Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Application Of K-means Clustering Algorithm In Consumer Shopping Behavior Segmentation In E-commerce

Andrean Japardi[1]*, Edi[2], Feriani Astuti Tarigan[3]

[1]*Informatics Engineering, STMIK Time, Medan, Indonesia*
[2,3]*Information Systems, STMIK Time, Medan, Indonesia*
andreanjapardi17@gmail.com[1*], edi_foe@yahoo.com[2] , ferianiastutitime@gmail.com[3]

Abstract

Companies are forced by fierce rivalry to be the best at satisfying customer requirements in order to keep clients from moving to rivals. Therefore, an algorithm—such as the K-means Clustering algorithm—is required to segment customer buying behavior so that businesses may more accurately satisfy demands. Using a simulated dataset from Kaggle that contains nine variables with information on consumer purchasing behavior, this study attempts to apply K-means Clustering to segment customer shopping behavior in e-commerce and assess its efficacy. Pre-processing, normalization, and the selection of three important numerical features—AvgTotalSpend, AvgSpendPerTrans, and LoyaltyScore—are all part of the analytic process. The elbow technique and silhouette score are used to determine the ideal number of clusters. The density between clusters is also evaluated using the Dunn index. Three separate consumer clusters are identified by the segmentation results: Cluster 0 has very low values for AvgTotalSpend, AvgSpendPerTrans, and LoyaltyScore; Cluster 1 has reasonably high values for all three attributes; and Cluster 2 has low LoyaltyScore but high AvgTotalSpend and AvgSpendPerTrans. According to these results, customers in Cluster 1 are more likely to make repeat purchases in the future, which offers insightful information for focused marketing campaigns.

Keywords: *consumer,segmentation,k-means,clustering,behavior*

1. Introduction

In the sales and distribution sector, competition between companies with comparable segmentation is common and usual. Due to intense competition, businesses must aim to be the most cutting-edge and best at satisfying customer requirements. in order to prevent customers from going to other businesses [1]. Many of these businesses, however, have struggled to separate consumer purchasing patterns, which has led to the loss of potential customers and may result in financial losses [2]. In order for these businesses to exactly address the demands of their customers, an algorithm that can segment based on consumer buying behavior is required. One such algorithm is the K-means Clustering algorithm. One of the most often used segmentation techniques is K-means Clustering. Finding the first centroid point and then figuring out the closest centroid distance is how K-means Clustering operates. To ensure that it stays the same, it calculates a new centroid after determining the closest centroid distance and then recalculates from the centroid distance calculation till the new centroid calculation. This study aims to test the effectiveness of the K-means Clustering algorithm in segmenting consumer shopping behavior based on a simulated e-commerce dataset and to apply the algorithm to segment consumer shopping behavior on those datasets.

2. Literature Review

2.1. Aggregation Data

Since it will generate new data that is more helpful in executing the segmentation process using K-means than data that is still in the form of raw data, the process of data aggregation based on certain variables as a reference for data aggregation is crucial [3].

2.2. RFM Process

RFM stands for recency, frequency, and monetary, where monetary is the total amount of money spent on a purchase, frequency is the frequency of a consumer's purchases, and recency is the last time they made a purchase [4].

2.3. Elbow Methods

One technique for figuring out the ideal number of clusters is the elbow approach. To choose the best clustering outcome, the elbow approach is used to compute and compare the SSE values for each cluster[8].

2.4. Silhouette Score

The ideal number of cluster values may also be determined using the silhouette score. This approach starts by figuring out the silhouette clusters' range. The Euclidean approach is then used to determine the silhouette value for each data point for each cluster value. Plotting is done once all cluster values have been determined. To find the greatest cluster value around 1, plotting is done. Excellent cluster separation is indicated by this, whereas cluster overlap is indicated by a low score. The number of clusters that yield the best cluster division is more clearly shown by this research[8].

2.5. Dunn Index

The distance between clusters is calculated using the Dunn index. The degree of separation between clusters is measured using the Dunn index. The better the separation between the generated clusters, the higher the Dunn index score. Clusters that are clearly separated from one another lessen the possibility of overlapping since they are not too near to one another. As a result, a high Dunn index score suggests that the final clusters are more homogenous and isolated from one another, which eventually enhances the analysis's quality[9].

2.6. Outlier

A number in a data collection that significantly deviates from the rest of the data is called an outlier. Measurement mistakes, natural variability, or specific conditions that do not exist in that portion of the data are the causes of these outlier numbers. Because K-means Clustering is highly sensitive to extreme values, outlier data can significantly affect its effectiveness. Outliers can lead the cluster center to be pulled into the extreme value, which can impede proper cluster formation. In order to guarantee that the clustering process can function as efficiently as possible and that the segmentation results that are produced can accurately represent the patterns of actual consumer buying behavior, it is crucial to find data that differs significantly from other data [5].

