The Implementation of Fuzzy C-Means Algorithm to Analyze Based on Price and Sold Variables in Tokopedia

Yuda Perwira[1] and Viddi Mardiansyah[2]

[1,2] Widyatama University, Bandung, Indonesia
viddi.mardiansyah@widyatama.ac.id

Abstract. Various studies suggest a correlation how much of the product has been sold and product price to make decisions about sales strategy. Tokopedia, an e-commerce corporation, maintains a sales database encompassing consumer behavior data. This data can be a valuable resource for vendors who wish to develop more effective sales strategies based on sales. Fuzzy C-Means (FCM) algorithms is frequently employed techniques to analyze with overlapping data such as dataset obtained from Tokopedia. The FCM method was implemented to analyze datasets from Tokopedia which is useful for making decisions related to sales strategies. FCM algorithms are regarded as good method in analyzing based on price and sold variables with evaluation score by Davies Bouldin Index is 0.8253028798553185.

Keywords: Fuzzy C-Means, Clustering, Analyze.

1 Introduction

The e-commerce sector in Indonesia is experiencing fast growth, projected to reach a turnover of IDR 545.7 trillion in 2022. Conducting a comprehensive market study is essential for effective business expansion in order to harness this potential. Tokopedia, a prominent e-commerce platform in Asia, boasts a user base of over 100 million and must discern its market potential in order to develop effective marketing strategies and improve customer satisfaction. In e-commerce market potential analysis, the fuzzy C-Means technique, which permits data points to be assigned to several clusters with varying degrees of membership, surpasses other approaches. The objective of this study is to offer a more thorough assessment of the commercial potential of Tokopedia [1][2][3].

V. Mardiansyah and B. A. Prasetyo (eds.), *Proceedings of the Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, Advances in Engineering Research 252,
https://doi.org/10.2991/978-94-6463-618-5_33

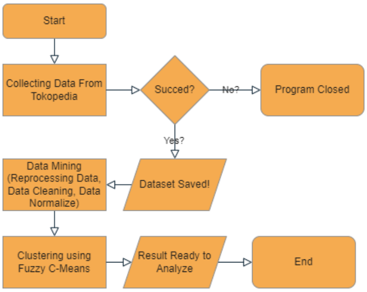
Start
Collecting Data From Tokopedia
Succed?
No?
Program Closed
Yes?
Dataset Saved!
Data Mining (Reprocessing Data, Data Cleaning, Data Normalize)
Clustering using Fuzzy C-Means
Result Ready to Analyze
End

Fig 1. Flowchart of Methods

2.1 Collecting Research Data

This study employs data scraping techniques to gather information from the Tokopedia website using Google Chrome. The data search was conducted on July 9, 2024, using the keyword “handphone”. The search results yielded approximately 500,000 pages. Because of the constraints of the data processing tools, this study has only considered data from the initial 50 pages of the search results. The study focuses on various aspects, including the store name, location, product name, price, quantity of products sold, and buyer ratings. For data extraction, the researchers thoroughly examined the HTML element found on the Tokopedia web page. They aimed to pinpoint the specific class element containing the relevant information. The class element that has been identified is then extracted and implemented in code using the Visual Studio Code.

Initialization and Preconfiguration. Selenium Currently, the researchers have created the source code to establish and customize data scraping. The source code incorporates a time sleep system to regulate the time intervals between operations, automates the interface with web browsers, and activates specific parts to pause the script execution until certain criteria are fulfilled. This is done to mitigate potential timing issues and to assure the stability of the automated scripts in the data scraping process. Subsequently, automated scripts were created using Selenium WebDriver. The script is structured into two primary portions: data search and retrieval [6].

Data Retrieval. The subsequent step involves processing HTML web pages and extracting data related to products and research objects. The researchers employed the BeautifulSoup library to accomplish this objective. HTML components that contain data requiring identification and extraction [7]. By utilizing the Pandas library, the data is stored in a data frame data structure and subsequently saved as a CSV file named “dataset”. The CSV file is displayed in Table 1. The data scraping process yielded almost 4,000 data points. However, Table 1 in this study only presents the first 4 data points because of page constraints.

