IRJEMS International Research Journal of Economics and Management Studies
Published by Eternal Scientific Publications
ISSN: 2583 – 5238 / Volume 3 Issue 4 April 2024 / Pg. No: 71-77
Paper Id: IRJEMS-V3I4P112, Doi: 10.56472/25835238/IRJEMS-V3I4P112

Original Article

Clustering E-Commerce Consumers through Machine Learning-Based Analysis of Clickstream Data

[1]**Mohammad Al Masalma**
[1]*Department of Electrical and Computer Engineering, King Abdilaziz University, Jeddah, Saudi Arabia.*

Received Date: 06 March 2024 Revised Date: 18 March 2024 Accepted Date: 25 March 2024 Published Date: 13 April 2024

Abstract: *In recent years, the e-commerce sector has significantly grown, with selling strategies becoming increasingly broad and complicated. Businesses aiming to enhance customer experience, optimize marketing efforts, and boost revenue growth must comprehend user behavior. In this research, we examine how well clustering algorithms work to identify significant trends in clickstream data from e-commerce. In order to conclude useful information that can aid in strategic decision-making in a business, we intend to aggregate and analyze user activities such as page visits, product views, and transactions. K-means is a well-known and mostly used clustering algorithm that is used because of its capacity to group data samples according to in-common behavioral patterns. Preprocessed clickstream data, which includes various parameters, including the quantity of clicks, average price viewed, and most popular product, color, location, and model photography, is subjected to each algorithm. We evaluate these algorithms' performance in terms of cluster quality, interpretability, and significance for e-commerce analytics through extensive testing and comparative research. This paper's findings identify various user clusters with different product preferences, browsing behaviors, and levels of involvement.*

Keywords: *Clickstream Data, Clustering Algorithms, Data analysis, Data mining, K-means, Time series Data.*

I. INTRODUCTION

The development of e-commerce has changed dramatically, altering conventional business structures and the methods in which goods and services are purchased and sold. Businesses must modify their tactics to succeed in this dynamic and cutthroat market as more and more customers resort to online platforms for their purchasing requirements. Businesses looking to increase their market share and obtain a competitive edge must comprehend the complex dynamics of e-commerce transactions and consumer behavior. Within this context, our study aims to investigate the complex domain of e-commerce by exploring the vast amount of data produced by online transactions and customer interactions. Our goal is to use cutting-edge data analytics approaches to find hidden patterns, trends, and insights that can help organizations navigate the challenges of the digital marketplace. Our study is centered on the analysis of large-scale e-commerce datasets using advanced machine learning algorithms and statistical techniques. We aim to extract meaningful information from raw data and turn it into usable insights using methods like Principal Component Analysis (PCA) and k-means clustering.

The objective of our research is to equip businesses with the necessary knowledge to make informed decisions and customize their marketing strategies to suit the changing needs and preferences of their target audience by identifying discrete clusters of consumer behavior, segmenting markets, and forecasting future trends.

Furthermore, our study involves practical applications and real-world development in addition to the field of data analysis. Through collaboration with industry participants and case studies, our objective is to verify the efficacy of our analytical methodologies and exhibit their concrete influence on organizational performance.

II. LITERATURE REVIEW

[1]The literature on consumer behavior analysis in the context of fast-fashion retail, particularly in the U.K., emphasizes the significance of clickstream data analysis for understanding online consumer engagement. Leveraging Adobe Analytics, this study utilizes a substantial dataset capturing a representative sample of website visits, enabling the construction of a comprehensive dataset encompassing entry, browsing, exit patterns, and revenue metrics. Past research has highlighted the importance of such measures in delineating consumer segments and informing marketing strategies. Employing cluster analysis techniques, particularly the Partitioning Around Medoids (PAM) algorithm, the study aims to identify distinct consumer segments robustly while addressing computational challenges inherent in big data analytics. Furthermore, the methodology incorporates statistical tests like the Kruskal-Wallis and Dunn tests to assess revenue differentials among clusters, offering a rigorous analytical framework for deriving actionable insights. So, this literature underscores the critical role of data-driven approaches in elucidating consumer behavior patterns and optimizing business performance in the fast-fashion retail sector.

