USER SEGMENTATION BASED ON PURCHASING HABITS AND PREFERENCES ON THE AMAZON PLATFORM USING K-MEANS CLUSTERING

Al Isra Denk Rimakka[1*], Rezty Amalia Aras[2]
Bisnis Digital, Institut Teknologi dan Bisnis Kalla, Makassar, Indonesia[1]
E-mail address: alisra@kallabs.ac.id[1], reztyamalia@kallabs.ac.id[2]

Received: 15, November, 2023 ***Revised: 04, December, 2023*** ***Accepted: 08, December, 2023***

ABSTRACT

As a large company, Amazon operates an online marketplace with a diverse user base exhibiting varied purchasing habits. This diversity challenges Amazon to provide tailored services and marketing strategies for each user with distinct characteristics. Therefore, this research aims to assist Amazon in segmenting its users based on their characteristics, enabling the implementation of targeted marketing strategies and service provision for each user. The study employs the K-Means Clustering method to segment Amazon platform users based on their purchasing behavior, site feature interactions, and preferences. The research process involves Knowledge Data Discovery (KDD) stages, including data processing, attribute selection, and applying the K-Means Clustering algorithm. The analysis results reveal five distinct user clusters, each with unique characteristics reflecting user behavior and preferences. These clusters depict variations in purchasing frequency, interactions with site features, and responses to product recommendations. This user segmentation provides valuable insights for Amazon to develop more focused marketing strategies, enhance personalized services, and improve overall customer satisfaction.

Keywords: Amazon, K-Means Clustering, User Segmentation, Purchasing Habits.

1. INTRODUCTION

The rapid development of technology from year to year has transformed the habits of society in various activities. People now are familiar with and utilize the internet, which makes it easier for them to access information about products they want to purchase. The traditional practice of direct face-to-face transactions between sellers and buyers, which was once common, has now changed (Adhani et al., 2020). Consumer behavior, influenced by technology, has undergone a significant shift from traditional offline shopping to online shopping on e-commerce platforms. E-commerce is one of the rapidly growing industries in the digital era, serving as a platform where consumers can access comprehensive product information and engage in buying and selling transactions through the use of internet technology.

Amazon is one of the globally renowned e-commerce platforms. The Amazon platform offers a diverse range of products to cater to the needs of its potential consumers. However, being a large company with a substantial consumer base, there are varied purchasing habits among consumers. To ensure customer retention, the company must provide differentiated treatment for each type of consumer. Therefore, consumer clustering is necessary to determine the

strategies that should be employed to retain these consumers (Maulana et al., 2022). Clustering is one of the methods used in customer segmentation. In clustering, the data grouping process is carried out based on the characteristics of the data, which are then placed into several clusters (Murpratiwi et al., 2021). One way to perform clustering is by using data mining methods.

Data mining is one of the data processing methods that encompasses various algorithms such as classification, clustering, association, estimation, and prediction, which then generate patterns for decision-making. One popular algorithm used in clustering is K-Means. The algorithm operates by selecting the number of clusters and determining initial centroids randomly. Each cluster will contain a group of data points close to the centroid of that cluster. This process is carried out iteratively until there are no further changes in the clusters (Akbar et al., 2023). The application of K-Means will be implemented in this research with the aim of segmenting users on the Amazon platform. It is hoped that this study can be utilized to understand the characteristics of purchasing habits on the Amazon platform, enabling the implementation of appropriate marketing strategies to ensure that Amazon consumers continue to make purchases on the platform.

2. THEORY

User segmentation is a process of categorizing consumer data into several groups with similar characteristics, such as age, gender, shopping habits, product preferences, or the level of consumer spending. User segmentation is carried out with the aim of understanding the company's market share. By utilizing user data, companies can implement targeted marketing strategies based on user characteristics, such as improving services, offering promotions, or developing products (Faradila Ilena Putri , Retno Damayanti, 2022).

Purchasing habits refer to consistent patterns or tendencies exhibited by a consumer in the process of buying goods or services. This encompasses a series of steps or stages followed by the consumer, starting from recognizing needs or desires, seeking information, making purchasing decisions, to evaluating the experience after the purchase. Several factors can shape purchasing habits, including cultural factors, social factors, personal factors, and psychological factors (Dr. Vladimir, 2017).

