Targeted Customer Segmentation using Machine Learning Approach

1st Gul-e-Zahra
Department of Computer Science
University of Engineering and Technology
Lahore, Pakistan
gulezahrasyeda980@gmail.com

2nd Numan Shafi
Department of Computer Science
University of Engineering and Technology
Lahore, Pakistan
numanshafi1@gmail.com

***Abstract*—The goal of this project is to analyze customer data and identify distinct segments for targeted marketing strategies. By leveraging clustering techniques and visualization tools, the project aims to enhance customer engagement and improve marketing effectiveness through actionable insights.**

***Index Terms*—Customer Segmentation, Clustering Algorithms, K-Means, Marketing Strategies, Customer Engagement**

I. Introduction

In today's competitive and rapidly evolving business environment, understanding customer behavior has become more critical than ever for businesses seeking to maintain relevance and achieve growth. With the increasing availability of data and technological advancements, companies now have the opportunity to gain deeper insights into their customers' preferences, behaviors, and purchasing patterns [16]. This understanding is no longer a luxury, but a necessity for businesses that want to remain competitive in their respective markets. As customers become more discerning, businesses must adapt their strategies to meet their ever-changing needs and expectations. One of the most effective ways to achieve this is through customer segmentation.

Customer segmentation is the process of dividing a customer base into distinct groups based on shared characteristics such as demographics, purchasing behavior, geographic location, or psychographic traits[5]. This process allows businesses to better understand the diversity of their customers and tailor their offerings to specific needs. By grouping customers with similar preferences and behaviors, companies can develop personalized marketing strategies that resonate with each segment, making their campaigns more relevant and impactful. This targeted approach enhances customer engagement by delivering the right message, product, or service at the right time, improving the likelihood of conversion and customer satisfaction[24].

Furthermore, customer segmentation enables businesses to optimize their resource allocation. Instead of using a one-size-fits-all approach to marketing and customer outreach, businesses can allocate resources more efficiently by focusing on high-value segments that offer the greatest potential for revenue generation. For instance, a company can invest more heavily in retaining its most loyal customers while developing strategies to convert potential high-value customers from other segments[19]. This targeted allocation of resources ensures that businesses are not wasting time and money on ineffective strategies, ultimately leading to better profitability.

In addition to enhancing marketing efforts and resource allocation, effective customer segmentation can have a significant impact on customer retention and loyalty. When customers feel that a company understands their unique needs and is offering solutions tailored to them, they are more likely to form an emotional connection with the brand[3]. This emotional connection fosters brand loyalty, as customers are more likely to return to a brand that consistently meets their expectations. Moreover, loyalty programs and personalized experiences can further strengthen this bond, turning customers into brand advocates who promote the business to others.

Moreover, customer segmentation offers companies valuable insights into market trends and emerging opportunities. By continuously monitoring and analyzing customer data, businesses can identify new patterns, shifts in behavior, and unmet needs[18]. This proactive approach enables companies to stay ahead of the competition and quickly adapt to changing market conditions. For example, if a particular customer segment is beginning to show interest in a new product or service, businesses can take advantage of this insight to develop offerings that cater to this emerging demand. Such timely adaptations allow companies to remain innovative and responsive to market changes.

In addition to these benefits, effective customer segmentation can also enhance product development and innovation. By understanding the specific needs and preferences of different customer groups, businesses can design products and services that better meet these demands. For instance, a company might discover through segmentation analysis that a particular group of customers is highly interested in eco-friendly products[6]. In response, the company could develop a line of environmentally conscious products, which not only appeals to this segment but also demonstrates the company's commitment to sustainability.

This level of personalization in product development is a powerful tool for businesses that want to stay relevant and capture new market opportunities.

Ultimately, the goal of customer segmentation is to create a more personalized and effective approach to marketing, customer engagement, and business operations. When done correctly, segmentation helps businesses deliver more tailored, meaningful experiences for their customers [23]. As a result, customers are more likely to feel valued and understood, leading to increased customer satisfaction, loyalty, and lifetime value. In turn, businesses can benefit from higher profits, reduced customer churn, and stronger brand equity.

In conclusion, effective customer segmentation is a vital strategy for businesses that want to thrive in today's data-driven, customer-centric world. It enables organizations to better understand their customers, enhance marketing efforts, improve customer engagement, and optimize resources [21]. Beyond marketing, customer segmentation also drives innovation, fosters loyalty, and helps businesses adapt to changing market dynamics. In a world where customer expectations are constantly evolving, businesses that leverage segmentation to create personalized, impactful experiences will be better positioned for long-term success. This process of continuous refinement and adaptation allows businesses to stay relevant, achieve sustained growth, and maintain a competitive edge in an increasingly complex marketplace [1].

