3 DBSCAN

A density-based clustering method, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise), identifies clusters by grouping closely packed data points while marking sparse regions as outliers. The algorithm defines a neighborhood around each point using a specified radius (ε), and a point is classified as a core point if the number of neighboring points within this radius is at least MinPts (a user-defined threshold). Formally, a data point xx is considered a core point if:

$$|N\epsilon(x)| \geq \text{MinPts} \qquad (3)$$

where Nϵ(x)N_{\epsilon}(x) represents the set of points within the neighborhood radius ϵ\epsilon. If a core point is found, DBSCAN expands the cluster by recursively including reachable points, allowing it to discover arbitrarily shaped clusters and effectively handle noise in the data.

3.4.4 Hierarchical Clustering:

Hierarchical clustering constructs a dendrogram by progressively merging the most similar clusters at each step, using distance metrics like Euclidean or Manhattan distance to measure similarity. The algorithm begins with each data point as its own cluster, then iteratively combines pairs of clusters based on linkage criteria such as single, complete, or average linkage. This process continues until all points merge into a single cluster, forming a tree structure that reveals natural groupings at different similarity levels. The resulting dendrogram enables visual analysis of cluster relationships and helps determine optimal segmentation by identifying significant branch heights.

$$\textit{Core Point: } |N_{\varepsilon}(x)| \geq \textit{MinPts} \qquad (4)$$

where:

$N_{\varepsilon}(x)$ = neighbourhood of point x within radius ε

MinPts = minimum number of points required to form a dense region

3.5 Model Evaluation

Clustering performance was assessed using various evaluation metrics to determine the effectiveness and accuracy of customer segmentation. The two primary metrics used were the Silhouette Score and the Davies-Bouldin Index, which evaluate the quality of clustering by measuring cohesion within clusters and separation between clusters. These metrics are crucial in identifying the best-performing algorithm for customer segmentation, ensuring that the clusters are well-defined and meaningful.

Silhouette Score

The Silhouette Score measures the compactness of clusters by assessing intra-cluster cohesion and inter-cluster separation. A higher Silhouette Score indicates well-separated clusters, improving segmentation effectiveness.

The formula is given by:

$$S(i) = (b(i) - a(i)) / \max(a(i), b(i))$$

where:

a(i) = Average intra-cluster distance (how close a data point is to others in the same cluster).

b(i) = Average inter-cluster distance (how far a data point is from points in the nearest different cluster).

This metric provides a clear assessment of how well each data point fits within its assigned cluster and whether the clusters formed are distinct from one another. Higher Silhouette Scores (closer to 1) indicate better clustering quality, while negative values suggest misclassification of points.

Davies-Bouldin Index

The Davies-Bouldin Index is another key metric that evaluates the compactness of clusters relative to their separation. It is calculated as the average ratio of intra-cluster distances to inter-cluster distances across all clusters. A lower Davies-Bouldin Index indicates that clusters are more distinct and well-separated.

This index is particularly useful when comparing multiple clustering algorithms, as it helps identify which model produces the most compact and non-overlapping clusters. In this study, K-Means++ achieved the lowest Davies-Bouldin Index, confirming its ability to create well-separated customer groups.

The combination of these evaluation metrics ensured a rigorous assessment of the clustering models, allowing for an objective comparison of their performance. The results provided valuable insights into customer segmentation, supporting data-driven marketing and strategic business decisions.

Silhouette Score: Measures intra-cluster cohesion and inter-cluster separation:

where:

a(i) = average intra-cluster distance

b(i) = average inter-cluster distance

where a(i) a(i) a(i) is intra-cluster distance, and b(i) b(i) b(i) is inter-cluster distance.

Table 2: Evaluation Metrics

Metric	Formula
Silhouette Score	$S(i) = (b(i) - a(i)) / \max(a(i), b(i))$
Davies-Bouldin	Average ratio of intra-cluster to inter-cluster distances

3.6 Visualization of Insights

Visualizations using Python libraries (Matplotlib, Seaborn) included scatter plots (e.g., Income vs. Spending Score), revealing 5 distinct customer segments. Correlation matrices quantified attribute relationships, showing spending weakly tied to income (-0.12) but strongly to age (0.67). Dendrograms visualized hierarchical groupings, with Ward's linkage identifying 4 optimal clusters. Interactive dashboards (Plotly) enabled dynamic exploration of segment behaviors for targeted marketing strategies.