Data Mining is one step in the Knowledge Discovery in Databases (KDD) process. There are several processes involved, including data cleaning, data integration, data selection, data transformation, pattern evaluation, and knowledge presentation. The data mining process framework consists of three stages: data collection, data transformation, and data analysis. Preprocessing is the initial step where data is collected to generate what is known as raw data, which is then transformed by converting the raw data into a format suitable for data mining. The results of this transformation are used by data analysts to create knowledge using techniques such as statistical analysis, machine learning, and information visualization (Firdaus, 2017).

K-Means Clustering is a non-hierarchical method that groups data into one or more clusters. The data within each cluster share similar characteristics, so data with different characteristics will be in a different cluster, resulting in low variability within a cluster. There are several steps involved in performing clustering using the K-Means method:

1. Choose the number of clusters or the value of K.
2. Initialize the k cluster centers; this can be done in various ways, with random initialization being the most common. The cluster centers are given initial values with random numbers.
3. Allocate all data/objects to the nearest cluster. The proximity between two objects is determined based on the distance between them. Similarly, the proximity of a data point to a specific cluster is determined by the distance between the data point and the cluster center. In this phase, the distance of each data point to each cluster center needs to be calculated. The distance between a data point and a specific cluster will determine which cluster the data point belongs to. To calculate the distance of all data points to each cluster center, Euclidean distance theory is used and formulated as follows:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (1)$$

where: D (i,j) = Distance of data point i to cluster center j
X ki = Data point i on attribute k
X kj = Cluster center j on attribute k

4. Recalculate the cluster centers with the current cluster membership. The cluster center is the average of all data points/objects in a specific cluster. If desired, the median (middle value) of that cluster can also be used. Therefore, the average (mean) is not the only measure that can be used.

Reassign each object using the new cluster center. If the cluster centers do not change anymore, then the clustering process is complete. Alternatively, return to step number 3 until the cluster centers no longer change (Eno Ketherin et al., 2018).

3. METHOD

In this research, the dataset was obtained from a collection of user data on the Amazon platform sourced from the Kaggle Repository. The dataset comprises 22 attributes with 602 instances (Ananda & Aras, 2021). The attributes in the dataset include age, gender, Purchase Frequency, Purchase Categories, Personalized Recommendation Frequency, Browsing Frequency, Product Search Method, Search Result Exploration, Customer Reviews Importance, Add to Cart Browsing, Cart Completion Frequency, Cart Abandonment Factors, Save for Later Frequency, Review Left, Review Reliability, Review Helpfulness, Personalized Recommendation Frequency, Recommendation Helpfulness, Rating Accuracy, Shopping Satisfaction, Service Appreciation, and Improvement Areas.

The attributes from the dataset will be processed using one of the data mining algorithms, namely K-Means Clustering. The stages in the data processing can be observed in the following flowchart.

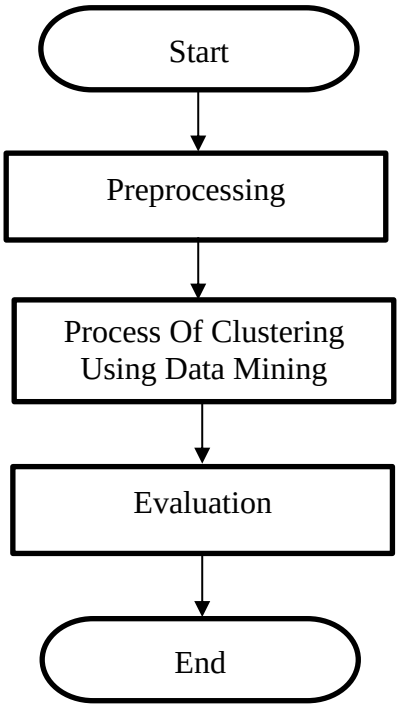

Figure. 1. Flowchart of proposed method

Data processing is carried out in the RapidMiner application, which is software that utilizes data mining algorithms to extract models from large datasets using a combination of statistical methods (Ananda & Aras, 2021). The first stage involves preprocessing. In the given dataset, there are missing values, so data cleaning will be performed using the Replace Missing Value operator to fill in the missing values with the dataset's mean.