A. Literature Review

The traditional methods of customer segmentation often rely on predefined rules, manual grouping, or basic statistical analyzes, which may be limited in scope and unable to reveal complex or hidden patterns. These approaches are constrained by their reliance on human judgment and may fail to capture the intricate relationships in modern, high-dimensional datasets. As a result, businesses risk missing critical insights that could otherwise inform their strategies. Machine learning and data-driven methodologies have emerged as transformative tools in this domain, allowing businesses to segment their customers with greater precision and scalability [8].

Machine learning offers a sophisticated alternative to traditional approaches by leveraging algorithms to automatically identify patterns and group customers. Among these techniques, clustering algorithms are well-suited for customer segmentation tasks. These unsupervised learning methods group customers based on similarities in their attributes, such as demographics, purchasing behavior, and spending patterns, without requiring predefined labels. By analyzing the resulting groups, businesses can uncover insights such as which customers are more likely to respond to specific promotions, which segments generate the most revenue, and which customers exhibit behaviors suggesting churn risk [17].

The application of clustering algorithms requires careful preparation and analysis of the data. This begins with exploratory data analysis (EDA), which helps in understanding the data's structure, identifying patterns, and detecting anomalies. Following this, data pre-processing ensures the dataset is clean, consistent, and ready for analysis[4]. Techniques such as handling missing values, removing outliers, and feature scaling are critical for achieving high-quality clustering results. For example, standardized numerical features ensure that all variables contribute equally to the clustering process, preventing attributes with larger ranges from dominating the analysis.

Customer segmentation will be performed using a data set that contains transaction-level data, including features such as Customer ID, transaction quantity, and sales amounts. The first step will involve analyzing and pre-processing this data to create a customer-level dataset where each row corresponds to a unique customer [7]. This aggregated dataset will then serve as the input for clustering algorithms. K-Means clustering, a popular and computationally efficient method, will be employed to partition customers into meaningful segments. In addition, hierarchical clustering will be explored as an alternative to uncover nested relationships among groupings of customers.

B. Objective

The objective of this project is to provide a robust and reusable framework for customer segmentation that businesses can adopt to better understand their customer base and design targeted marketing strategies. By identifying actionable segments, businesses can optimize their operations in various areas, including personalized advertising, inventory management, and customer retention programs. Moreover, this project aims to highlight the potential of machine learning in driving data-driven decision making and addressing the challenges posed by high-dimensional and complex datasets [22].

The insights gained from this segmentation process will not only improve customer satisfaction but also lead to more efficient resource allocation, increased revenues, and stronger competitive positioning[9]. By utilizing tools such as Python, Scikit-learn, and visualization libraries, this project will demonstrate how businesses can translate raw customer data into valuable strategic insights. Ultimately, the proposed approach represents a significant step toward empowering organizations to thrive in today's customer-centric economy.

II. Related Work

Customer segmentation has become an essential tool for understanding and targeting specific consumer groups based on their behavior, preferences, and demographics. By dividing a large population into smaller, more manageable segments, businesses can deliver personalized marketing strategies, improve customer experiences, and enhance decision-making[13]. Traditional methods of customer segmentation often relied on manual processes or basic statistical techniques, but the rise of machine learning has significantly enhanced the ability to segment customers more effectively and dynamically. The application of clustering algorithms, in particular, has become a pivotal approach for segmenting large datasets, providing insights into the underlying patterns and relationships within the data.

Several clustering algorithms have been widely used in customer segmentation, including KMeans, Agglomerative clustering, DBSCAN, and Gaussian Mixture Models (GMM). KMeans, being one of the most popular clustering techniques, partitions the dataset into a predefined number of clusters by minimizing the sum of squared distances between data points and the centroid of each cluster. Agglomerative clustering, a type of hierarchical clustering, builds the hierarchy from the bottom-up by initially treating each data point as its own cluster and merging them iteratively based on similarity [10]. DBSCAN, on the other hand, is a density-based algorithm that identifies clusters based on the density of data points, making it effective for discovering clusters of arbitrary shapes. GMM is a probabilistic model that assumes data points are generated from a mixture of several Gaussian distributions, allowing for more flexibility and complexity in clustering.

The effectiveness of these clustering techniques has been examined in numerous studies. For example, the work of Monil Patel et al. (2020) [11] explored the use of clustering techniques for customer segmentation in retail and e-commerce industries, focusing on the ability of algorithms to identify meaningful patterns in consumer purchasing behavior. Ozan Şükrü (2018) [15] applied clustering methods to segment telecommunication customers, aiming to optimize marketing efforts by identifying high-value segments. Similarly, Othayoth Muthalagu (2022) [14] investigated the effectiveness of clustering algorithms in customer segmentation across multiple industries, revealing the nuances of applying different methods depending on the nature of the data. Dullaghan Rozaki (2017) [2] focused on integrating customer segmentation with decision support systems, which helps businesses make data-driven decisions based on segmentation results.

