**Customer Segmentation: Analyze customer data and segment them into distinct groups based on their purchasing behavior** M.Atish

*Department of Computer Science Lovely Professional University* Punjab, India

mondi@lpu.in

**1.ABSTRACT**

Customer segmentation is a crucial technique that helps businesses analyze and categorize their customer base into smaller, homogeneous groups based on behavioral patterns and demographic traits. This study employs **K-means clustering**, a popular unsupervised machine learning algorithm, to segment customers based on attributes such as Annual Income, Spending Score, and Age. To ensure optimal clustering, the **elbow method** is used to determine the appropriate number of clusters. The methodology incorporates **Euclidean distance** for cluster assignment and evaluates cluster compactness through **inertia**. The results uncover three distinct customer groups, providing actionable insights for personalized marketing strategies. This study highlights the role of machine learning in optimizing resource allocation and improving customer engagement through data-driven segmentation.

**Keywords**

K-means clustering, machine learning, customer segmentation, elbow method, Euclidean distance

**2.INTRODUCTION**

"Customer segmentation is a key strategy for businesses to understand and categorize their customers based on behavior and demographics. Using \*\*K-Means clustering\*\*, this project analyzes data attributes such as

\*\*Annual Income\*\*, \*\*Spending Score\*\*, and \*\*Age\*\* to identify distinct customer groups. The Elbow Method is applied to determine the optimal number of clusters, ensuring meaningful segmentation. These insights help businesses create targeted marketing strategies, enhance customer engagement, and maximize profitability."

2.1Overview

In today’s competitive business environment, understanding customer behavior is paramount to achieving sustainable growth. Businesses are increasingly relying on data analytics to derive meaningful insights from vast datasets. **Customer segmentation**, the process of dividing customers into groups based on shared traits, is a powerful tool for designing targeted marketing campaigns, improving customer retention, and maximizing profitability.

Traditional segmentation techniques often rely on simplistic demographic data, leading to generalizations that fail to capture intricate behavioral patterns. Advances in machine learning, particularly **K-means clustering**, have revolutionized customer segmentation by enabling the identification of nuanced patterns in high-dimensional data.

2.2 Objective

The primary objective of this study is to segment customers into distinct groups using K-means clustering. By analyzing Annual Income, Spending Score, and Age, the study aims to provide actionable insights that can enhance marketing strategies and improve customer engagement.

2.3 Relevance of Features

**Annual Income:** Reflects a customer’s purchasing power and economic capacity.

**Spending Score:** A score assigned by businesses, indicating customer spending habits.

**Age:** Offers insights into generational preferences and purchasing tendencies.

By combining these features, the study provides a holistic view of customer behavior, facilitating effective segmentation.

**2.4 TOPICS COVERED**

This paper addresses the following key topics:

Data Preprocessing: Handling feature scaling using z-score normalization.

Clustering with K-Means: Cluster formation based on centroid optimization.

Elbow Method: Identifying the optimal number of clusters through variance analysis.

Mathematical Concepts: Incorporating formulas for Euclidean distance and inertia.

Visualization and Interpretation: Using scatter plots to analyze clustering outcomes.

**3. FORMULAS USED**

**3.1 Euclidean Distance**

The Euclidean distance formula is fundamental to K-means clustering. It calculates the straight-line

distance between two points in multidimensional space:

d(p,q)=∑i=1n(qi−pi)2d(p,q)=i=1∑n​(qi​−pi​)2​

Where:

pp and qq are two points in n-dimensional space

qi and pi​ are the ith coordinates of qq and pp, respectively

**3.2 Inertia**

Inertia measures the within-cluster variance, which the K-means algorithm minimizes to create compact clusters. It is calculated as:

Inertia=∑i=1k∑x∈Ci∣∣x−μi∣∣2Inertia=i=1∑k​x∈Ci​∑​∣∣x−μi​∣∣2

Where:

* k is the number of clusters
* Ci​ represents the ith cluster
* x is a data point in Ci​
* μi​ is the centroid of Ci​

Lower inertia values indicate tighter clusters, making the segmentation more meaningful.