Table 1. Data Scraping Results

Num.	Store Name	Location	Products Name	Price (Rp)	Products Sold (Quantity)	Buyer Ratings
1	iSmile Official Store	Central Jakarta	Apple iPhone 15 Pro Max Garansi Resmi - 256GB…	21.049.000	1.000+	5.0
2	iSmile Official Store	Central Jakarta	Apple iPhone 15 Garansi Resmi - 128GB 256GB 512GB	13.099.000	2.000+	5.0
3	vivo Indonesia	Tangerang Regency	vivo Y22 (4/128) - Helio G85, 50MP Camera, Spl…	1.367.000	2.000+	4.9
4	Digitech Mall	Central Jakarta	Redmi Note 13 5G 8/256 GB 8GB 256GB Garansi …	2.250.000	500+	4.9

2.2 Data Mining

Researchers perform data mining through three stages to organize the data, making data analysis easier. Initially, researchers identify missing values and duplicate data for exclusion. The next step involves converting object data (alphanumeric) into integers (numeric). This procedure occurs because the clustering process with FCM only accepts integer data types [8]. During the last step, the data is normalized to have a consistent structure and format. The normalized data is shown in Table 2.

Tabel 2. Normalization Data Results

.	Store Name	Location	Products Name	Price	Products Sold	Rating
0	0.648688	0.195122	0.523664	0.897343	0.270833	0.923077
1	0.648688	0.195122	0.443650	0.280797	0.604167	0.923077
2	0.989796	0.682927	0.969356	0.168478	0.541667	0.923077
3	0.648688	0.195122	0.428328	0.254831	0.270833	1.000000
4	0.989796	0.682927	0.972761	0.710145	0.750000	0.923077

2.3 Clustering

At this stage, the researcher is developing the source code to extract the columns that will be processed using the FCM algorithm. The objective of using FCM is to minimize the defined objective function as follows [4]:

1. Provide the data to be clustered X as a matrix of dimensions n x m (n = number of data samples, m = specific characteristics of each data point). Given the i-th sample data (i=1, 2, …, n) and the j-th attribute (j=1, 2, ..., m), Xij is defined.

$$x = \begin{bmatrix} x11 & \cdots & x1m \\ \vdots & \ddots & \vdots \\ xn1 & \cdots & xnm \end{bmatrix}$$

2. Determine the values of:
   a. Number of clusters = c;
   b. Degree of fuzziness = w;
   c. Maximum iterations = MaxIter;
   d. Minimum expected error = ε;
   e. Initial objective function = P₀ = 0;
   f. Initial iteration = t = 1;

3. Produce random numbers µik for i = 1, 2,..., c, to be used as elements in the starting matrix U. Compute the total of each column that represents an attribute:

$$Qi = \sum_{k=1}^{c} \mu\text{ik}$$

   With i=1, 2, ..., n. Calculate:

$$\mu\text{ik} = \frac{\mu\text{ik}}{Qi}$$

4. With the k-th cluster center: V_{kj}, where k = 1, 2, ..., c; and j = 1, 2, ..., m.

$$V_{\text{kj}} = \frac{\sum_{i=1}^{n}((\mu_{\text{ik}})^{w} * X_{\text{ij}}}{\sum_{i=1}^{n}(\mu_{\text{ik}})^{w}}$$

5. Calculate the objective function at the k-th iteration, P_t.

$$P_{\text{t}} = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(X_{\text{ij}} - \text{V}_{\text{kj}})\right]^{2} (\mu_{\text{ik}})^{w}\right)$$

6. Calculate the changes in the partition matrix:

$$\mu_{\text{ik}} = \frac{\left[\sum_{j=1}^{m}(X_{\text{ij}} - \text{V}_{\text{kj}})\right]^{\frac{-2}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{\text{ij}} - \text{V}_{\text{kj}})\right]^{\frac{-2}{w-1}}}$$

Check the stopping condition: if: (| P_t - P_{t-1} | < ε) or (t > MaxIter) then stop; if not: t = t + 1, repeat step 4.