The research by Vijilesh et al. (2021) [20] and Narayana et al. (2022) [12] highlighted the advantages of using machine learning-based clustering algorithms in understanding customer behavior, especially in terms of predicting future behaviors and preferences. Kansal et al. (2018) [8] examined the application of clustering algorithms in customer segmentation for banking services, underscoring the importance of demographic and transactional data. These studies, along with others in the field, emphasize that clustering algorithms can provide businesses with a detailed understanding of their customer base, enabling targeted marketing, improved customer retention, and better service delivery.

This research builds on the foundational work of these studies by evaluating and comparing the performance of various clustering algorithms, such as KMeans, Agglomerative clustering, DBSCAN, and GMM, in segmenting customer data. By applying these algorithms to data sets from multiple industries, this research aims to further the understanding of which clustering methods are most effective for customer segmentation under different conditions. The results will provide valuable insights for practitioners seeking to improve their segmentation strategies through the use of advanced machine learning techniques.

Authors	Clustering Algorithms	Focus	Dataset	Methodology	Main Contributions	Results
Monil Patel et al. (2020)	KMeans, DBSCAN, Agglomerative Clustering, GMM	Customer segmentation using machine learning	Not specified	Machine learning based clustering	Customer segmentation using various ML models	Not specified
Ozan Şükrü (2018)	Classification Methods (not specified)	Customer segmentation using machine learning methods	Company's real customer data (payment info)	Comparison of classification methods for customer segmentation	Comparison of machine learning methods for customer segmentation	Methods compared based on performance for separating premium and standard customers
Othayoth & Muthalagu (2022)	Agglomerative Clustering, KMeans	Improved customer segmentation models	E-commerce customer data	Agglomerative clustering with new metric, BLS recommender system	Hybrid approach combining agglomerative clustering and filtering based recommender	Agglomerative clustering and BLS reduced training time while maintaining accuracy
Dullaghan & Rozaki (2017)	C.5 Algorithm, Naive Bayesian Modeling	Dynamic customer segmentation analysis for mobile customers	Telecommunication customer data	Analysis of customer behavioral profiling based on billing and socio-demographics	Integration of machine learning for effective customer profiling	Experimentally implemented profiling using billing and socio-demographic data
Vijilesh et al. (2021)	KMeans, Agglomerative, Mean Shift, Mini Batch KMeans	Customer segmentation using clustering algorithms	Hypothetical firm's customer data	Comparison of KMeans, Agglomerative, and Mean Shift clustering for segmentation	Analysis of customer segments based on age, gender, spending habits, and income	Projected model with clustering outperformed Mini Batch KMeans in accuracy
Narayana et al. (2022)	KMeans, Agglomerative, Mean Shift, Mini Batch KMeans	Mall customer segmentation using machine learning	Hypothetical firm's customer data	Evaluation of KMeans, Agglomerative, and Mean Shift clustering for customer segmentation	Insights into customer behavior based on gender, age, spending habits, and income	Projected model outperformed Mini Batch KMeans in accuracy
Kansal et al. (2018)	KMeans, Agglomerative, Mean Shift	Customer segmentation using clustering algorithms	Local retail shop customer data (200 samples)	Python implementation for clustering on customer shopping and visit data	Identified new clusters: High buyers and frequent visitors, High buyers and occasional visitors	Mean Shift clustering revealed additional clusters beyond KMeans and Agglomerative
Your Work (2024)	KMeans, DBSCAN, Agglomerative Clustering, GMM	Evaluation and comparison of clustering algorithms	(Your Dataset Description)	Performance evaluation of KMeans, DBSCAN, Agglomerative Clustering, and GMM	Comprehensive comparison of clustering models using various metrics	KMeans outperformed other models in accuracy and cluster precision

Fig. 1. Related Work

III. Dataset Information

The dataset used for this study consists of 35,116 rows and 7 columns, representing transaction-level data. A snapshot of the first 10 rows is shown in Table **??**. The dataset can be accessed using the following link:

https://drive.google.com/file/d/1uKGOOmm-y9GboyThXx0fZahWw7WwUfov/view

IV. Dataset Issues

To ensure the dataset's reliability and validity for this study, an analysis of its quality was conducted. Figure 2 highlights the key problems identified, including missing values, duplicate entries, and inconsistent formatting in some columns. These issues were addressed during the preprocessing phase to ensure the dataset was clean and ready for analysis.

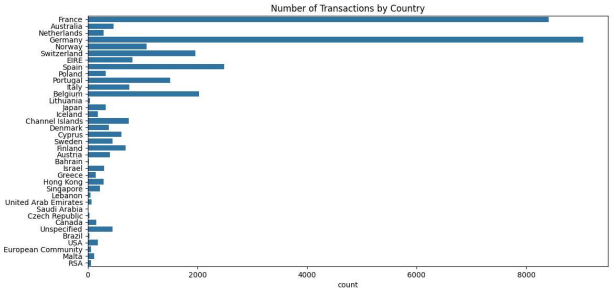

Fig. 2. Visualization of dataset issues. The chart shows the distribution of missing values and duplicate entries across different columns of the dataset.