After replacing missing values, data transformation will be carried out because there are attributes with nominal types that need to be converted into numeric types using the Nominal to Numeric operator. This transformation is necessary as the K-Means algorithm can only process numeric data. In this stage, attribute selection will be performed since not all attributes are relevant to the research objective. Attribute selection will be done using the Select Attributes operator. The following are the attributes suitable for processing in this research:

a. Purchase_Frequency
b. Browsing_Frequency
c. Customer_Reviews Importance
d. Add_to_Cart_Browsing
e. Cart_Completion_Frequency
f. Saveforlater_Frequency
g. Review_Left
h. Review_Reliability
i. Review_Helpfulness
j. Recommendation_Helpfulness
k. Rating_Accuracy
l. Shopping_Satisfaction

The above attributes focus on purchasing frequency, interactions with site or application features, as well as user satisfaction and trust in reviews and recommendations. By utilizing

these attributes, K-Means Clustering will aid in grouping users into segments based on their purchasing behavior and preferences on the Amazon platform. In this stage, the clustering process will be performed using the K-Means operator. The researcher will specify three clusters (K=3) with centroid values randomly selected.

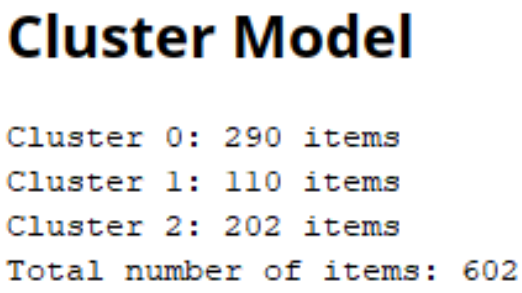

Figure. 2. Cluster Dataset

In the above image, it can be observed that several clusters have been formed, accommodating different amounts of data. There are three clusters formed according to the specified K during the clustering process. The largest amount of data is in cluster 0 with 290 data points, and the smallest is in cluster 1 with 110 data points. The remaining data is in cluster 2, totaling 202 data points, with a grand total of 602 data points.

After clustering with the formation of clusters based on the specified K, it cannot be immediately deemed as good clustering. The clustering results can be evaluated to determine the optimal number of clusters. One evaluation method for K-Means Clustering is the Davies Bouldin method. The determination of the number of K in this method can be seen from the smallest Davies Bouldin Index.

Table 1. The determination of the number of K.

DBI K	Column A (t)
3	-2.636
4	-2.520
5	-2.502

From the research results, the cluster with the smallest Davies Bouldin Index (DBI) has a value of 5. This indicates that the suitable number of clusters for segmentation in this study is 5 clusters.

4. RESULTS AND DISCUSSION

From the analysis using K-Means Clustering on Amazon platform user data, 5 clusters (cluster 0 to cluster 4) have been formed, each with different characteristics. The following are conclusions drawn from each cluster based on the dominant attributes in each cluster:

Cluster 0:

a. This cluster contains users who tend not to add products to the cart while browsing on Amazon (Add_to_cart = No).
b. Users in this cluster are less likely to provide reviews about products (Review_reliability = Rarely).
c. However, if users in this cluster add products to the cart, they are more likely to complete the purchase (Add_to_cart_browsing = Yes).

Cluster 1:
a. This cluster has users who tend not to add products to the cart while browsing on Amazon (Add_to_cart_browsing = No).
b. Users in this cluster are less likely to provide reviews about products (Review_reliability = Rarely).
c. However, if users in this cluster find reviews from other customers helpful, they are more likely to provide reviews as well (Review_helpfulness = Sometimes is on Average 143.14% larger).

Cluster 2:
a. This cluster has a habit of always saving products in the "Save for Later" feature on Amazon (Saveforlater_frequency = Always is on Average 1,178.37% larger).
b. Users in this cluster tend not to provide reviews about products (Review_reliability = Never is on Average 352.57% smaller).
c. Also, they rarely provide reviews about products (Review_reliability = Rarely is on Average 251.20% smaller).

Cluster 3:
a. This cluster tends to frequently receive personalized product recommendations from Amazon (Personalized_recommendation_frequency = Yes is on average 159.75% larger).
b. Users in this cluster tend to provide reviews about products less frequently (Review_reliability = Rarely is on average 135.64% larger).
c. However, they tend to find these recommendations helpful Recommendation_helpfulness = Yes is on average 126.25% larger).