The dataset utilized in this study[2] originates from the Kaggle data repository, encompassing clickstream data from an online retailer specializing in apparel for pregnant women. Spanning five months of records from 2008, the dataset includes various attributes such as product categories, website photo placements, IP address countries of session origins, and product prices, among others (Śapczyński M., Białowąs S., 2013). Comprising 165,474 rows and 14 columns, the dataset adheres to pertinent data protection regulations, ensuring compliance and data integrity. Leveraging this rich dataset, the study employs the K-means algorithm, a widely used unsupervised machine learning approach, to cluster the data into distinct groups. K-means iteratively partitions the unlabelled dataset into K clusters based on centroid computations, with the optimal K often determined using methods like the Elbow Method. This iterative process *allows* for the identification of underlying patterns and structures within the data, facilitating further analysis and insights into consumer behavior and preferences; the proposed method uses K-means after finding the most relevant features, they filter all features that do not affect the clustering process and ended with the only single feature which is Category(main page). The clustering was conducted depending on this feature. **Figure 1** shows this study's final clustering.

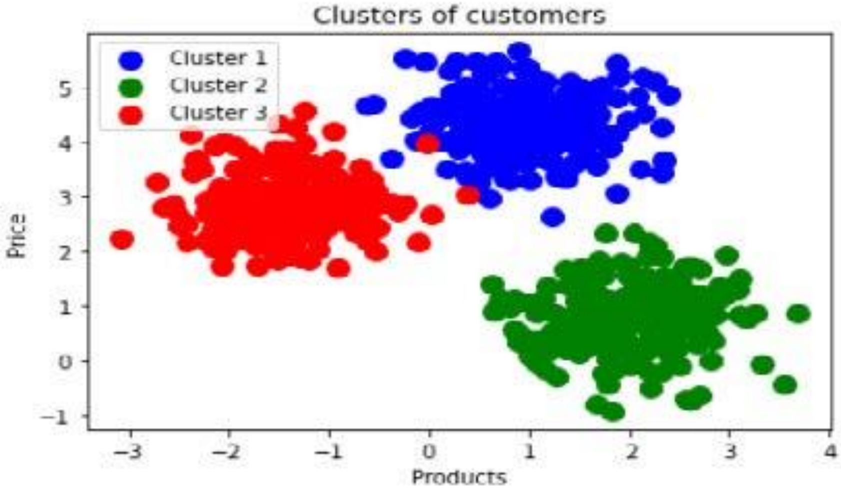

Figure 1: Clustering Result

In this case study [3], clickstream data collected from one of China's largest e-commerce platforms, akin to Amazon.com, provides a rich dataset for systematic analysis. The platform caters to a diverse range of consumer needs, offering products ranging from household necessities to electronics, apparel, and cosmetics. With approximately eight percent of China's online shoppers using this platform and average daily traffic of three million users, the dataset presents a comprehensive snapshot of online consumer behavior. Data collection focused solely on clickstream data from PC users, recording requested URLs and timestamp requests. Each user session was assigned a unique session ID to organize records effectively. Despite lacking detailed page content, the dataset's size and scope offer ample opportunities for analysis. Preprocessing efforts aimed to enhance data quality and relevance, including user and category identification, time estimation, and data cleaning. These preprocessing steps ensured that the dataset was well-prepared for subsequent analysis. Additionally, parameter analysis was conducted to calibrate algorithm parameters, such as the general and rough thresholds, to optimize clustering results. Leveraging optimized parameters, the rough leader clustering algorithm classified users into distinct clusters, revealing insightful interest patterns. Visual representations of these patterns highlighted dominant user interests, ranging from women's apparel to food and household necessities and electronics. These findings offer valuable insights into user preferences and behavior, with implications for product offerings and customer experience enhancements within the e-commerce platform.

III. METHODOLOGY

The methodology used in this research can be divided into three main stages; in the first stage, the data is cleaned and aggregated then the data is fed into our algorithms, the results are evaluated, and then a comprehensive data analysis is conducted.

A) Dataset Description:

The dataset "e-shop clothing 2008" encompasses a comprehensive record of user interactions and behaviors on an e-commerce platform. It comprises various attributes capturing user engagement and transactional patterns during browsing and purchasing activities. The dataset includes the following variables:

1. YEAR: The year of the recorded data (2008).
2. MONTH: Ranging from April (4) to August (8), indicating the month of the session.
3. DAY: The day number of the month.