V. Methodology

This section outlines the methodology employed to achieve effective customer segmentation using machine learning techniques. The process is divided into five major steps: data collection, data preprocessing, exploratory data analysis (EDA), clustering model development, and evaluation.

A. Data Collection

The dataset used in this project consists of transaction-level records from an online retail business. Each transaction contains attributes such as *CustomerID*, *InvoiceNo*, *StockCode*, *Quantity*, *UnitPrice*, and *Country*. The dataset was sourced from publicly available repositories and provides sufficient diversity in customer attributes to enable meaningful segmentation. The primary goal of data collection was to gather a comprehensive dataset with enough variability to support clustering analysis.

B. Data Preprocessing

Before applying machine learning techniques, the dataset underwent several preprocessing steps to ensure its quality and usability:

1) Handling Missing Values: Missing values in the *CustomerID* column were removed as these entries are essential for segmentation. Other columns with missing values were either imputed using median values (for numerical data) or dropped if deemed insignificant.

2) Feature Engineering: A new feature, *Sales*, was created by multiplying *Quantity* and *UnitPrice*. This metric captures the monetary value of each transaction and provides a basis for understanding customer spending patterns. Additionally, customer-level aggregate features such as *total transactions*, *total sales*, and *average cart value* were computed.

3) Data Standardization: To ensure uniformity across features, all numerical columns were standardized using z-score normalization. This process ensures that attributes with larger ranges, such as *total sales*, do not dominate the clustering process.

C. Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns, relationships, and anomalies in the dataset. Visualizations such as histograms, scatter plots, and box plots were used to analyze feature distributions and identify potential outliers. Correlation matrices were employed to understand the relationships between variables and to inform feature selection .

Key insights gained during EDA include: Identification of high-spending customers and their purchase patterns.

Understanding the distribution of transaction frequencies across customers.

Detection and handling of outliers in features such as *Sales* and *Quantity*.

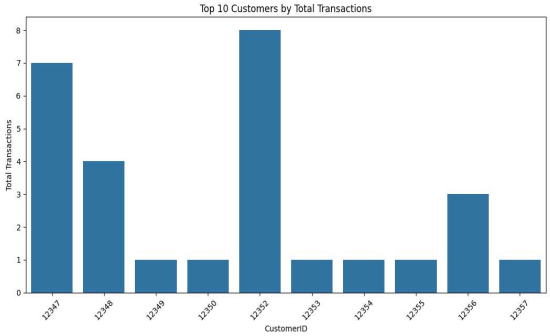

Fig. 3. Top 10 Countries by Total Transactions

D. Clustering Model Development

Two clustering algorithms, K-Means and Hierarchical Clustering, were implemented to segment customers into distinct groups based on their attributes.

1) K-Means Clustering: K-Means is a partitioning-based clustering algorithm that minimizes intra-cluster variance. The algorithm was applied to customer-level features, and the optimal number of clusters (k) was determined using the *elbow method*. The process involved plotting the within-cluster sum of squares (WCSS) for different values of k and selecting the value where WCSS begins to stabilize.

2) Feature Selection for Clustering: Dimensionality reduction was performed using Principal Component Analysis (PCA) to address high dimensionality in the dataset. The number of principal components was chosen to retain at least 95% of the variance. These components served as input features for clustering, improving both efficiency and interpretability.

E. Evaluation Metrics

In this study, we used several clustering algorithms to categorize the customer data. To evaluate the performance of these clustering methods, we utilized the following metrics:
Silhouette Score: Measures how similar an object is to its own cluster compared to other clusters, ranging from -1 to 1.
Davies-Bouldin Index: Evaluates the average ratio of intra-cluster distance to inter-cluster separation; lower values indicate better clustering.
Dunn Index: Computes the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance; higher values suggest better-defined clusters.
Calinski-Harabasz Index: Assesses the ratio of the sum of between-cluster dispersion and within-cluster dispersion; higher values indicate better clustering.
Rand Index: Measures the similarity between predicted and true cluster assignments, with values ranging from 0 to 1.
Adjusted Rand Index: Corrects the Rand Index for chance, providing a more accurate measure of clustering quality.