Cluster 4:
a. This cluster has a habit of making purchases on Amazon multiple times a week (Purchase_frequency = Multiple times a week is on average 546.50% larger).
b. Users in this cluster are less likely to complete purchases after adding products to the cart (Cart_completion_frequency = Never is on average 542.53% larger).
c. Also, they rarely engage in browsing on Amazon (Browsing_frequency = Rarely is on average 506.58% larger).

5. CONCLUSIONS AND SUGGESTIONS

Overall, this research has a strong focus on customer segmentation on the Amazon platform through the implementation of the K-Means Clustering method. This approach helps unravel diverse user behavior and preferences into more focused and understandable groups. Through the Knowledge Data Discovery (KDD) steps, including data preprocessing, attribute selection, and the application of the K-Means algorithm, this research yields valuable insights.

The analysis results in 5 user groups with significantly different characteristics. The first cluster represents users who are reluctant to add products to the cart while exploring the site, but most of them will complete the purchase if a product is added. The second cluster includes users with a similar pattern, with an emphasis on product reviews. The third cluster reflects users who frequently use the "Save for Later" feature and generally do not provide reviews. The fourth cluster describes a group more reliant on personalized product recommendations, while the fifth cluster consists of active purchasers but less likely to complete purchases after adding products to the cart.

The evaluation results of the Davies Bouldin Index confirm that the optimal number of clusters for this research is 5. Overall, this research provides an in-depth understanding of Amazon user behavior and preferences, with significant strategic implications. Amazon can use this insight to refine marketing strategies, enhance personalized services, and ensure that the customer experience is continuously improved.

6. ACKNOWLEDGEMENTS

The authors would like to thank kaggle machine learning for preparing a public dataset for researchers to use and the authors would also like to thank the members of the Intelligent Systems research group for inspiring discussions and sharing meaningful knowledge.

REFERENCES

Adhani, L. K., Dharmastiti, R., & Trapsilawati, F. (2020). Pengaruh Waktu Sebelum Dan Selama Pandemi Covid-19 Terhadap Perilaku Konsumen Belanja Online. *Perspektif Keilmuan Teknik Industri Pada Era New Normal*, 50–55.

Akbar, M. N., Azizah Salsabila, Aldi Perdana Asri, & Muhammad Syawir. (2023). Analisis Clustering Untuk Segmentasi Pengguna Kartu Kredit Dengan Menggunakan Algoritma K-Means Dan Principal Component Analysis. *AGENTS: Journal of Artificial Intelligence and Data Science*, *3*(1), 16–24. https://doi.org/10.24252/jagti.v3i1.56

Ananda, N., & Aras, R. A. (2021). Clustering Pengeluaran Tahunan Berbagai Macam Produk Menggunakan Metode K-Means. *Seminar Nasional Sains Dan Teknologi Informasi SENSASI 2021*, 143–147.

Dr. Vladimir, V. F. (2017). Landasan Keputusan Manajemen. *Gastronomía Ecuatoriana y Turismo Local.*, *1*(69), 5–24.

Eno Ketherin, B., Anjani Arifiyanti, A., & Sodik, A. (2018). Analisa Segmentasi Konsumen Menggunakan Algoritma K-Means Clustering. *Sains Dan Teknologi Terapan*, 51–58.

Faradila Ilena Putri , Retno Damayanti, K. (2022). *Penerapan Algoritma K-Means*. *2*(mei 2022), 408–418.

Firdaus, D. (2017). Penggunaan Data mining dalam kegiatan pembelajaran. *Jurnal Format Volume 6 Nomor 2 Tahun 2017*, *6*(2), 91–97.

Franklin, S. W., & Rajan, S. E. (2014). Computerized screening of diabetic retinopathy employing blood vessel segmentation in retinal images. *Biocybernetics and Biomedical Engineering*, *34*(2), 117–124. https://doi.org/10.1016/j.bbe.2014.01.004

Maulana, E. B., Hadiana, A. I., & Umbara, F. R. (2022). Segmentasi Pengunjung Pusat Perbelanjaan Menggunakan Metode K-Means Clustering. *Seminar Nasional Sistem Informasi (Senasif) 2022*, *6*(1), 3093–3102.

Murpratiwi, S. I., Agung Indrawan, I. G., & Aranta, A. (2021). Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *18*(2), 152. https://doi.org/10.23887/jptk-undiksha.v18i2.37426