4. ORDER: Sequence of clicks during one session.
5. COUNTRY: Country of origin of the IP address, categorized into different countries and domains.
6. SESSION ID: A unique identifier for each session.
7. PAGE 1 (MAIN CATEGORY): Main product categories, such as trousers, skirts, blouses, or sale items.
8. PAGE 2 (CLOTHING MODEL): Information about the specific product code.
9. COLOUR: Color of the product is categorized into various colors.
10. LOCATION: Photo location on the page, divided into six parts.
11. MODEL PHOTOGRAPHY: Frontal or side view of the product photography.
12. PRICE: Price of the product in US dollars.
13. PRICE 2: Indicates whether the price of a product is higher than the average price for the entire product category.
14. PAGE: Page number within the e-store website.

The dataset provides valuable insights into consumer buying behaviors and preferences across different product categories, colors, and pricing ranges. It serves as a rich resource for analyzing user behavior dynamics and identifying trends in e-commerce activities. If used, please cite the original publication by Łapczyński and Białowąs (2013).

B) *Dataset Aggregation:*

In order to extract meaningful insights from the dataset and facilitate analysis, we performed data aggregation. Each session in the original dataset was represented by multiple rows, capturing various interactions and attributes. However, this granular level of detail was not conducive to our analysis objectives, as we aimed to analyze each session as a single entity.

To address this, we aggregated the data by consolidating all the rows associated with each session into a single record. This transformation allowed us to treat each session as a coherent unit, thereby simplifying subsequent analysis and interpretation.

During the aggregation process, we applied specific aggregation functions to different attributes. For categorical variables such as “PAGE 1 (MAIN CATEGORY),” “COLOUR,” “LOCATION,” and “MODEL PHOTOGRAPHY,” we retained the most frequent value within each session. This approach helped capture the dominant characteristics of each session in terms of product category, color preference, photo location, and model photography style.

For numerical variables like “PRICE,” we calculated the average price within each session. This aggregation method enabled us to summarize the pricing information associated with each session, providing a representative value that reflects the overall price level observed during the session.

By aggregating the dataset in this manner, we transformed the raw data into a more structured and analytically manageable format, laying the foundation for our subsequent analysis of user behavior and purchasing patterns on the e-commerce platform.

C) *Principle Component Analysis (PCA)*

Instead of working with a bunch of complex features, PCA finds a smaller set of simpler directions that capture most of the important variations in the data. It does this by identifying the directions where the data points are most spread out and then using those directions to represent the data in a more manageable way[4].

PCA was pivotal in preparing the dataset for K-means clustering. It served a dual purpose: dimensionality reduction and feature transformation. Standardization was initially applied to ensure consistent scaling across features. Subsequently, PCA was performed on the standardized data, focusing on retaining two principal components that explained the majority of the variance. This deliberate reduction to two principal components facilitated computational efficiency and enhanced interpretability. By summarizing the original feature space into two uncorrelated variables, PCA effectively captured the essential characteristics of the data while reducing its dimensionality. This streamlined representation enabled a more efficient and insightful application of the subsequent K-means clustering algorithm.

a. K-means: Imagine you have a pile of unsorted objects. K-means clustering helps you organize them into groups (k) based on how similar they are. It does this by putting objects close together if they share features and separating objects that are very different[5].
Choosing the right number of clusters (k) in K-means clustering can be tricky. The elbow method helps with this by creating a graph. **Fig .2** shows how much better the clustering gets (measured by WCSS) as you increase the number of clusters. The ideal number of clusters (k) is the ‘elbow’ of the curve, the point where adding more clusters does not significantly improve the grouping[6]
In determining the optimal number of clusters for our analysis, we applied the elbow method, a popular technique used to identify the appropriate number of clusters in K-means clustering. By plotting the within-cluster sum of squares (WCSS) against the number of clusters, we observed a distinct “elbow” point where the rate of decrease in WCSS slows down

significantly. This elbow point indicates the optimal number of clusters beyond which the marginal gain in clustering quality diminishes. Upon applying the elbow method to our dataset, we identified the elbow point at three clusters, suggesting that partitioning the data into three distinct clusters would adequately capture the underlying structure and variability in the dataset. This informed our decision to proceed with three clusters in our subsequent analysis. **Fig. 2** shows the graph of the Elbow method and indicates that the optical cluster number is 3.