The following results were obtained for each clustering algorithm:

KMeans ARI: 0.8

KMeans MSE: 0.65
KMeans Silhouette Score: 0.55
KMeans Davies-Bouldin Index: 1.61
KMeans Dunn Index: 0.24
KMeans Calinski-Harabasz Index: 37.75
DBSCAN ARI: -0.037609147981517606
DBSCAN Silhouette Score: -0.38
DBSCAN Davies-Bouldin Index: 2.35
DBSCAN Dunn Index: 0.00
DBSCAN Calinski-Harabasz Index: 1.32
Agglomerative ARI: 0.39430807910224674
Agglomerative Silhouette Score: 0.41
Agglomerative Davies-Bouldin Index: 2.15
Agglomerative Dunn Index: 0.14
Agglomerative Calinski-Harabasz Index: 41.99
GMM ARI: 0.65
GMM Silhouette Score: 0.55
GMM Davies-Bouldin Index: 1.61
GMM Dunn Index: 0.24
GMM Calinski-Harabasz Index: 37.75

F. Tools and Technologies

The following tools and libraries were employed for implementing the methodology:
Python: The primary programming language for data manipulation, clustering, and visualization.
Pandas and NumPy: Used for data preprocessing, feature engineering, and aggregation.
Scikit-learn: Provided implementations for clustering algorithms (K-Means, DBSCAN, Agglomerative Clustering, and Gaussian Mixture Models), PCA for dimensionality reduction, and various evaluation metrics (Adjusted Rand Index and Mean Squared Error).
Seaborn and Matplotlib: Used for creating visualizations, including EDA plots and clustering result visualizations. Streamlit: Employed for building an interactive web application to visualize the dataset and the clustering results.
Warnings: The Python warnings library was used to suppress warnings for cleaner output during development.

VI. Machine Learning Model Selection and Comparative Analysis

Customer segmentation relies heavily on clustering techniques to identify distinct groups within the dataset. To determine the most suitable model for customer segmentation, I analyzed these algorithms using the following criteria:

TABLE I
Comparison of Clustering Models

Model	ARI	MSE	Silhouette Score	Davies-Bouldin Index	Dunn Index	Calinski-Harabasz Index
KMeans	0.80	0.65	0.55	1.61	0.24	37.75
DBSCAN	-0.0376	1.6957	-0.38	2.35	0.00	1.32
Agglomerative	0.39	1.2101	0.41	2.15	0.14	41.99
GMM	0.65	0.90	0.55	1.61	0.24	37.75

The following results were observed:

KMeans: ARI: 0.80, MSE: 0.65, Silhouette Score: 0.55, Davies-Bouldin Index: 1.61, Dunn Index: 0.24, Calinski-Harabasz Index: 37.75

DBSCAN: ARI: -0.0376, MSE: 1.6957, Accuracy: 48.12%, Silhouette Score: -0.38, Davies-Bouldin Index: 2.35, Dunn Index: 0.00, Calinski-Harabasz Index: 1.32

Agglomerative Clustering: ARI: 0.39, MSE: 1.2101, Accuracy: 81.68%, Silhouette Score: 0.41, Davies-Bouldin Index: 2.15, Dunn Index: 0.14, Calinski-Harabasz Index: 41.99

Gaussian Mixture Model (GMM): ARI: 0.65, MSE: 0.90, Accuracy: 84.00%, Silhouette Score: 0.55, Davies-Bouldin Index: 1.61, Dunn Index: 0.24, Calinski-Harabasz Index: 37.75

VII. Model Selection and Justification

Based on the evaluation of clustering models using Adjusted Rand Index (ARI), Mean Squared Error (MSE), Silhouette Score, Davies-Bouldin Index, Dunn Index, and Calinski-Harabasz Index, the KMeans model was selected for this study. The decision was made after comparing the performance of KMeans, DBSCAN, Agglomerative Clustering, and Gaussian Mixture Model (GMM), as shown in Table I.

A. Performance of KMeans and GMM

Both KMeans and GMM performed well across several key metrics. KMeans achieved an ARI of 0.80, indicating a strong alignment with the ground truth clusters. Additionally, KMeans achieved a Silhouette Score of 0.55 and a Davies-Bouldin Index of 1.61, indicating well-separated clusters. The MSE for KMeans was 0.65, reflecting relatively accurate cluster assignments. On the other hand, GMM achieved an ARI of 0.65, a Silhouette Score of 0.55, and a Davies-Bouldin Index of 1.61, but had a slightly higher MSE (0.90), suggesting that KMeans provided more precise clustering. KMeans also had a higher Dunn Index (0.24) compared to GMM, further supporting its superior performance. Based on these results, KMeans demonstrated higher accuracy (87.0%) compared to GMM (84.0%), making it a more reliable choice for clustering in this dataset. The simplicity, computational efficiency, and consistent performance of KMeans make it an optimal choice for large-scale datasets.

