Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 5, No. 3, pp. 280-294, 2024

Comparative Analysis of K-Means and K-Medoids Clustering in Retail Store Product Grouping

Sekar Ghaida Muthmainah[1*], Asep Id Hadiana[2], Melina[3]

[1,2,3]Department of Informatics, Faculty of Science and Informatics, Universitas Jenderal Achmad Yani, Jl. Terusan Jend. Sudirman, Cimahi, West Java, 40525, Indonesia

**Corresponding author email: sekar.ghaida@student.unjani.ac.id*

Abstract

The retail business is growing very rapidly with increasing business competition. The application of information technology is one strategy for understanding consumer product purchasing patterns and grouping sales products. This research aims to analyze and compare the K-Means and K-Medoids Clustering techniques for retail data based on the Davies Bouldin Index value and computing time. K-Means is an algorithm that divides data into k clusters based on centroids, while K-Medoids Clustering uses objects with medoids representing clusters as centroid centers. Clustering in both methods produces an optimal number of clusters of 3 clusters. The results of this research show that K-Means produced 358 data in Cluster 1, 292 data in Cluster 2, and 367 data in Cluster 3 with a DBI of 0.7160. Meanwhile, K-Medoids produced 295 data in Cluster 1, 360 data in Cluster 2, and 362 data in Cluster 3 with a DBI of 0.7153. In addition, this study calculated the average computation from 5 experiments, namely K-Means with an average time of 0.024278/s and K-Medoids of 0.05719/s. Based on the lower DBI, K-Medoids have better results in clustering, but the K-Means method is better in terms of computational efficiency. It is hoped that the results of this research will provide valuable insights for retail business people in analyzing sales data.

Keywords: Retail business, Clustering, Davies-Bouldin Index, K-Means, K-Medoids.

1. Introduction

Currently, the retail business has developed rapidly in line with the growth of consumer needs which increases every year (Ong et al., 2020). A retail business is a business that sells goods directly to consumers by breaking down several products into smaller ones and including goods and services (Anjani, 2019; Vebyanti YPontoh et al., 2024). One of the challenges in retail is competition between competitors so the use of information technology in the retail business is important in facing competition in the global market (Ong et al., 2020). Information technology allows retail businesses to manage data and analyze large amounts of data quickly and efficiently. However, in managing retail businesses there are obstacles such as the difficulty of carrying out good management in analyzing large amounts of data (Septiani et al., 2024). The application of information technology can help decision-making in increasing sales, better understanding consumer product purchasing patterns, and knowing sales patterns in product groupings (Sani, 2018; Takdirillah, 2020). The role of big data and analysis in retail business is important in managing retail business (Bradlow et al., 2017). Analysis in the retail business is needed to process and analyze large amounts of data using data mining techniques to find relationships between data that are related to analyzing patterns (Diana et al., 2023).

Data Mining is the right technique for analyzing data. This technique is used to explore and filter data that allows retail business people to make decisions quickly (Sani, 2018). Data Mining, known as Knowledge Discovery in Databases (KDD), is a process of finding patterns by processing large amounts of data (Fatmawati & Windarto, 2018; Hadi & Diana, 2020; Tarigan, 2023). One method in Data Mining is Clustering (Sani, 2018). Clustering is a data mining technique that is used to examine relationships between data by grouping based on the similarity of data entered into the same cluster, while data that has different characteristics is included in another cluster (Diana et al., 2023; Gupta et al., 2021; Utomo, 2021). Clustering or grouping is shown to find hidden patterns in the relationships between data. Clustering is widely used in data analysis, machine learning, prediction, computing, and other studies such as economics (Arora et al., 2016; Gupta et al., 2021). There are two popular algorithms in clustering, namely the K-Means algorithm and K-Medoids Clustering. The difference between the two is that K-Means uses the average value (dividing n data points) to determine the centroid center, while K-Medoids uses an object that is the medoid/object that represents the cluster as the centroid center (Gupta et al., 2021; Hoerunnisa et al., 2024). The K-Medoids algorithm uses the total deviation/closeness distance to form clusters between medoids and non-medoids repeatedly until it converges (Mayadi et al., 2023). K-Medoids is a development method of K-Means Clustering to overcome the problem of outlier data sensitivity, even though it has higher computational complexity (Intan et al., 2023; Mousavi et al., 2020). One metric for measuring the distance of data to the cluster center in both methods is the Euclidean Distance metric (Ramadhani et al., 2022).

Several previous studies that examined the comparison of K-Means and K-Medoids include research conducted by Wargijono Utomo to cluster the spread of Covid-19 which produced a Davies-Bouldin Index value for K-Means $k = 5$ of 0.064 and K-Medoids $k = 2$ of 0.411. This study produced results that K-Means is better than K-Medoids Clustering based on the Davies-Bouldin Index value (Utomo, 2021). In addition, there is previous research conducted by Reza Gustrianda, et al. who clustered product data. This study produced the best results with K-Means with a K-Means result of 0.430 using the Davies Bouldin-Index value and a DBI value for K-Medoids of 1.392 (Gustrianda & Mulyana, 2022). Other research was conducted by Preeti Arora who analyzed the comparison of K-Means and K-Medoids on Big Data which resulted in K-Medoids being better in terms of computing time, not sensitive to outlier data, but the complexity of K-Medoids was higher than K-Means (Arora et al., 2016).

Based on previous research, it can be seen that the performance of the K-Means and K-Medoids algorithms varies and has different results depending on the characteristics of the data and the context of its application. Some studies examine the comparison of the two algorithms, but there is still a lack of studying and comparing the two algorithms in grouping retail store products. Therefore, this study aims to analyze and compare the effectiveness of K-Means and K-Medoids in grouping retail store products. The use of the Davies-Bouldin Index as a cluster validity metric can also assist analysis in testing cluster results against retail store products. The results of the study are expected to provide valuable insights for retail business actors in choosing the optimal clustering algorithm for analyzing sales data.

2. Literature Review

2.1. Clustering

One of the data mining techniques is clustering, which is grouping several data objects into a cluster so that each cluster contains different data from other clusters (Nadiyah et al., 2024). Clustering can be done by selecting a cluster center which then divides the data into several groups based on the similarity of attributes from a set of data by calculating the distance between two records, the results of which can then form a pattern to increase the performance of a company or business (Khanbabaei et al., 2019). In data mining, there are two types of clustering methods, namely hierarchical clustering with nonhierarchical or partition-based clustering (Setiawan, 2016). Hierarchical clustering is an approach with a method of one whole data and two or more data can be grouped by forming a hierarchy. Meanwhile, partition-based clustering is an approach to grouping clusters with predetermined groups by dividing the data set into clusters so that each data only becomes one cluster or group (Nadiyah et al., 2024). The difference between hierarchical clustering and partition-based clustering is that hierarchical has clusters that are part of other clusters. This clustering has the usefulness of finding distribution patterns in data sets in the process of directed data analysis (Nabila et al., 2021). To produce better patterns or results, clustering can be combined with two methods to make mining capacity more optimal and cover the weaknesses of other methods (Zou, 2020).