K-means clustering played a crucial role in our analysis approach, especially after reducing the dataset's dimensions through PCA. This clustering technique partitions the data into a set number of clusters by minimizing the variance within each cluster. Through iterative updates of cluster centroids, K-means effectively identifies distinct clusters within the dataset. We determined the number of clusters either based on prior knowledge or using methods like the elbow method. The resulting clusters provided valuable insights into underlying patterns and groupings within the data, aiding further analysis and interpretation. Moreover, K-means facilitated the assignment of data points to specific clusters, enabling segmentation based on cluster membership for subsequent analysis. In the suggested methodology, the features are scaled, then they are fed into the K-means algorithm, and then they are clustered into 3 clusters. **Fig. 3** represents our clusters after they are projected into Principal Components.

b. Data analysis and Interpreting Clusters: Analyzing data is similar to sorting through a sand pile. Even with all of this specific information (data points), it might be challenging to perceive the wider picture. Using a sieve is similar to aggregation. It facilitates the grouping of related data points, which makes it simpler to identify patterns and trends. In this manner, mounds of data can be transformed into understandable and useful insights.[7]
As we have multidimensional data, we cannot visualize the data in a meaningful way in order to understand what is behind our clusters and what attributes contributed to clustering our data points in the way they are.
For this reason, a new column is added in order to assign each data point to its corresponding cluster, and this is to help in identifying the similarities in the sessions and to find the common attributes that made the algorithm cluster the data points in the way they are clustered.