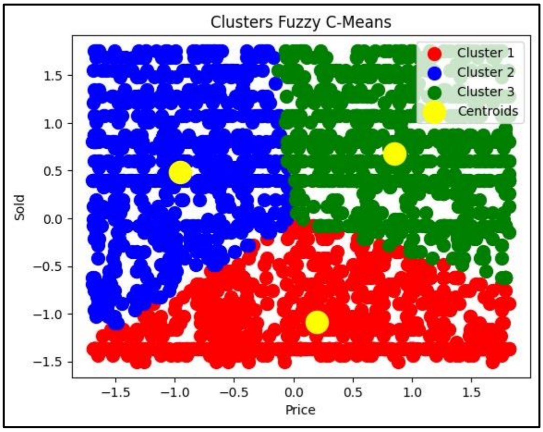

Fig. 1. Fuzzy C-Means Clustering Visualization

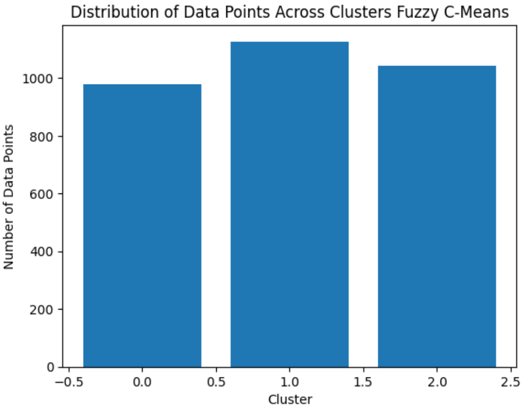

Fig. 2. Fuzzy C-Means Clustering Bar Chart

Following the clustering process, the researchers stored the fuzzy matrix, which is composed of a row that represents a single data point within a data set, and a column that represents a cluster. The matrix values indicate the level of eligibility for each data point within each cluster. As the rating increases, so does the eligibility. Then, the researchers utilize the source code to initialize, define, and label each cluster. Subsequently, these cluster labels are employed to categorize data groupings, as depicted in Figure 1. Afterwards, the researchers augmented the source code to compute the quantity of data in each cluster.

3 Result and Discussion

3.1 Analyze of Clusters

The results presented in figure 2 demonstrate that the data can be classified into three clusters: cluster 1, cluster 2, and cluster 3. Cluster 1 is distinguished by its products having comparatively lower nominal pricing in comparison to the other clusters, resulting in low sales levels. Cluster 2 exhibits product attributes that are situated within a more elevated price bracket in comparison to Cluster 1, while also achieving greater sales levels than both Cluster 1 and 3. Conversely, Cluster 3 distinguishes itself by offering products at far higher prices than the other clusters, but with comparatively lower sales volumes, although not below those of Cluster 1.

4 Conclusion

From the procedures and outcomes of the examined tests in this work, it can be inferred that the Fuzzy C-Means approach is suitable for analyzing the correlation between pricing variables and sales variables. In conclusion, the performance results obtained utilizing Fuzzy C-Means in this work are satisfactory, if not perfect. The reason for this situation is the outcomes derived from the assessment of clustering validity using the Davies Bouldin Index (DBI), which is 0.8253028798553185. The obtained value is more nearly equal to 1 than to 0. One may consider this outcome to be somewhat satisfactory as it does not attain the value of 1.

This research is confined to a database source and a single type of scraping method: Tokopedia. Future research will delve into various data sources and alternative scraping techniques. Its goal is to enhance and fortify the analysis of abnormal data.

References

1. Statista, “E-commerce in Indonesia - Statistics & Facts,” 2023. [Online]. Available: https://www.statista.com/statistics/1170463/indonesia-e-commerce-market-size/. [Accessed: 17 Jul 2024].

2. A. Wardhana, M. Pradana, H. Shabira, D. M. A. Buana, D. W. Nugraha, and K. Sandi, “The Influence of Consumer Behavior on Purchasing Decision Process of Tokopedia E-Commerce Customers in Indonesia,” IEOM Society International, 2021. [Online]. Available: https://ieomsociety.org/singapore2021/papers/998.pdf. [Accessed: 17 Jul 2024].

3. S. S. Prasetyo, M. Mustafid, and A. R. Hakim, “Penerapan fuzzy c-means kluster untuk segmentasi pelanggan e-commerce dengan metode recency frequency monetary (RFM),” *Jurnal Gaussian*, vol. 9, no. 4, pp. 421-433, Dec. 2020

4. F. H. Setiawan, “Penerapan Fuzzy C-Means dan Apriori untuk Rekomendasi Promosi Produk Berdasarkan Segmentasi Konsumen.” Semarang, Indonesia, 2011.

5. Y. Li, J. Qi, X. Chu, and W. Mu, “Customer Segmentation Using K-Means Clustering and the Hybrid Particle Swarm Optimization Algorithm,” The Computer Journal, vol. 66, no. 4, pp. 941-962, Apr. 2023. doi: 10.1093/comjnl/bxab206.

6. B. García, M. Munoz-Organero, C. Alario-Hoyos, and C. D. Kloos, “Automated driver management for Selenium WebDriver,” Empirical Software Engineering, vol. 26, no. 5, p. 107, 2021. doi: 10.1007/s10664-021-09975-3.

7. Polidoro, F., Giannini, R., Conte, R. L., Mosca, S., & Rossetti, F. (2015). “Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation.” Statistical Journal of the IAOS, 31(2), 165-176.

8. Google Developers, “Machine Learning: What is Clustering?” 2022. [Online]. Available: https://developers.google.com/machine-learning/clustering/overview. [Accessed: 01-Jul-2024].