2.2. Min-Max Normalization

Min-Max Normalization is one of the data normalization methods before the data is processed by data mining by performing linear transformation on real data (Nishom, 2019). Normalization is carried out at the Data Transformation stage in KDD (Knowledge Discovery in Databases) which aims to map the data range into the same range as the range 0 and 1. The equation for Min-Max normalization is given in equation (1) (Melina et al., 2022).

$$x' = \frac{x - v_{min}}{v_{max} - v_{min}} \tag{1}$$

where

x' : normalized data
x : data per column
v_{min} : minimum value in a data column
v_{max} : maximum value in a data column.

2.3. K-Means Clustering

K-Means Clustering is an iterative clustering method that has high efficiency and can use sample analysis and variables with various data types (Zou, 2020). K-Means is a clustering method that is included in partitional-based clustering which can be started by selecting the number of k or clusters randomly which become the center or centroid of the cluster which is then calculated the distance of each data using Euclidean Distance so that the closest distance is found (Putra & Wadisman, 2018). The stages of K-Means are

a) Initialize or determine k or the cluster group formed
b) Random selection and determination of k as the centroid center
c) Calculate the distance of each data using the Euclidean Distance equation, using equation (2).

$$d(X_i, X_g) = ||x - y|| = \sqrt{\textstyle\sum_{j=1}^{p}(X_{ij} - X_{gj})^2} \tag{2}$$

where
$d(X_i, X_g)$: distance of data to the centroid
X_{ij} : data variable/k-th data attribute
X_{gj} : k-th center point on data attribute

d) Grouping of cluster member data with the closest distance to the centroid center based on the minimum distance with equation (3).

$$a_{ij}\begin{cases}1, & s = \min\{d\left(X_i, C_{kj}\right)\}\\ 0, & others\end{cases} \tag{3}$$

where
a_{ij} : the value of a member point X_i into a centroid center C_{kj}
s : the shortest distance from X_i to a centroid center C_{kj} after comparison

e) Determine the new centroid center by calculating the average of objects at similar centroids, using equation (4).

$$C_{kj} = \frac{1}{n_k}\sum d_k \tag{4}$$

where
C_{kj} : k-th centroid center on variable j ($j = 1, 2, \ldots\ p$)
n_k : amount of data in cluster k
d_k : data on cluster k

f) Iterate through steps 3 to 6 to ensure that the centroid changes are within the specified limits and there is no cluster shift.

2.4. K-Medoids Clustering

K-Medoids Clustering is a grouping method using centralized objects (medoids) in the cluster as the center of the cluster from the average object. K-Medoids Clustering is a partitioning method that is a development of K-Means to overcome sensitivity to outliers (Intan et al., 2023; Murpratiwi et al., 2021), which is an advantage of K-Medoids. However, K-Medoids have a disadvantage in computation that tends to be more complex because K-medoids use objects as medoids rather than the average at the center of the cluster (Mousavi et al., 2020). The steps for K-Medoids Clustering are

a) Initialize as many cluster centers as the number of cluster groups formed.
b) Distribute each data into clusters using Euclidean Distance using equation (5).

$$d(X_i, X_g) = ||x - y|| = \sqrt{\textstyle\sum_{j=1}^{p}(X_{ij} - X_{gj})^2} \tag{5}$$

c) Select random objects in each cluster as new candidate medoids.
d) Calculate the distance of each object in the cluster with the new candidate medoids.
e) Calculate the total deviation (S) with the total value of the new distance – the total old distance. If S is less than 0 then the object is exchanged with the cluster data as the formation of a new set of k objects as medoids.

f) Iterate steps 3 to 5 until no changes occur in medoids.

2.5. Davies-Bouldin Index

Davies Bouldin Index or DBI is an index to measure cluster evaluation of grouping/clustering which was introduced in 1979 by Davis L. Davies and Donald W. Bouldin (Tarigan, 2023). The best value of Davies Bouldin Index has a result close to 0, not negative. The equation of Davies Bouldin Index is given in equation (6) (Tempola & Assagaf, 2018).

$$DBI = \frac{1}{k}\sum_{i=1,i=j}^{k} \max(R_{i,j}) \tag{6}$$

where
k : number of clusters
$R_{i,j}$: ratio or comparison between i-th cluster and j-th cluster.

A good cluster has a small cohesion value and the largest possible separation. The equation for calculating $R_{i,j}$ or the rasio between clusters is given in equation (7).

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{7}$$

SSW or Sum of Square Within Cluster is a metric of the i-th cluster cohesion. This cohesion is the sum of the distances of the data to the centroid. The equation for the Sum of Square Within Cluster (SSW) is given in equation (8)

$$SSW_i = \frac{1}{n}\sum_{j=1}^{k} d(x_{j,} c_i) \tag{8}$$

where
n : amount of data on the cluster
c_i: i-th centroid
d : distance calculation using Euclidean Distance

As for $SSB_{i,j}$ or Sum of Square Between Cluster is a metric of separation between clusters by calculating the distance between clusters. The equation of Sum of Square Between Cluster (SSB) is given in equation (9).

$$SSB_{i,j} = d\left(c_i, c_j\right) \tag{9}$$

where
c_i : i-th centroid
c_j : j-th centroid

3. Materials and Methods

3.1. Materials

This study analyzed the data needed for the research process. The secondary data used in this study was collected using documentation techniques. Secondary data comes from company documentation or related parties obtained through websites or journals.

3.2. Method

There is a research framework that explains the stages used in compiling the research. In this research, the stages carried out include literature studies, the Knowledge Discovery in Databases (KDD) process in the form of data collection, data selection, data preprocessing, data transformation, and the application of the K-Means and K-Medoids Clustering algorithms. The stages of this research also include system analysis design and evaluation.

1) Literature Study: This stage is carried out by reading and processing research materials so that they can be used as a bibliography. Literature study is carried out by reading journals or books according to relevant research topics so that they can be used as references for solving problems in research.
2) Data Collection: A process of collecting data in data mining that can be stored in the form of files.
3) Data Selection: A process of selecting attributes or instruments selected for data processing.

4) Data Preprocessing: One of the important processes of the research method to avoid data duplication, inaccurate data, or data with typographical errors and unnecessary data. Incorrect data is cleaned to maintain the accuracy of the results and complex data reduces its complexity.
5) Data Transformation: Data transformation is a stage of data coding or adjusting the form of data so that it can be adjusted to the data mining process or Knowledge Discovery in Database (KDD).
6) Determining the Optimal Number of *k*: At this stage, the optimal number of *k* or clusters is determined using the Davies-Bouldin Index metric.
7) Clustering Implementation: At this stage, clustering is implemented or applied using K-Means Clustering and K-Medoids Clustering after determining the optimal number of *k* using the Davies-Bouldin Index value.
8) Evaluation and Results: At this stage, a pattern is examined so that it does not conflict with the hypothesis. This stage is also the stage of data evaluation against clustering using the Davies-Bouldin Index or DBI method by evaluating the cohesion and separation values. This cohesion value is the amount of data proximity to the centroid, and separation is the distance of data between centroids. In the results, a comparison of the k-means and k-medoids methods is carried out based on the DBI value and a comparison of computational time efficiency.