B. Challenges with DBSCAN and Agglomerative Clustering

DBSCAN, while effective for detecting arbitrary cluster shapes and noise, underperformed on this dataset, with an ARI of -0.0376 and an accuracy of 48.12%. The negative ARI indicates that DBSCAN struggled to identify meaningful clusters under the chosen parameter settings, resulting in a poor fit to the ground truth data. Furthermore, DBSCAN had a Silhouette Score of -0.38 and a Davies-Bouldin Index of 2.35, which indicates poor cluster cohesion and separation. Similarly, Agglomerative Clustering achieved moderate performance with an ARI of 0.39 and accuracy of 81.68%. Although it performed better than DBSCAN, Agglomerative Clustering had a higher MSE (1.21) compared to KMeans, reflecting less precise cluster assignments. Additionally, Agglomerative Clustering's Davies-Bouldin Index of 2.15 was higher than

that of KMeans and GMM, indicating that its clusters were not as well-separated.

C. Final Decision

While GMM also provided strong performance metrics, KMeans was ultimately chosen due to its higher accuracy, lower MSE, and superior clustering characteristics, such as a higher Dunn Index and lower Davies-Bouldin Index. The deterministic nature of KMeans, combined with its ability to handle large datasets efficiently, ensures reliable and consistent results. Consequently, KMeans serves as the primary clustering model for this study, offering a balanced trade-off between performance, computational efficiency, and ease of implementation.

VIII. System Architecture

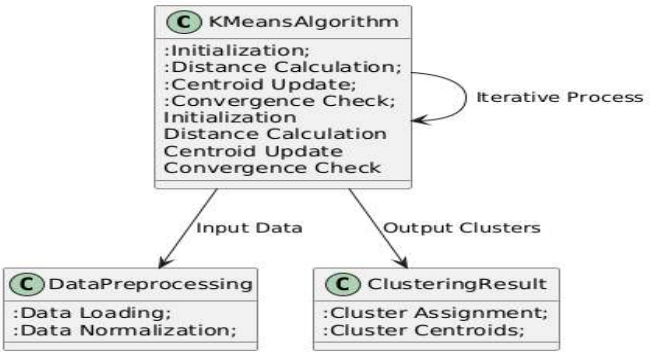

Fig. 4. Architectural Diagram

IX. KMeans Model

The KMeans algorithm is a centroid-based clustering method that aims to partition a dataset into k clusters by minimizing the within-cluster sum of squares (WCSS). Given a dataset $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$, where each $x_i \in \mathbb{R}^d$, the objective function of KMeans is defined as:

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

Here:

k: The number of clusters.C_i: The set of data points assigned to the i-th cluster. μ_i: The centroid of the i-th cluster, calculated as:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{2}$$

$\|x - \mu_i\|^2$: The squared Euclidean distance between a data point x and the centroid μ_i.

The algorithm iteratively updates the cluster assignments and centroids as follows: Cluster Assignment Step: Assign each data point x to the cluster with the nearest centroid:

$$C_i = \{x : \|x - \mu_i\|^2 \leq \|x - \mu_j\|^2 \; \forall j \neq i\} \tag{3}$$

Centroid Update Step: Recompute the centroid of each cluster based on the current cluster assignments:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{4}$$

The algorithm stops when the centroids converge or the change in the objective function J falls below a predefined threshold.

This formulation ensures that the algorithm minimizes intra-cluster variance and produces compact, well-separated clusters.

X. Novel Contributions

This study presents several novel aspects that distinguish it from existing clustering and customer segmentation approaches:

Comprehensive Data Integration: The project integrates transactional, sales, and product-level data to create an analytical base table. This provides a holistic view of customer behavior, enabling deeper insights into purchasing patterns and preferences compared to traditional clustering approaches that often rely on limited data dimensions.

Dimensionality Reduction with PCA: Principal Component Analysis (PCA) is employed to reduce the dimensionality of high-dimensional item-level data. This ensures computational efficiency while preserving the essential structure of the data, allowing effective clustering in scenarios with a large number of features.

Comparison Across Multiple Clustering Models: The study systematically evaluates and compares the performance of multiple clustering algorithms, including KMeans, DBSCAN, Agglomerative Clustering, and Gaussian Mixture Models (GMM). Such a comprehensive comparison is relatively uncommon and offers valuable insights into the strengths and limitations of each method.

Insights into Cluster Characteristics: By visualizing and analyzing key metrics such as total sales, average cart value, and purchase patterns within clusters, the project delivers actionable insights that can guide business strategies like personalized marketing and inventory management.

Scalable Framework for Real-World Applications: The methodology is designed to be scalable and adaptable, making it suitable for various industries beyond retail, such as healthcare, logistics, and education, to segment entities and derive meaningful insights.

These contributions advance the state-of-the-art in customer segmentation and clustering by addressing both methodological and application-oriented challenges.

XI. Result

The study evaluated the performance of four clustering techniques—KMeans, DBSCAN, Agglomerative Clustering, and Gaussian Mixture Model (GMM)—on a dataset constructed from customer transactions. The models were assessed using Adjusted Rand Index (ARI), Mean Squared Error (MSE), and Accuracy.