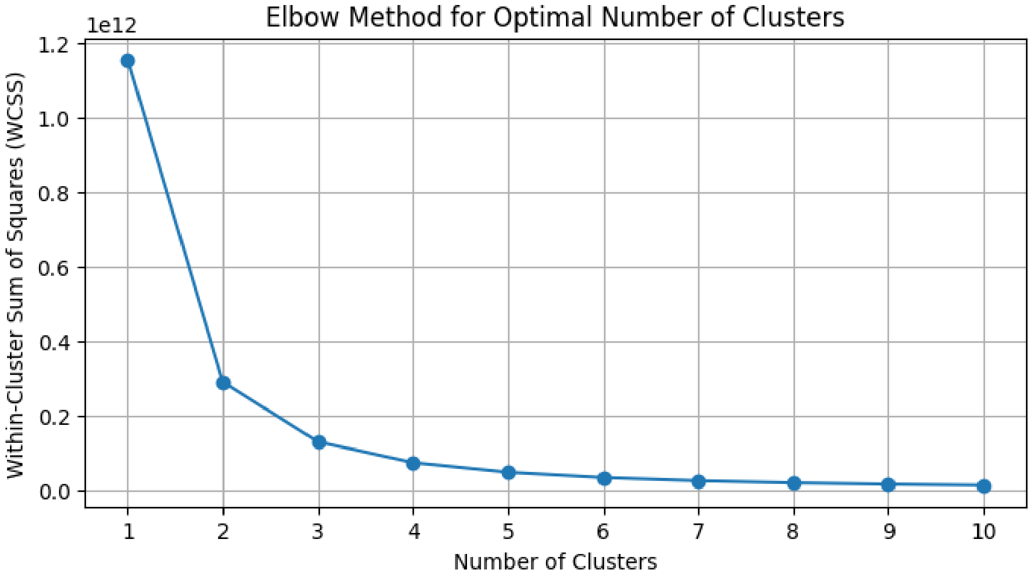

Fig. 2 Elbow method implementation result

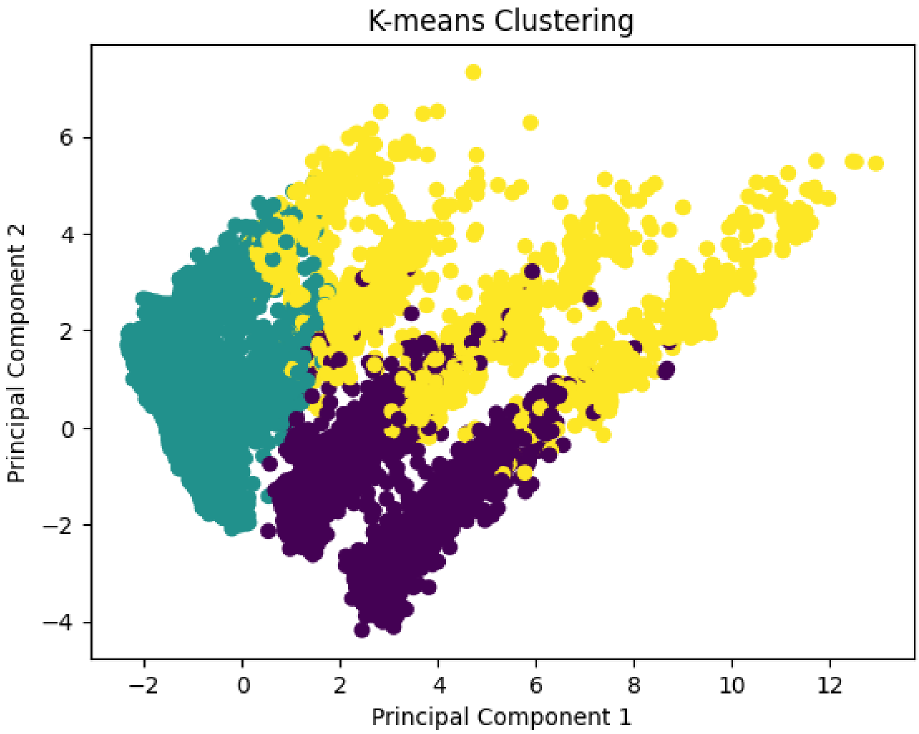

Fig. 3 K-means clustering result

IV. RESULTS AND DISCUSSION

K-means has clustered our data points into three clusters, and this has already been identified using the elbow method, as mentioned previously.

Two main features, namely location and Model photography, were analyzed due to the importance and the contrition they made during the clustering process.

As can be seen in **Fig. 4** the Y-axis indicates the number of clusters, and the X-axis represents the percentage of sessions that are concentrated on either en-face or profile images of the products since 1 indicates en-face image and 2 indicates profile image.

Cluster 0 shows that 100% of sessions are interested in the products that have end face images; on the contrary, cluster 1 shows completely the inverse, while the third cluster shows that around 4 % of the sessions checked the profile images of products and 96 % checked the en-face images.

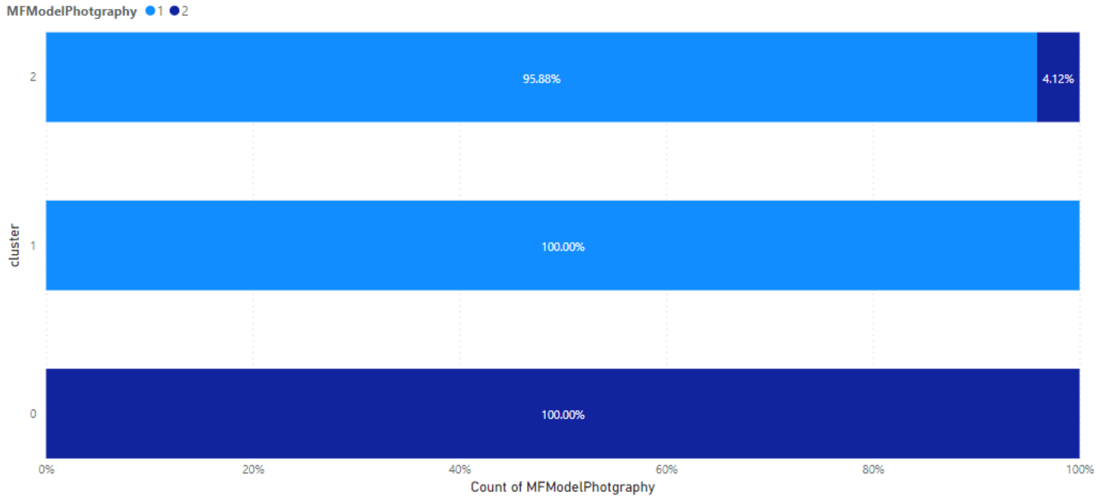

Fig. 4 Most Frequent Model Photography by cluster

In **Fig. 5,** the X-axis shows the percentage distribution of sessions over 6 different locations of the product on the web page

1: top left
2: top in the middle
3: top right
4: bottom left
5: bottom in the middle
6: bottom right

cluster 0 shows that around 31% of sessions are interested in the products that are placed on the middle bottom. Also, about 19 % are top left; in cluster 1, we can notice that most sessions are approximately distributed over different locations of the page; in cluster 2 there is a focus on the products that are located on the top left and top in the middle also middle in the bottom.