3.2.1 Data Collection

The data collection method used in this study uses a documentation technique where the data source comes from Kaggle.com in the form of grocery data in 2020 with a total of 1085 data with attributes, namely item_id, item_name, total_initial_items_stock, transactions, items_sold, remaining_items, item_purchase_price, purchase_price, selling_price, profit, average sales (one year), and year. Table 1 explains the description of the research data attributes.

Table 1: Data Attribute Description

Attribute	Attribute Description
Item id	ID of grocery items
Item name	Name of grocery items
Total initial items stock	Total stock before sales transaction
Transactions	Number of goods transactions
Items sold	Total of goods sold
Remaining items	Remaining goods that have not been sold
Item purchase price	Price of grocery goods
Purchase price	Price set when purchasing goods/capital
Selling price	Price set for sales to customers
Profit	Income earned as profit from the total price issued
Average sales (one year)	Average sales in one year
Year	Year of grocery goods obtained

3.2.2 Data Selection

This stage is the data selection stage before data processing and data transformation are carried out in the study. Data Selection is done by selecting attributes that are not needed. There are 6 attributes selected in the study, including total initial items stock, transactions, items sold, remaining items, profit, and average sales in one year. The selected attributes are described in Table 2.

Table 2: Attributes After Data Selection

Attributes	Attribute Description
Total initial items stock	Total goods stock before sales transaction
Transactions	Number of goods transactions
Items sold	Total of goods sold
Remaining items	Remaining unsold items
Profit	Income earned as a profit from the amount of price issued
Average sales (one year)	Average sales in one year

The attributes are selected based on their relevance to the analysis objectives and the attributes are not redundant or have similar explanations in the dataset. Data that is not selected for the clustering process includes several attributes, namely Item ID because it cannot affect the analysis results, Item Name which is not relevant to the research and is only descriptive information, Item Purchase Price and Selling Price which are redundant data on the profit attribute, and Year is not selected because information about the attribute is listed on the average sales in one year. The data that has been selected with the selected attributes can be seen in Table 3.

Table 3: Data Selection Results

Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)
6859	26	6751	108	2400	260
6926	30	6358	568	300	2
6764	3619	6233	531	1400	212
6781	2280	5937	844	1500	3
6758	2919	3556	3202	800	1
6755	1688	3476	3279	2800	2
6840	28	3400	3440	1300	121
6982	28	3400	3582	2500	227
6926	15	3400	3526	2000	3
…	…	…	…	…	…
6921	1	1	6920	2000	1

3.2.3 Data Preprocessing

Data Preprocessing or data cleaning is a stage of data cleaning by removing duplicate data, incomplete data, and irrelevant data to maintain the consistency and accuracy of the data. This stage also ensures that there are no missing values in preventing errors in the analysis process. In this study, there were missing values or null data as cleaning data was carried out so the data from the cleaning amounted to 1080 data.

3.2.4 Data Tranformation

This stage is a transformation stage by adjusting the data by changing the data form into a new data form so that it is following the application of data mining. Data normalization is carried out using min-max normalization so the data range can be normalized between range 0 and 1 through mapping. The results of normalization can be seen in Table 4.

Table 4: Data Transformation

Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)
0.434	0.007	1	0.000	0.793	1.000
0.703	0.008	0.942	0.067	0.069	0.004
0.052	1.000	0.923	0.061	0.448	0.815
0.120	0.630	0.879	0.107	0.483	0.008
0.028	0.807	0.527	0.449	0.241	0.000
0.016	0.466	0.515	0.460	0.931	0.004
0.357	0.007	0.504	0.484	0.414	0.463
0.928	0.007	0.504	0.504	0.828	0.873
0.703	0.004	0.504	0.496	0.655	0.008
…	…	…	…	…	…
0.683	0.000	0.000	0.989	0.655	0.000

4. Results and Discussion

4.1. Data Mining

1) Determining the Optimal Number of Clusters

The initial stage in the clustering process is to determine the optimal number of clusters using the Davies Bouldin Index (DBI) in both methods. Determining the optimal number of clusters or k is needed to affect the clustering results to prevent the formation of non-optimal clusters (Murpratiwi et al., 2021). The number of clusters tested using the Davies Bouldin Index is $\boldsymbol{k = 2}$, $\boldsymbol{k = 3}$, $\boldsymbol{k = 4}$, and $\boldsymbol{k = 5}$ to determine the smallest value as an indicator in determining the best number of clusters. Table 5 is the determination of the optimal number of clusters in the study.

Table 5: Determining the Optimal Number of Clusters

Amount *k*	**DBI *K-Means***	**DBI *K-Medoids***
2	1.000	1.000
3	**0.716**	**0.715**
4	0.754	0.864
5	0.848	0.927

Based on Table 5, the best Davies-Bouldin Index (DBI) results are determined by results approaching 0 non-negative or the smallest value, so that the optimal number of clusters for K-Means Clustering and K-Medoids Clustering in this study is *k = 3* or a total of 3 clusters.

2) K-Means Clustering

The first step in K-Means Clustering is to determine *k* or cluster groups formed as many as clusters. The data selected in this study is divided into 3 clusters so that ***k* = 3**. Then the initial centroid value is selected randomly. The randomly selected centroids are given in Table 6.

Table 6: Determining the Initial Centroid of K-Means

Centroid	**Row**	**Total initial items stock**	**Transactions**	**Items sold**	**Remaining items**	**Profit**	**Average sales (one year)**
C1	*R2*	0.703	0.008	0.942	0.067	0.069	0.004
C2	*R5*	0.028	0.807	0.527	0.449	0.241	0
C3	*R8*	0.928	0.007	0.504	0.504	0.828	0.873

There are *C1*, *C2*, and *C3* which are the cluster divisions, and Row which is the data row number to facilitate the clustering process. Then find the distance between the data and the cluster center using Euclidean Distance. The distance calculation on *C1* is as follows.

$$d(R_1, C_1) = \sqrt{\begin{matrix}(0.434-0.703)^2+(0.007-0.008)^2+(1-0.924)^2\\ +(0.000-0.067)^2+(0.793-0.069)^2+(1-0.004)^2\end{matrix}}$$

= 1.263

$$d(R_2.C_1) = \sqrt{\begin{matrix}(0.703-0.703)^2+(0.008-0.008)^2+(0.924-0.924)^2\\ +(0.067-0.067)^2+(0.069-0.069)^2+(0.004-0.004)^2\end{matrix}}$$