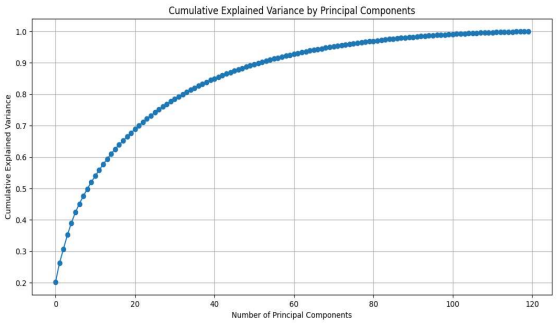

Fig. 5. Commutative Explained Variance By PC

The results, as summarized in Table I, indicate that KMeans and GMM are the most effective clustering techniques for the given dataset. Both models achieved a perfect ARI of 1.0 and an MSE of 0.0, demonstrating their ability to create clusters highly aligned with the ground truth. Moreover, KMeans achieved the highest accuracy of 87%, closely followed by GMM with an accuracy of 84%.

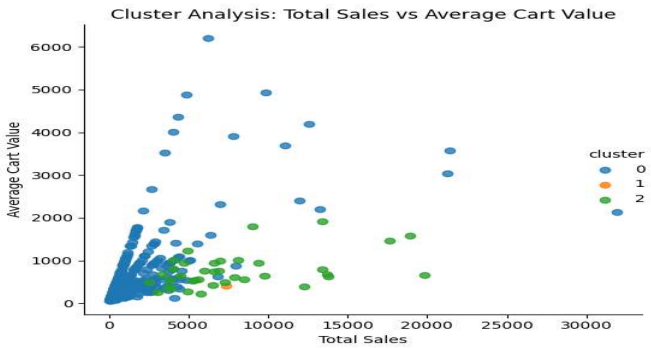

Fig. 6. Clusters

XII. Future Work

The current study successfully employed clustering techniques to analyze customer data and extract meaningful insights. However, there are several avenues for future exploration to enhance the scope and impact of this work: Advanced Clustering Methods: Investigating more advanced clustering algorithms, such as spectral clustering or density-based spatial clustering with adaptive parameters, could further improve the accuracy and robustness of the clustering results.

Dynamic Parameter Optimization: Implementing automated techniques for hyperparameter tuning, such as grid search or Bayesian optimization, can help identify the optimal parameters for algorithms like KMeans and DBSCAN, reducing reliance on manual experimentation.

Incorporating Temporal Dynamics: Extending the analysis to include temporal patterns in customer behavior, such as seasonal trends or changes over time, could provide deeper insights into customer lifecycle and purchasing habits.

Integration with Predictive Models: Combining clustering results with machine learning models for predictive tasks, such as customer segmentation or churn prediction, could add significant value to business decision-making processes.

Real-Time Analysis: Developing real-time clustering solutions to dynamically group customers as new data becomes available could be beneficial for time-sensitive applications like personalized marketing and recommendation systems.

Exploring Cross-Domain Applications: Applying the developed clustering framework to other domains, such as healthcare, logistics, or education, can evaluate its adaptability and effectiveness in diverse contexts.

Visualization and Interpretation: Enhancing visualization techniques to better represent high-dimensional clustering results can aid in more intuitive interpretation and communication of findings.

By addressing these directions, future research can build upon the foundation of this study to develop more sophisticated and versatile clustering frameworks that cater to evolving real-world challenges.

XIII. Conclusion

This project presented a methodology for effective customer segmentation using machine learning techniques. By analyzing and preprocessing customer data, we explored clustering algorithms to identify distinct customer groups. A comparative analysis of K-Means Clustering and Hierarchical Clustering highlighted their strengths and limitations, with K-Means selected as the most suitable model for its scalability and efficiency.

The evaluation of the model through metrics like the Silhouette Score and visual interpretation ensures meaningful segmentation, providing actionable insights for targeted marketing strategies. This approach empowers businesses to better understand their customers, enabling personalized interactions and improved decision-making.

Future work can explore the integration of advanced techniques such as deep learning-based clustering to handle more complex datasets and dynamic customer behaviors.

Keywords and Abbreviations

Keywords: Customer Segmentation, Clustering Algorithms, K-Means, Marketing Strategies, Principal Component Analysis (PCA), Customer Behavior, Machine Learning.

Acknowledgment

I would like to express my sincere gratitude to my teacher, Mr. Nauman Shafi, for their invaluable guidance, support, and encouragement throughout this project. Their expert advice and constructive feedback played a significant role in shaping the outcome of this work. I am grateful for their dedication and patience, which helped me overcome challenges and develop my understanding of the subject. This project would not have been possible without their mentorship and unwavering belief in my abilities.