**4.METHODOLOGY**

This study begins by identifying key customer features such as annual income and spending score for clustering. Clustering techniques are chosen based on their suitability for segmentation tasks. The dataset is reviewed to ensure completeness and relevance. Metrics like WCSS are used to determine the optimal number of clusters. These steps ensure the results are accurate and actionable.

**4.1Data Preprocessing**

The dataset includes attributes such as Annual Income, Spending Score, and Age. To ensure fair contribution from each feature, the data is standardized using **z-score normalization**, which scales each feature to have a mean of 0 and a standard deviation of 1.

4.2 K-Means Clustering

K-means clustering is an iterative algorithm that minimizes within-cluster variance by assigning data points to the nearest cluster centroid. The algorithm recalculates centroids iteratively until no further optimization is possible.

Steps in K-means clustering:

1. Initialize k cluster centroids randomly.
2. Assign each data point to the nearest centroid based on **Euclidean distance**.
3. Recalculate centroids as the mean of points in each cluster.
4. Repeat steps 2 and 3 until centroids stabilize.

4.3 Elbow Method

To determine the optimal number of clusters, the elbow method is used. It involves plotting the inertia values for various k values and identifying the "elbow point," where the curve begins to flatten. This point balances model complexity with clustering accuracy.

**5.RESULTS AND ANALYSIS**

**5.1 Elbow Plot and Optimal Clusters**

The elbow plot revealed that three clusters were optimal for this dataset. Beyond three clusters, the reduction in inertia was minimal, indicating diminishing returns from additional clusters.

A diagram of a customer segmentation

Description automatically generated

**Table 1:** **Inertia Values for Different k**

|  |  |
| --- | --- |
| **Number of Clusters (k)** | **Inertia** |
| **1** | **22000.5** |
| **2** | **14050.3** |
| **3** | **9200.1** |
|  |  |

**5.2 Cluster Profiles**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Characteristics** | **Average Income** | **Spending Score** | **Age** |
| **Cluster 0** | **Moderate income and moderate spending habits** | **$50,000** | **55** | **30** |
| **Cluster 1** | **High income, high spending, older demographics** | **$85,000** | **75** | **40** |
| **Cluster 2** | **Low income and low spending, diverse age group** | **$25,000** | **30** | **28** |

**5.3 Gender Distribution**

|  |  |  |
| --- | --- | --- |
| **Cluster** | **Male (%)** | **Female (%)** |
| **Cluster 0** | **52** | **48** |
| **Cluster 1** | **60** | **40** |
| **Cluster 2** | **49** | **51** |

The gender distribution table provides additional insights into demographic patterns within clusters.

**6.DISCUSSION**

This segmentation process provides actionable insights for businesses by revealing distinct customer groups with unique characteristics:

Cluster 0: This group consists of customers with moderate income and spending habits. These individuals are generally price-sensitive and prefer value-based products. Businesses targeting this segment should focus on loyalty programs, discounts, and mid-range products. Offering reward systems that encourage repeat purchases can help strengthen brand loyalty among these customers. Additionally, promotional campaigns centered around affordability and value can resonate strongly with this group.

Cluster 1: This group includes high-income customers with high spending scores. These individuals are typically older and value premium products and personalized services. To appeal to this segment, businesses should prioritize exclusive offerings, premium quality products, and high-touch customer service. Luxury brands can particularly benefit by targeting this segment with VIP memberships, early access to new collections, and personalized shopping experiences. Furthermore, leveraging data-driven insights to offer tailored recommendations can improve customer satisfaction and long-term engagement.