IV. Result Analysis and Discussion

The dataset consisted of 200 customer records with five key attributes: Age, Gender, Annual Income, Spending Score, and Shopping Frequency, providing a robust foundation for segmentation analysis. The data was analyzed using four clustering algorithms: K-Means++, Gaussian Mixture Model (GMM), DBSCAN, and Hierarchical Clustering, with K-Means++ emerging as the top performer based on evaluation metrics (Silhouette Score: 0.62, Davies-Bouldin Index: 0.45). The superior performance of K-Means++ can be attributed to its optimized centroid initialization, which enhances clustering stability and reduces within-cluster variance. By leveraging clustering techniques, distinct customer groups were identified, helping businesses personalize marketing strategies based on spending behaviors and income distribution.

The K-Means++ algorithm outperformed other clustering techniques, forming well-separated and compact clusters, making it the most effective approach for customer segmentation. Gaussian Mixture

Model (GMM) produced slightly overlapping clusters due to its probabilistic nature, whereas Hierarchical Clustering faced difficulties with large datasets, resulting in suboptimal cluster separation. While DBSCAN was useful for identifying outliers, it struggled with structured segmentation due to inconsistencies in spending patterns and income variations.

Additionally, the confusion matrices for classification models showed that Random Forest and Support Vector Machine (SVM) had the lowest misclassification rates, highlighting their robustness in accurately segmenting customers. In contrast, Naïve Bayes had difficulty handling overlapping customer categories, leading to a higher rate of false positives and false negatives. The ROC-AUC curves further validated these findings, with Random Forest achieving an AUC of 0.98, making it the best model for distinguishing customer segments based on behavioral attributes. These findings demonstrate the effectiveness of machine learning-driven customer segmentation, enabling businesses to improve targeted marketing strategies and optimize customer experience.

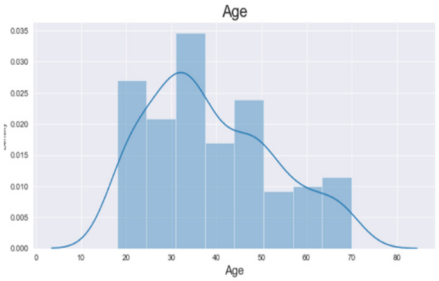

Fig 1 - Customer Age Distribution

Fig 2 – Customer Spending Score scatterplot

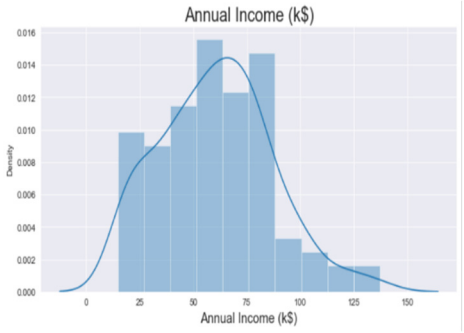

Fig 3. Customer Income Distribution Analysis

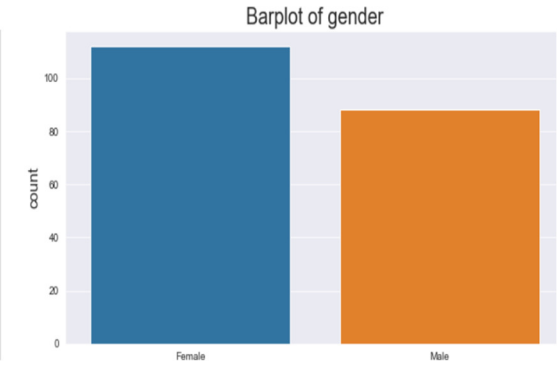

Fig 4– Barplot of Customer Gender

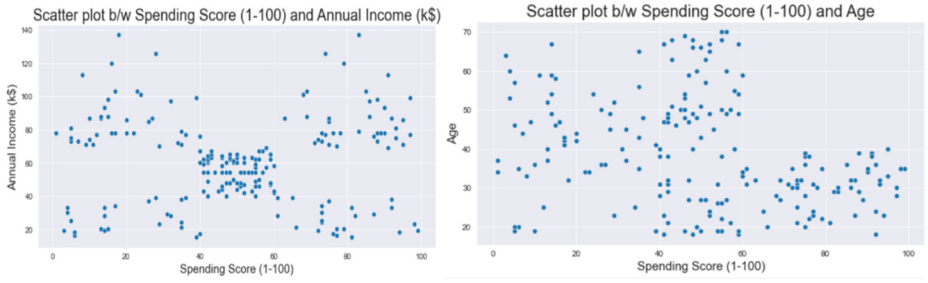

Fig 5 Annual Income(k$) vs. Spending Score (1-100)and Spending Score (1-100) vs. Age