= 0.000

$$d(R_3.C_1) = \sqrt{\begin{matrix}(0.052-0.703)^2+(1-0.008)^2+(0.923-0.924)^2\\ +(0.061-0.067)^2+(0.448-0.069)^2+(0.815-0.004)^2\end{matrix}}$$

= 1.486

Distance calculation on *C2*

$$d(R_1.C_2) = \sqrt{\begin{matrix}(0.434-0.028)^2+(0.007-0.807)^2+(1-0.527)^2\\ +(0.000-0.449)^2+(0.793-0.241)^2+(1-0.000)^2\end{matrix}}$$

= 1.592

$$d(R_2.C_2) = \sqrt{\begin{matrix}(0.703-0.028)^2+(0.008-0.807)^2+(0.942-0.527)^2\\ +(0.067-0.449)^2+(0.069-0.241)^2+(0.004-0.000)^2\end{matrix}}$$

= 1.201

$$d(R_3.C_2) = \sqrt{\begin{matrix}(0.052-0.028)^2+(1-0.807)^2+(0.923-0.527)^2\\ +(0.061-0.449)^2+(0.448-0.241)^2+(0.815-0.000)^2\end{matrix}}$$

= 1.026

Distance calculation on *C3*

$$d(R_1.C_3) = \sqrt{\begin{matrix}(0.434-0.928)^2+(0.007-0.007)^2+(1-0.504)^2\\+(0.000-0.504)^2+(0.793-0.828)^2+(1-0.873)^2\end{matrix}}$$
$$= 0.873$$

$$d(R_2.C_3) = \sqrt{\begin{matrix}(0.703-0.928)^2+(0.008-0.007)^2+(0.942-0.504)^2\\+(0.067-0.504)^2+(0.069-0.828)^2+(0.004-0.873)^2\end{matrix}}$$
$$= 1.328$$

$$d(R_3.C_3) = \sqrt{\begin{matrix}(0.052-0.928)^2+(1-0.007)^2+(0.923-0.504)^2\\+(0.061-0.504)^2+(0.448-0.828)^2+(0.815-0.873)^2\end{matrix}}$$
$$= 1.508$$

The results of the distance calculation in iteration 1 can be seen in Table 7. The closeness of the data can be seen from the minimum cost. It can be seen that the 1st data has a value of $C3 < C1$ and $C3 < C2$ so it can be seen that the 1st data in iteration 1 is included in Cluster 3.

Table 7: Euclidean Distance Calculation Iteration 1 K-Means

Total initial items stock	Transactins	Items sold	Remaining items	Profit	Average sales (one year)	*C1*	*C2*	*C3*	Cluster
0.434	0.007	1	0	0.793	1	1.263	1.592	0.873	Cluster 3
0.703	0.008	0.942	0.067	0.069	0.004	0.000	1.201	1.328	Cluster 1
0.052	1	0.923	0.061	0.448	0.815	1.486	1.026	1.508	Cluster 2
0.12	0.63	0.879	0.107	0.483	0.008	0.951	0.582	1.485	Cluster 2
0.028	0.807	0.527	0.449	0.241	0	1.201	0.000	1.600	Cluster 2
0.016	0.466	0.515	0.46	0.931	0.004	1.327	0.770	1.345	Cluster 2
0.357	0.007	0.504	0.484	0.414	0.463	0.903	0.997	0.816	Cluster 3
0.928	0.007	0.504	0.504	0.828	0.873	1.328	1.600	0.000	Cluster 3
0.703	0.004	0.504	0.496	0.655	0.008	0.848	1.129	0.910	Cluster 1
0.494	0.271	0.494	0.498	0.172	0.463	0.849	0.853	0.926	Cluster 1
…	…	…	…	…	…	…	…	…	…
0.659	0	0	0.988	0.655	0	1.443	1.338	1.163	Cluster 3
0.554	0	0	0.984	0.724	0	1.476	1.313	1.182	Cluster 3
0.683	0	0	0.989	0.655	0	1.443	1.349	1.158	Cluster 3

The next step is to calculate the new centroid center by calculating the average of the objects. The new centroid center in the 2nd iteration can be seen in Table 8.

Table 8: New Centroid Center Iteration 2 K-Means

Centroid	Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)
C1	0.839	0.027	0.107	0.890	0.098	0.024
C2	0.259	0.018	0.017	0.957	0.519	0.005
C3	0.677	0.009	0.013	0.976	0.748	0.014

Then calculate the distance of the data with the centroid using Euclidean Distance and determine the cluster with the minimum cost or distance which can be seen in Table 9.

Table 9: Euclidean Distance Calculation Iteration 2 K-Means

Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)	*C1*	*C2*	*C3*	Cluster
0.434	0.007	1	0	0.793	1	1.786	1.726	1.721	Cluster 3
0.703	0.008	0.942	0.067	0.069	0.004	1.181	1.431	1.467	Cluster 1
0.052	1	0.923	0.061	0.448	0.815	1.915	1.814	1.942	Cluster 2
0.12	0.63	0.879	0.107	0.483	0.008	1.496	1.364	1.507	Cluster 2
0.028	0.807	0.527	0.449	0.241	0	1.288	1.128	1.363	Cluster 2
0.016	0.466	0.515	0.46	0.931	0.004	1.384	0.962	1.095	Cluster 2

0.357	0.007	0.504	0.484	0.414	0.463	0.920	0.832	0.948	Cluster 2
0.928	0.007	0.504	0.504	0.828	0.873	1.252	1.319	1.128	Cluster 3
0.703	0.004	0.504	0.496	0.655	0.008	0.801	0.816	0.694	Cluster 3
0.494	0.271	0.494	0.498	0.172	0.463	0.825	0.942	1.047	Cluster 1
…	…	…	…	…	…	…	…	…	…
0.659	0	0	0.988	0.655	0	0.604	0.424	0.098	Cluster 3
0.554	0	0	0.984	0.724	0	0.703	0.361	0.128	Cluster 3
0.683	0	0	0.989	0.655	0	0.597	0.447	0.097	Cluster 3

In the 2nd iteration, there is still cluster shifting compared to the 1st iteration, so the iteration is repeated until the next iteration until there is no cluster shifting.

3) K-Medoids Clustering

The first step in K-Medoids Clustering is to determine the number of clusters. In this study, the data is divided into 3 clusters so that ***k = 3***. Table 10 is a random determination of the initial cluster center $k = 3$.