2.7. Boxplot

A statistical graph called a boxplot is used to display data dispersion. The graph's form, which resembles a box with lines extending above and downward, is what gives it the name boxplot. A popular tool in many domains, particularly quality management, is the boxplot. to aid in the identification of outliers by researchers. The boxplot displays a rectangular box in the center of the data. The box's top and lower borders represent the upper quartile (Q3) and lower quartile (Q1). Outliers are defined as 1.5 times the interquartile range (Q3-Q1) from the box's upper edge.[6].

2.8. Flask

Python created the web framework Flask. Websites are created with it. Because the Flask framework provides fundamental capabilities like a database and does not require any special tools or libraries, it falls into the micro-framework category. Flask is separated into two categories: template files that include Jinja templates, such as HTML pages, and static files that contain status codes needed for a website, such as CSS code, JavaScript, and picture files. Flask has the benefit of being flexible, allowing developers to build modifications in accordance with project requirements without being constrained by a convoluted framework [7].

3. Methods

3.1. Dataset

The dataset used in this study is the UrbanMart Transactions Dataset (2023-2024) obtained from the Kaggle platform. This dataset contains e-commerce consumer transaction data such as TransactionID, CustomerID, TransactionDate, TransactionValue, PaymentMethod, CustomerGender, CustomerAgeGroup, Region, and ProductCategory.

3.2. Pre-processing

In order to increase data quality and make the cluster findings more indicative of consumer buying behavior, the raw data is first handled through a pre-processing stage before the clustering procedure is executed. Data aggregation based on CustomerID, which combines transactions coming from the same customer into a single data item, is one of the primary pre-processing steps. Three primary numerical characteristics are derived from this aggregation's data. The average amount spent by each customer is known as AvgTotalSpend, the average amount spent by each customer per transaction is known as AvgSpendPerTrans, and the loyalty score indicates how frequently and how much a customer spends on shopping.

3.3. Determining the number of clusters

Three evaluation methods—the elbow method, the silhouette score, and the Dunn index—will be employed to ascertain the ideal number of clusters. The elbow technique uses a point that drastically lowers to form an elbow in order to calculate the ideal cluster value. The ideal cluster value is the point just at the elbow's junction. By examining the score produced by the silhouette approaching 1 or -1, the silhouette score establishes the ideal cluster value. The cluster value is ideal when the number is near 1, while it is not optimal when the number is near -1. Internal density and cluster separation are measured using the Dunn Index. The more effective the segmentation and the greater the separation between clusters, the higher the Dunn Index value.

3.4. Clustering Process

Three previously acquired numerical features—TotalSpend, AvgSpendPerTrans, and LoyaltyScore—are used in the clustering method. This clustering procedure will employ the kmeans++ approach, which is superior at creating clusters.

3.5. Visualization and Evaluation of Results

This research will make use of 3D PCA, pie charts, and boxplots to show and assess the distribution between the created clusters. Pie charts are used to see how the clusters are distributed, boxplots are used to find potential outliers, and 3D PCA is used to see how the clusters are distributed.

3.6. System Implementation

To make it easier for users to complete the segmentation process efficiently, the entire process—including pre-processing, figuring out how many clusters to employ, clustering, and visualization—is done with the Python programming language and incorporated into a Flask-based website.

4. Results

As seen in the illustration below, the research findings are displayed as an analysis figure:

Cluster Analysis Results

	Cluster	TopProductCategory	DominantAgeGroup	DominantGender	DominantRegion	PaymentMethod	AvgTotalSpend	AvgSpendPerTrans	LoyaltyScore
0	0	Beauty	26-35	Female	Medan	Bank Transfer	52941665.19	2320117.81	1.92
1	1	Electronics	26-35	Female	Surabaya	Bank Transfer	71365659.02	2570458.40	23.14
2	2	Beauty	26-35	Female	Surabaya	Bank Transfer	72809724.12	2741741.35	2.19

Fig. 1: Cluster Analysis Results

The cluster findings are shown not just in tabular form but also as pie charts and 3D PCA graphics, as shown in the following image:

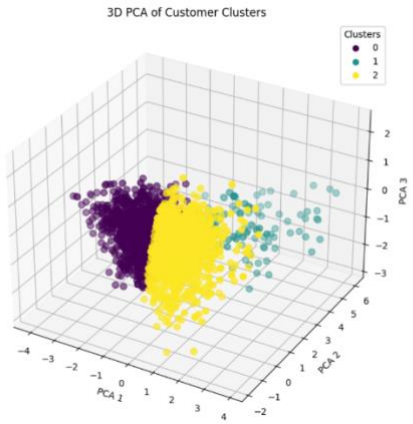

Fig. 2: PCA 3D

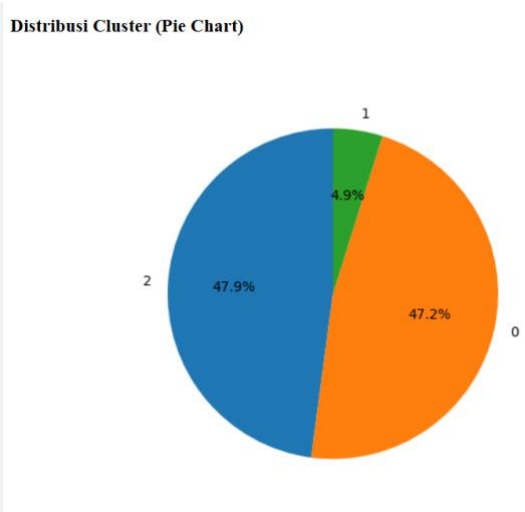

Fig. 3: Pie Chart

Cluster 1 is evident from the above figures, with a relatively high AvgTotalSpend of 71,365,659 rupiah, AvgSpendPerTrans of 2,570,458 rupiah, and LoyaltyScore of 23.14. Cluster 2's average total expenditure is 72,809,724 rupiah, its average spend per transfer is 2,741,741 rupiah, and its loyalty score is 2.19. Cluster 0's average total expenditure is 52,941,665 rupiah, its average spend per transfer is 2,320,117 rupiah, and its average loyalty score is 1.92.

5. Conclusion

Using a simulated e-commerce dataset, this research effectively segmented customer buying behavior using the K-Means Clustering technique. According to the clustering results, customers might be divided into a number of groups based on several attributes, including spending caps and loyalty to a retailer. A reasonably excellent and clearly differentiated segmentation was produced by the K-Means algorithm, according to evaluation utilizing the elbow technique, silhouette score, and dunn index. As a result, the research goal of using and evaluating K-Means Clustering's capacity to segment customer purchasing behavior has been accomplished.

References

[1] H. A. "Penerapan Data Mining Menggunakan Metode K-Means Clustering Untuk Pengelompokkan Data Pelanggan (Studi Kasus : PT Pinus Merah Abadi)," *Jurnal Web Informatika Teknologi (J-WIT),* vol. 6, no. 1, pp. 1-8, 30 June 2021.

[2] K. T. S. V. and V. R. , "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability,* vol. 14, no. 12, 13 June 2022.

[3] I. P. V. S. Hadi Sukmawati and Y. W. , "Implementasi Algoritma K-Means Pada Klasterisasi Tingkat Kasus Stunting Di Kabupaten Batang," *Jurnal Infortech,* vol. 6, no. 2, pp. 101-106, 6 December 2024.

[4] B. T. Kristanti, A. J. and E. P. Mandyartha, "Implementasi *K-Means Clustering* Dalam Segmentasi Pelanggan Berdasarkan Usia, Pendapatan Dan Model *RFM* (Studi Kasus: Lantikya Store Jombang)," *JITET (Jurnal Informatika dan Teknik Elektro Terapan),* vol. 12, no. 3, 7 August 2024.

[5] A. F. Zabidi, "Penerapan Algoritma K-Means untuk Pengelompokan Koleksi Perpustakaan dengan Data Mining," *Media Jurnal Informatika,* vol. 16, no. 2, p. 233–242, December 2024.

[6] G. R. and Y. A. , "Perbandingan Kinerja Algoritma K-Means dan Agglomerative Clustering untuk Segmentasi Penjualan Online pada Customer Retail," *Jurnal Informatika: Jurnal Pengembangan IT (JPIT),* vol. 9, no. 1, p. 92–96, January 2024.

[7] D. F. Ningtyas and N. S. , "Implementasi Flask Framework Pada Pembangunan Aplikasi Purchasing Approval Request," *Jurnal Janitra Informatika dan Sistem Informasi,* vol. 1, no. 1, p. 19–34, 16 April 2021.

[8] R. I. N. and A. , "Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest," *Journal of Electrical and Electronics Engineering,* vol. 7, no. 1, January 2022.

[9] R. A. Sary, N. S. and W. A. , "Application of K-Means++ with Dunn Index Validation of Grouping West Kalimantan Region Based on Crime Vulnerability," *Jurnal Ilmu Matematika dan Terapan,* vol. 18, no. 4, p. 2283–2292, 14 October 2024.