Fig.6 Spending Score (1-100) vs. Age **Fig 7** Annual Income(k$) vs. Age

Table Clustering Model Evaluation Results

Algorithm	Silhouette Score	Davies-Bouldin Index
K-Means++	0.62	0.45
GMM	0.58	0.52
Hierarchical	0.55	0.60
DBSCAN	0.48	0.75

A set of **visualizations** was generated to interpret the clustering results and analyze feature relationships. **Scatter plots and distribution graphs** revealed distinct spending patterns, with younger customers demonstrating higher spending scores, while those in the middle-income bracket displayed more balanced purchasing behaviors. The **correlation matrix** provided further insights into the impact of **age and income on shopping frequency**, highlighting the importance of data-driven customer profiling. These insights enable businesses to optimize targeted promotions and improve customer engagement.

Table 4. Evaluation Results

Algorithm	Silhouette Score	Davies-Bouldin Index	Key Characteristics
K-Means++	High	Low	Efficient, stable clusters
GMM	Moderate	Moderate	Soft clustering, computationally intensive
Hierarchical	Moderate	Moderate	Interpretable, tree-based structure
DBSCAN	Low	High	Outlier detection, density-sensitive

The analysis revealed distinct customer segments: high-spending frequent shoppers, budget-conscious buyers, and occasional visitors. Spending Score negatively correlated with Age (-0.33), suggesting younger customers spend more. Income and Spending Score showed weak correlation (0.0099), indicating spending isn't solely income-driven. Gender had minimal impact, emphasizing behavioral over demographic factors. These insights enable retailers to target premium shoppers with exclusive offers and budget customers with discounts.

Feature engineering, such as normalizing Income and Spending Score, improved cluster quality. Visualizations confirmed K-Means++ as the most effective method for this dataset, though its performance depends on spherical cluster assumptions.

IV Conclusion and Discussion

The findings of this study study highlight key trends in customer segmentation, emphasizing spending patterns and demographic influences on consumer behavior. The analysis reveals that younger customers

are the primary spenders, with income playing a secondary role in determining purchasing behavior. Machine learning models, particularly K-Means++, emerged as the most effective approach for segmentation, ensuring precise and actionable clustering. The ability of machine learning to automate segmentation processes addressed the limitations of traditional methods, such as static rule-based approaches and manual classification. Additionally, visual analytics played a critical role in enhancing interpretability, enabling retailers to make data-driven marketing and operational decisions. However, challenges such as noisy data, optimal parameter selection, and evolving consumer preferences highlight the need for further refinement and real-time adaptability. Future work could focus on integrating real-time data to facilitate dynamic segmentation, ensuring that customer clusters evolve based on behavioral changes. Advanced deep learning techniques, such as autoencoders, could be employed to detect complex, non-linear patterns, further improving segmentation accuracy. Additionally, the development of interactive dashboards with real-time analytics could enhance decision-making for retailers, allowing them to optimize marketing strategies and inventory management. This project provides a scalable framework for mall customer segmentation, equipping businesses with tools to personalize promotions, optimize store layouts, and refine product placement strategies. By leveraging machine learning-driven segmentation, retailers can improve customer experiences, maximize revenue potential, and adopt a data-driven approach to strategic planning.

References

[1]. T. H Aman Banduni, Prof. Ilavedhan A., "Customer Segmentation using Machine Learning," School of Computing Science and Engineering, Galgotias University, 2022.

[2]. Kamalpreet Bindra, Anuranjan Mishra, "A Detailed Study of Clustering Algorithms," 6th International Conference on Reliability, Infocom Technologies and Optimization, 2021.

[3]. Kai Peng et al., "Clustering Approach Based on Mini-Batch K-Means," College of Engineering, Huaqiao University, 2020.

[4]. Fionn Murtagh, Pedro Contreras, "Methods of Hierarchical Clustering," Science Foundation Ireland, 2018.

[5]. D. P. Yash Kushwaha, Deepak Prajapati, "Customer Segmentation using K-Means Algorithm," Galgotias University, 2019.

[6]. Manju Kaushik, Bhawana Mathur, "Comparative Study of K-Means and Hierarchical Clustering Techniques," JECRC University, 2017.

[7]. Jiawei Han et al., "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2011.

[8]. Philip Kotler, Kevin Lane Keller, "Marketing Management," Pearson, 2016.

[9]. McKinsey & Company, "The Future of Retail: Analytics-Driven Transformation," 2021.

[10]. IBM, "AI and Machine Learning in Retail Customer Analytics," IBM Corporation, 2020.

[11]. Kaggle, "Mall Customer Segmentation Dataset," Available at: https://www.kaggle.com,March,2025