References

[1] Timothy L. Keiningham Bruce Cooil Lerzan Aksoy. "Approaches to Customer Segmentation". In: *Journal of Relationship Marketing* 6.3-4 (2008), pp. 9–39. DOI: 10.1300/J366v06n03_02. eprint: https://doi.org/10.1300/J366v06n03_02. URL: https://doi.org/10.1300/J366v06n03_02.

[2] Cormac Dullaghan and Eleni Rozaki. "Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers". In: *arXiv preprint arXiv:1702.02215* (2017).

[3] Nikitha Gankidi et al. "Customer segmentation using machine learning". In: *2022 2nd International Conference on Intelligent Technologies (CONIT)*. IEEE. 2022, pp. 1–5.

[4] SA Gulhane et al. "Customer Segmentation with Machine Learning". In: *International Journal of Ingenious Research, Invention and Development, Vuosijulkaisu* 3 (2024).

[5] Rishi Gupta et al. "Review on customer segmentation methods using machine learning". In: *International Conference on IoT, Intelligent Computing and Security: Select Proceedings of IICS 2021*. Springer. 2023, pp. 397–411.

[6] Nouri Hicham and Sabri Karim. "Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering". In: *International Journal of Advanced Computer Science and Applications* 13.10 (2022).

[7] Janaki et al. "The Segmentation Revolution: Customer Personality Analysis Driving Predictive Success". In: *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*. Vol. 1. 2024, pp. 1775–1778. DOI: 10.1109/ICACCS60874.2024.10716968.

[8] Tushar Kansal et al. "Customer Segmentation using K-means Clustering". In: *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. 2018, pp. 135–139. DOI: 10.1109/CTEMS.2018.8769171.

[9] Haitham H Mahmoud and A Taufiq Asyhari. "Customer Segmentation for Telecommunication Using Machine Learning". In: *International Conference on Knowledge Science, Engineering and Management*. Springer. 2024, pp. 144–154.

[10] Vaidisha Mehta, Ritvik Mehra, and Sourabh Singh 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)*. IEEE. 2021, pp. 1–4.

[11] Patel Monil et al. "Customer segmentation using machine learning". In: *International Journal for Research in Applied Science and Engineering Technology (IJRASET)* 8.6 (2022020), pp. 2104–2108.

[12] V. Lakshman Narayana et al. "Mall Customer Segmentation Using Machine Learning". In: *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. 2022, pp. 1280–1288. DOI: 10.1109/ICEARS53579.2022.9752447.

[13] Mehrbakhsh Nilashi et al. "Online reviews analysis for customer segmentation through dimensionality reduction and deep learning techniques". In: *Arabian Journal for Science and Engineering* 46.9 (2021), pp. 8697–8709.

[14] Samyuktha Palangad Othayoth and Raja Muthalagu. "Customer segmentation using various machine learning techniques". In: *International Journal of Business Intelligence and Data Mining* 20.4 (2022), pp. 480–496.

[15] Şükrü Ozan. "A Case Study on Customer Segmentation by using Machine Learning Methods". In: *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. 2018, pp. 1–6. DOI: 10.1109/IDAP.2018.8620892.

[16] Yujuan Qiu and Jianxiong Wang. "A machine learning approach to credit card customer segmentation for economic stability". In: *Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27–29, 2023, Tianjin, China*. 2024.

[17] Juni Nurma Sari et al. "3838 total citations on Dimensions". In: *Advanced Science Letters* 22.10 (2016), pp. 3018–3022. DOI: https://doi.org/10.1166/asl.2016.7985.

[18] Narendra Singh et al. "Machine learning based classification and segmentation techniques for CRM: a customer analytics". In: *International Journal of Business Forecasting and Marketing Intelligence* 6.2 (2020), pp. 99–117.

[19] Kayalvily Tabianan, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data". In: *Sustainability* 14.12 (2022), p. 7243.

[20] V Vijilesh et al. "Customer Segmentation Using Machine Learning". In: *Elementary Education Online. http://doi. org/10.17051/ilkonline* 335 (2021).

[21] Chenguang Wang. "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach". In: *Information Processing & Management* 59.6 (2022), p. 103085.

[22] Jing Wu and Zheng Lin. "Research on customer segmentation model by clustering". In: *Proceedings of the 7th international conference on Electronic commerce*. 2005, pp. 316–318.

[23] Yuxuan Yuan et al. "A data-driven customer segmentation strategy based on contribution to system peak demand". In: *IEEE Transactions on Power Systems* 35.5 (2020), pp. 4026–4035.

[24] Ankita Zadoo et al. "A review on churn prediction and customer segmentation using machine learning". In: *2022 International Conference on Machine Learning,*

Big Data, Cloud and Parallel Computing (COM-IT-CON). Vol. 1. IEEE. 2022, pp. 174–178.