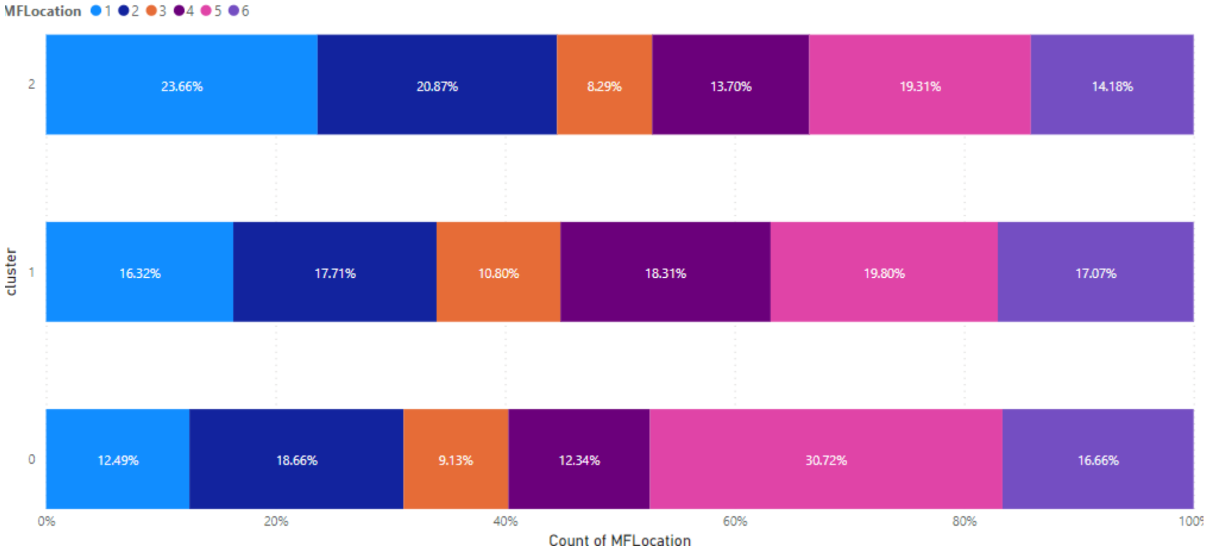

Fig. 5 Most Frequent Model Photography by Cluster

V. CONCLUSION

All things considered, our study has painstakingly made its way through three crucial phases: dataset definition, aggregation, and sophisticated analysis methods such as PCA and K-means clustering. With the use of the extensive "e-shop clothing 2008" dataset, we conducted a thorough investigation into user interactions and behaviors in the context of e-commerce.

We reduced the size of the dataset by carefully combining the data, which allowed for a logical examination of the user session. We reduced the dataset's dimensionality using PCA while maintaining its essential features, setting the stage for more in-depth understanding. K-means clustering then identified different user groups, providing insight into a range of customer preferences, especially with regard to homepage positioning and product photography styles. Our study's ability to identify patterns and trends not only improves our comprehension of online consumer behavior but also provides useful information for e-commerce platform optimization, allowing for improved user customization and better overall purchasing experiences.

V. REFERENCES

[1] M. Zavali, E. Lacka, and J. De Smedt, "Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data," *IEEE Trans Eng Manag*, vol. 70, no. 4, pp. 1353–1364, Apr. 2023, doi: 10.1109/TEM.2021.3070069.

[2] J.-Z. Lezu, G. Cassani, and A. Hendrickson, "Analyzing and Clustering Clickstream Data for Marketing Personalization," 2021.

[3] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using clickstream data," *Electron Commer Res Appl*, vol. 14, no. 1, pp. 1–13, Jan. 2015, doi: 10.1016/j.elerap.2014.10.002.

[4] J. Shlens, "A Tutorial on Principal Component Analysis," Apr. 2014, [Online]. Available: http://arxiv.org/abs/1404.1100

[5] E. Umargono, J. E. Suseno, and V. Gunawan, "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula," 2020.

[6] S. Nanjundan, S. Sankaran, C. R. Arjun, and G. P. Anand, "Identifying the number of clusters for K-Means: A hypersphere density based approach."

[7] A. Jain and A. Kumawat, "ANALYSIS OF CLICKSTREAM DATA," *International Research Journal of Engineering and Technology*, 2022, [Online]. Available: www.irjet.net