Table 10: K-Medoids Initial Medoid Determination

Centroid	Row	Total initial items stock	Transactions	Items sold	Remainin items	Profit	Average sales (one year)
C1	*R2*	0.703	0.008	0.942	0.067	0.069	0.004
C2	*R5*	0.028	0.807	0.527	0.449	0.241	0
C3	*R8*	0.928	0.007	0.504	0.504	0.828	0.873

The Row column is the n-th row of data to facilitate the clustering process. The initial medoid with $k = 3$ is *C1*, *C2* and *C3*. In determining the initial medoid center, the K-Medoids stage initializes the same cluster center as the initial medoid center in K-Means. Then the next stage in K-Medoids Clustering is the calculation of Euclidean Distance to find the distance between data and the center of the cluster/medoid. The results in the 1st iteration of K-Medoids have the same results as K-Means because they have the same initial cluster center. The calculation of the distance in iteration 1 of K-Medoids can be seen in Table 11. Cluster grouping is determined by finding the minimum value between data distances/finding the smallest value, for example in the data in Table 11 the 1st data has a value of *C3<C1* and *C3<C2,* thus producing group cluster 3.

Table 11: Euclidean Distance Calculation Iteration 1 K-Medoids

Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)	*C1*	*C2*	*C3*	Min	Cluster
0.434	0.007	1	0	0.793	1	1.263	1.592	0.873	0.873	Cluster 3
0.703	0.008	0.942	0.067	0.069	0.004	0.000	1.201	1.328	0.000	Cluster 1
0.052	1	0.923	0.061	0.448	0.815	1.486	1.026	1.508	1.026	Cluster 2
0.12	0.63	0.879	0.107	0.483	0.008	0.951	0.582	1.485	0.582	Cluster 2
0.028	0.807	0.527	0.449	0.241	0	1.201	0.000	1.600	0.000	Cluster 2
0.016	0.466	0.515	0.46	0.931	0.004	1.327	0.770	1.345	0.770	Cluster 2
0.357	0.007	0.504	0.484	0.414	0.463	0.903	0.997	0.816	0.816	Cluster 3
0.928	0.007	0.504	0.504	0.828	0.873	1.328	1.600	0.000	0.000	Cluster 3
0.703	0.004	0.504	0.496	0.655	0.008	0.848	1.129	0.910	0.848	Cluster 1
0.494	0.271	0.494	0.498	0.172	0.463	0.849	0.853	0.926	0.849	Cluster 1
…	…	…	…	…	…	…	…	…	…	…
0.659	0	0	0.988	0.655	0	1.443	1.338	1.163	1.163	Cluster 3
0.554	0	0	0.984	0.724	0	1.476	1.313	1.182	1.182	Cluster 3
0.683	0	0	0.989	0.655	0	1.443	1.349	1.158	1.158	Cluster 3

The total proximity/total cost of iteration 1 is obtained by summing the minimum of iteration 1.

$$Total\ Cost_1 = 0.878 + 0.000 + 1.026 + 0.582 + \cdots + 1.158 = 1279.639$$

The next step is to determine the new cluster center randomly. Table 12 is the determination of the new cluster center/medoid in iteration 2 by taking the 6th data as R6, the 7th data as R7, and the 10th data as $R10$.

Table 12: New Medoid Iteration 2 K-Medoids

Centroid	Row	Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)
C1	R6	0.016	0.466	0.515	0.46	0.931	0.004
C2	R7	0.357	0.007	0.504	0.484	0.414	0.463
C3	R10	0.494	0.271	0.494	0.498	0.172	0.463

The distance calculation using Euclidean Distance in iteration 2 can be seen in Table 13.

Table 13: New Medoid Iteration 2 K-Medoids

Total initial items stock	Transactions	Items sold	Remaining items	Profit	Average sales (one year)	*C1*	*C2*	*C3*	Min	Cluster
0.434	0.007	1	0	0.793	1	1.358	0.958	1.119	0.958	Cluster 2
0.703	0.008	0.942	0.067	0.069	0.004	1.327	0.903	0.849	0.849	Cluster 3
0.052	1	0.923	0.061	0.448	0.815	1.226	1.248	1.141	1.141	Cluster 3
0.12	0.63	0.879	0.107	0.483	0.008	0.704	0.969	0.935	0.704	Cluster 1
0.028	0.807	0.527	0.449	0.241	0	0.770	0.997	0.853	0.770	Cluster 1
0.016	0.466	0.515	0.46	0.931	0.004	0.000	0.898	1.027	0.000	Cluster 1
0.357	0.007	0.504	0.484	0.414	0.463	0.898	0.000	0.384	0.000	Cluster 2
0.928	0.007	0.504	0.504	0.828	0.873	1.345	0.816	0.926	0.816	Cluster 2
0.703	0.004	0.504	0.496	0.655	0.008	0.874	0.620	0.745	0.620	Cluster 2
0.494	0.271	0.494	0.498	0.172	0.463	1.027	0.384	0.000	0.000	Cluster 3
…	…	…	…	…	…	…	…	…	…	…
0.659	0	0	0.988	0.655	0	1.118	0.934	1.016	0.934	Cluster 2
0.554	0	0	0.984	0.724	0	1.044	0.924	1.037	0.924	Cluster 2
0.683	0	0	0.989	0.655	0	1.133	0.942	1.021	0.942	Cluster 2

Calculations are made to calculate the proximity/total cost by adding the minimum cost and calculating the cost difference between the new total cost and the old total cost.

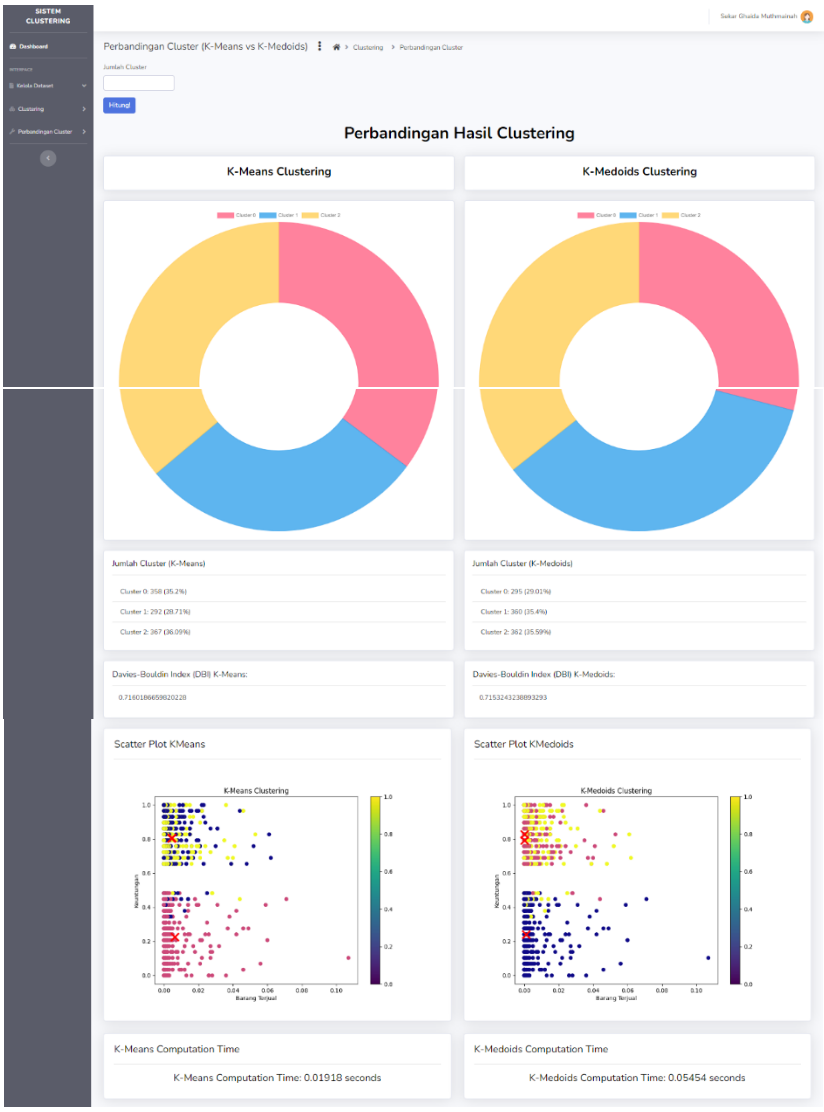
$$Total\ Cost_2 = 0.958 + 0.849 + 1.141 + 0.704 + \cdots + 0.942 = \ 1011.632$$