Cluster 2: Customers in this group have low income and low spending scores. These individuals prioritize affordability and often make purchasing decisions based on budget constraints. Effective strategies for this cluster include offering discounts, bundled deals, and entry-level products. Retailers can enhance their appeal by advertising cost-effective solutions and creating campaigns that emphasize value for money. Targeted messaging that highlights savings and practical benefits can effectively capture the attention of this group.

By tailoring marketing strategies to each cluster, businesses can enhance engagement, allocate resources more efficiently, and improve overall profitability. The ability to identify and understand such diverse customer groups underscores the power of machine learning in business analytics.

**7. CONCLUSION**

This study demonstrates the effectiveness of K-means clustering for customer segmentation, showcasing its ability to group customers into meaningful clusters based on their purchasing behavior and demographic characteristics. The analysis of features like Annual Income, Spending Score, and Age revealed three distinct customer segments, each with actionable insights for targeted marketing.

The segmentation process provides businesses with a roadmap for crafting personalized marketing strategies that align with the needs and preferences of different customer groups. For instance, while high-income customers may appreciate premium experiences, price-sensitive customers are more likely to respond to value-driven promotions. These insights can help businesses enhance customer satisfaction, drive repeat purchases, and maximize revenue.

In addition to its practical applications, this study highlights the importance of combining data preprocessing techniques with clustering algorithms to ensure accurate and unbiased results. The use of the elbow method to determine the optimal number of clusters further adds to the robustness of the segmentation process.

**Future Directions**

While the current study focused on three primary features, future research could explore additional dimensions such as:

* Product Preferences: Identifying clusters based on specific categories of products or services preferred by customers.
* Geographic Locations: Incorporating location-based data to account for regional preferences and market variations.
* Purchase Frequency: Analyzing buying patterns to identify habitual versus sporadic buyers.

Advanced clustering techniques, such as hierarchical clustering or Gaussian Mixture Models, could also be used to uncover deeper patterns within customer behavior. Furthermore, integrating time-series analysis could provide insights into evolving customer preferences, enabling businesses to adapt their strategies dynamically.

**8.REFERENCES**

1. Shih, C., & Lin, J. (2021). *Customer Segmentation Using Machine Learning Techniques*. Journal of Marketing Analytics, 29(2), 157–172.
2. Lee, Y., et al. (2020). *Optimizing Clustering for Retail Analytics*. IEEE Transactions on Knowledge and Data Engineering, 32(5), 1209–1223.
3. Xu, R., & Wunsch, D. (2005). *Survey of Clustering Algorithms*. IEEE Transactions on Neural Networks, 16(3), 645–678.
4. Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier, Chapter 8: Cluster Analysis.
5. Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O’Reilly Media.
6. Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
7. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
8. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. Pearson Education.
9. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.
10. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Wiley.
11. MacQueen, J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.
12. Bholowalia, P., & Kumar, A. (2014). *EBK-Means: A Clustering Technique Based on Elbow Method and K-Means in WSN*. International Journal of Computer Applications, 105(9), 17–24.
13. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.
14. Rokach, L., & Maimon, O. (2005). *Clustering Methods*. In *Data Mining and Knowledge Discovery Handbook* (pp. 321–352). Springer.
15. Lloyd, S. (1982). *Least Squares Quantization in PCM*. IEEE Transactions on Information Theory, 28(2), 129–137.
16. Berkhin, P. (2006). *A Survey of Clustering Data Mining Techniques*. In *Grouping Multidimensional Data* (pp. 25–71). Springer.
17. Kriegel, H. P., Kröger, P., & Zimek, A. (2011). *Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering*. ACM Transactions on Knowledge Discovery from Data, 3(1), 1–58.
18. Theodoridis, S., & Koutroumbas, K. (2008). *Pattern Recognition*. Elsevier.
19. Reddy, C. K., & Aggarwal, C. C. (2013). *Data Clustering: Algorithms and Applications*. CRC Press.
20. Wu, X., et al. (2008). *Top 10 Algorithms in Data Mining*. Knowledge and Information Systems, 14(1), 1–37.