$$\begin{aligned} S &= Total\ Cost_2 - \ Total\ Cost_1 \\ &= \ 1011.632 - \ 1279.639 \\ &= -268.007 \end{aligned}$$

The proximity difference is obtained by calculating the difference between the new total cost and the old total cost. If the difference is less than 0 or $S < 0$ then the iteration is continued. The calculation in this study was continued until the 4th iteration where the difference was more than 0.

4.2. Implementation

The implementation of the software built in this study uses a personal computer or stand-alone, by building web-based software using the Python language and Flask as a framework with tools using Visual Studio Code and Google Chrome and Microsoft Edge as web browsers as shown in Figure 1.



Perbandingan Cluster (K-Means vs K-Medoids)

Jumlah Cluster

Perbandingan Hasil Clustering

K-Means Clustering

K-Medoids Clustering

Jumlah Cluster (K-Means)

Cluster 0: 358 (35.2%)

Cluster 1: 292 (28.71%)

Cluster 2: 367 (36.09%)

Jumlah Cluster (K-Medoids)

Cluster 0: 295 (29.01%)

Cluster 1: 360 (35.4%)

Cluster 2: 362 (35.59%)

Davies-Bouldin Index (DBI) K-Means:

0.7160186659820228

Davies-Bouldin Index (DBI) K-Medoids:

0.7153243238893293

Scatter Plot KMeans

Scatter Plot KMedoids

K-Means Computation Time

K-Means Computation Time: 0.01918 seconds

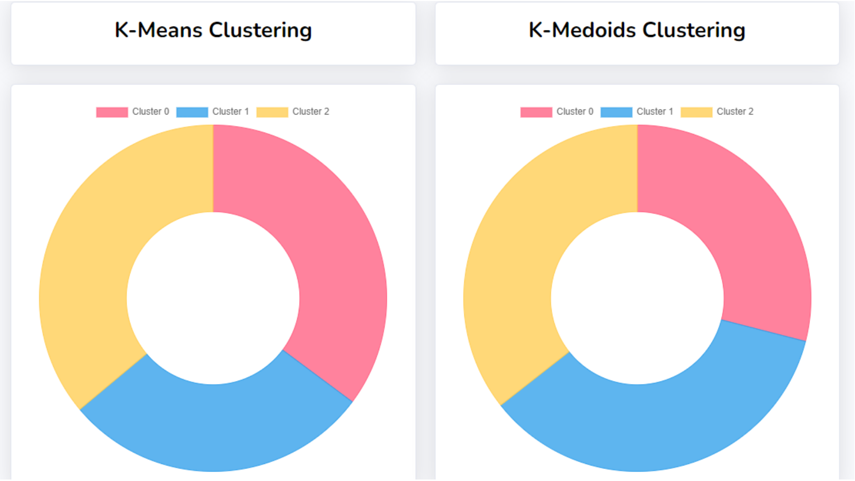
K-Medoids Computation Time

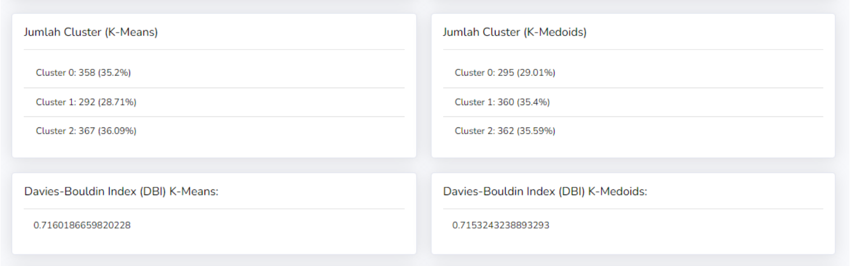
K-Medoids Computation Time: 0.05454 seconds

Figure 1: Clustering Comparison Interface

4.3. Evaluation And Results

This study uses cluster evaluation using the Davies Bouldin Index which is an index for testing cluster results. In this study, there are K-Means results with a total of *k = 3*, namely cluster 1 as many as 358 data, cluster 2 as many as 292 data, and cluster 3 as many as 367 data. The results of clustering k = 3 using K-Medoids have a total of 295 data cluster 1, 360 data cluster 2, and 362 data cluster 3. The cluster results can be seen in Figure 2.





Jumlah Cluster (K-Means)

Cluster 0: 358 (35.2%)

Cluster 1: 292 (28.71%)

Cluster 2: 367 (36.09%)

Jumlah Cluster (K-Medoids)

Cluster 0: 295 (29.01%)

Cluster 1: 360 (35.4%)

Cluster 2: 362 (35.59%)

Davies-Bouldin Index (DBI) K-Means:

0.7160186659820228

Davies-Bouldin Index (DBI) K-Medoids:

0.7153243238893293

Figure 2: Clustering Chart Results

The DBI results on K-Means *k=3* have a value of 0.7160, while the DBI results on K-Medoids *k=3* have a value of 0.7153 which is shown in Table 14.

Table 14: Davies Bouldin Index Results

DBI *K-Means* Results	0.7160
DBI *K-Medoids* Results	0.7153

The clustering results are said to be optimal if the DBI index value has the smallest possible value or is close to 0 and not negative, so based on the DBI results between the two methods, it is found that the use of K-Medoids Clustering is better than K-Means Clustering based on the Davies-Bouldin Index metric evaluation. In addition, there is also the computation time of each method producing an average k-means computation time of 0.024278/s and k-medoids with 0.05719/s which is shown in Table 15, so the results obtained that K-Means is better based on efficiency against computation time. This is because the average calculation in determining the centroid is faster than selecting objects on new medoids or the centroid calculation has a simpler computation.

Table 15: Computation Time *k=3*

Attempt	***K-Means Computation Time***	***K-Medoids Computation Time***
1	0.03074/s	0.06439/s
2	0.02951/s	0.06594/s
3	0.01533/s	0.04868/s
4	0.02859/s	0.04397/s
5	0.01722/s	0.06296/s
Average	**0.024278/s**	**0.05719/s**

This study shows that the best results between the K-Means or K-Medoids methods depend on the analysis needs. K-Means has the advantage of more efficient computing time because the computation in determining the centroid with the mean is simpler than determining the medoid. K-Medoids have an advantage over K-Means based on the Davies Bouldin Index value to test the quality of clusters in the grocery dataset even though the values are not that different.

In addition, there are also scatter plot results in the clustering results shown in Figure 3. The x-axis or feature 1 in the scatter plot shows 'Items Sold' and the y-axis or feature 2 is 'Profit'. These attributes are selected based on the relevance between the two features to determine product grouping. These attributes can interpret the products in demand, besides being profitable.

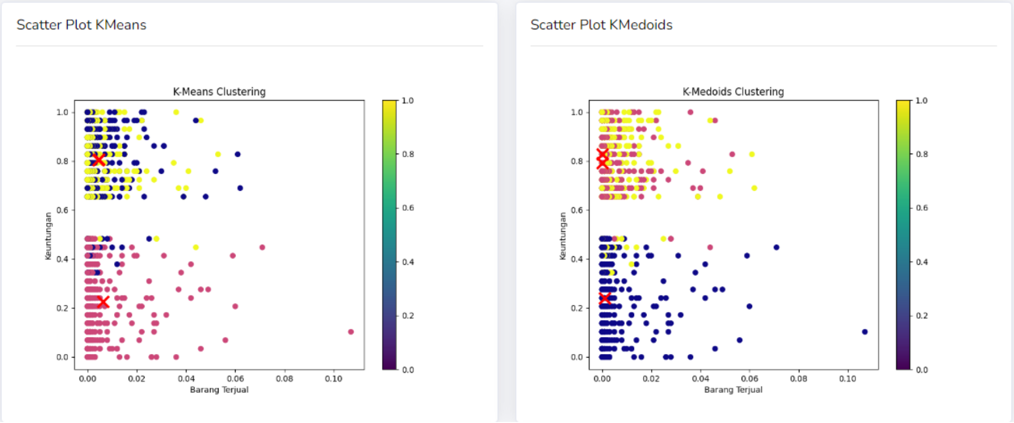


Figure 3: Scatter Plot K-Means and K-Medoids Items Sold and Profit

The distribution of data on the selected features based on the clustering results in both methods tends to show that most of the data on ‘Items Sold' is very low because it is close to the value of 0. However, in the 'Profit' feature, there is clustering data that is classified as low (close to 0) and high (close to 1). The clustering results on "Items Sold - low" and "Profit - high" allow for a low quantity of items sold with high profits where some products are sold at high prices. The results on "Items sold - low" and "Profit - low" interpret that there are goods that have a sales volume that is not very good with low profits.

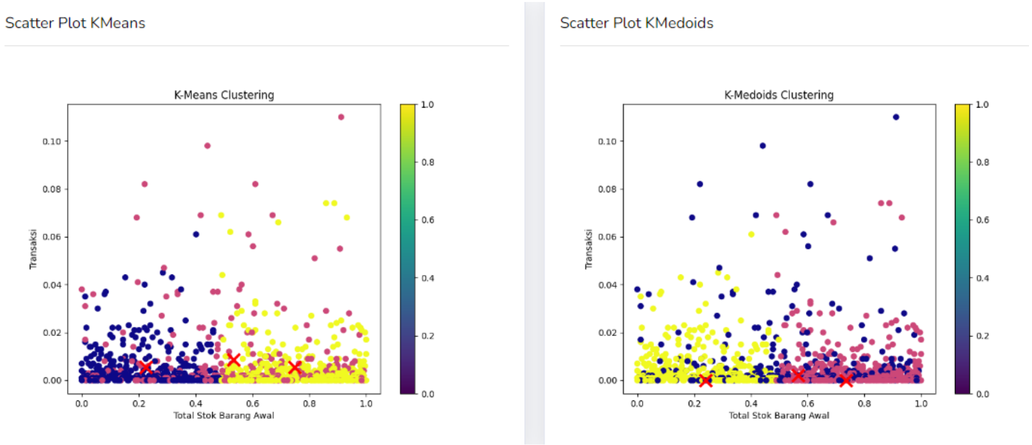


Figure 4: Scatter Plot K-Means and K-Medoids Total Initial Stock and Transactions

There are also Scatter plot results in Figure 4 in visualizing the relationship between the variables ‘Total Initial Stock’ as the x-axis or feature 1 and ‘Transaction’ as the y-axis or feature 2 to find patterns in product grouping. The scatter plot interprets that the clustering results in both methods have a high ‘Transaction” pattern (located at the top of the plot) with a “Total Initial Stock” that tends to be high where some items in the initial stock have a lot of/high stock and have good sales. However, in the scatter plot results, there is also a cluster density in the total initial stock of goods with low transactions.

5. Conclusion

This study obtained the optimum number of clusters of 3 clusters using the Davies Bouldin Index. The grouping pattern is based on 6 attributes that have been carried out in the Data Selection stage, namely total initial items stock, transactions, items sold, remaining items, profit, and average sales. The results of clustering produce a Davies Bouldin Index value for K-Means of 0.7160 with cluster 1 totaling 358 data, cluster 2 totaling 292 data, and cluster 3 totaling 367 product data, while the Davies Bouldin Index value for K-Medoids is 0.7153 with cluster 1 totaling 295 data, cluster 2 totaling 360 data, and cluster 3 totaling 362 data. This shows that the use of K-Medoids is better than K-Means based on testing the cluster results using the DBI value. The average computation time performed by K-Means in 5 trials is 0.024278/s and K-Medoids is 0.05719/s which indicates that K-Means has faster computation time efficiency than K-Medoid. Data distribution in product grouping is visualized using a scatter plot using the features of items sold and profits with the results of data on items sold tending to be very low, but the profit data has low and high values. The results on the initial item total stock feature and transactions have several high stocks with high transactions/sales. Based on the description, it can be concluded that this study has succeeded in showing the effectiveness of clustering techniques in retail data analysis, with K-Medoids slightly better optimal in clustering, while K-Means is optimal in computational time efficiency. The results of this study are expected to help retail business actors in optimizing sales strategies and product management in the retail business. This study can be developed by analyzing outliers in more depth to obtain their influence on clustering results, adding comparisons to other clustering methods such as DBSCAN for better clustering analysis, and other validity evaluation metrics besides the Davies Bouldin Index, in testing cluster quality to obtain a more diverse perspective.

References

Anjani, R. G. (2019). The Role of Information Systems in Retail Operations. *Journal of Economics and Management Information Systems (JEMSI)*, *1*(September), 60–69. https://doi.org/10.31933/JEMSI

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *International Conference on Information Security & Privacy (ICISP)*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. *Journal of Retailing*, *93*(1), 79–95. https://doi.org/10.1016/j.jretai.2016.12.004

Diana, Y., Hadi, F., Ekonomi, F., Bisnis, D., Putra Indonesia, U., Padang, Y., Lubuk, J. R., & Padang, B. (2023). Sales Analysis Using K-Medoids Algorithm to Optimise Sales of Goods. *JOISIE Journal Of Information System And Informatics Engineering*, *7*(1), 97–103.

Fatmawati, K., & Windarto, A. P. (2018). Data Mining: Application of Rapidminer with K-Means Cluster on Dengue Fever (Dbd) Affected Areas by Province. *Computer Engineering, Science and System Journal*, *3*(2), 173. https://doi.org/10.24114/cess.v3i2.9661

Gupta, A., Sharma, H., & Akhtar, A. (2021). A Comparative Analysis Of K-Means And Hierarchical Clustering. *EPRA International Journal of Multidisciplinary Research (IJMR)*, *7*(8). https://doi.org/10.36713/epra2013

Gustrianda, R., & Mulyana, D. I. (2022). Application of Data Mining in Product Selection with K-Means and K-Medoids Algorithm Methods. *Jurnal Media Informatika Budidarma*, *6*(1), 27. https://doi.org/10.30865/mib.v6i1.3294

Hadi, F., & Diana, Y. (2020). Clustering Building Material Sales Using K-Means Algorithm. *JOISIE (Journal Of Information Systems And Informatics Engineering)*, *4*(1), 22. https://doi.org/10.35145/joisie.v4i1.629

Hoerunnisa, A., Dwilestari, G., Dikananda, F., Sunana, H., & Pratama, D. (2024). Comparison of K-Means and K-Medoids Algorithms in Clustering Analysis of Crime Prone Areas in Indonesia. *Jurnal Mahasiswa Teknik Informatika*, *8*(1). https://doi.org/https://doi.org/10.36040/jati.v8i1.8249

Intan, S. F., Elvira, W., Rahayu, S., & ... (2023). Comparison of K-Means and K-Medoids Algorithms for Student Expenditure Grouping. *SENTIMAS : National Seminar on Research and Community Service*, 35–40. https://journal.irpi.or.id/index.php/sentimas/article/view/543

Khanbabaei, M., Alborzi, M., Sobhani, F. M., & Radfar, R. (2019). Applying clustering and classification data mining techniques for competitive and knowledge-intensive processes improvement. *Journal of Corporate Transformation*, *26*(2), 123–139. https://doi.org/10.1002/kpm.1595

Mayadi, Setiawati, S., & Priatna, W. (2023). Grouping MBKM Survey Results Using K-Mean and K-Medoids Clustering. *Jurnal Media Informatika Budidarma*, *7*. https://doi.org/10.30865/mib.v7i1.5003

Melina, Napitupulu, H., Sambas, A., Murniati, A., & Adimurti Kusumaningtyas, V. (2022). Artificial Neural Network-Based Machine Learning Approach to Stock Market Prediction Model on the Indonesia Stock Exchange During the COVID-19. *Engineering Letters*, *30*(3). https://www.researchgate.net/publication/362983602

Mousavi, S., Boroujeni, F. Z., & Aryanmehr, S. (2020). Improving customer clustering by optimal selection of cluster centroids in K-means and K-medoids algorithms. *Journal of Theoretical and Applied Information Technology*, *8*(10), 3807–3814.

Murpratiwi, S. I., Agung Indrawan, I. G., & Aranta, A. (2021). Analysis of Optimal Cluster Selection in Retail Store Customer Segmentation. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *18*(2), 152. https://doi.org/10.23887/jptk-undiksha.v18i2.37426

Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). Data Mining Analysis for Clustering Covid-19 Cases in Lampung Province with K-Means Algorithm. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(2), 100. http://jim.teknokrat.ac.id/index.php/JTSI

Nadiyah, Arifin, N. H. I., & Karim, A. (2024). Application of K-Means Algorithm for Clustering Service Assessment Based on Nurul Jadid University Student Satisfaction Index. *Jurnal Advance Research Informatika*, *2*, 23–30. https://doi.org/10.24929/jars.v2i2.3431

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Ong, J. O., Sutawijaya, A. H., & Saluy, A. B. (2020). Modern Retail Business Model Innovation Strategy in the Industry 4.0 Era. *Jurnal Ilmiah Manajemen Bisnis*, *6*(2), 201–210. https://doi.org/https://dx.doi.org/10.22441/jimb.v6i2.8891

Putra, R. R., & Wadisman, C. (2018). Implementation of Data Mining Potential Customer Selection Using K-Means Algorithm. *INTECOMS (Journal of Information Technology and Computer Science*, *151*(2), 10–17. https://doi.org/10.31539/intecoms.v1i1.141

Ramadhani, S., Azzahra, D., & Z, T. (2022). Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(1), 24–33. https://doi.org/10.31849/digitalzone.v13i1.9292

Sani, A. (2018). Application of the K-Means Clustering Method in Companies. *Jurnal Ilmiah Teknologi Informasi*, *May*, 1–7.

Septiani, S., Musthofa, & Seviawani, P. (2024). Using Big Data for Service Personalisation in E-Commerce Businesses. *ADI Bisnis Digital Interdisiplin Jurnal*, *5*(1), 51–57. https://doi.org/10.34306/abdi.v5i1.1098

Setiawan, R. (2016). Application of Data Mining Using K-Means Clustering Algorithm to Determine New Student Promotion Strategy (Case Study: Politeknik Lp3i Jakarta). *Jurnal Lentera ICT*, *3*(1), 76–92.

Takdirillah, R. (2020). Application of Data Mining Using Apriori Algorithm on Transaction Data to Support Sales Strategy Information. *Edumatic : Jurnal Pendidikan Informatika*, *4*(1), 37–46. https://doi.org/10.29408/edumatic.v4i1.2081

Tarigan, D. A. (2023). Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification. *Kajian Ilmiah Informatika Dan Komputer (KLIK)*, *4*(1), 545–552. https://doi.org/10.30865/klik.v4i1.964

Tempola, F., & Assagaf, A. F. (2018). Clustering of Potency of Shrimp in Indonesia with K-Means Algorithm and Validation of Davies-Bouldin Index. *International Conference on Science and Technology (ICST 2018)*, *1*. https://doi.org/10.2991/icst-18.2018.148

Utomo, W. (2021). The Comparison of K-means And K-medoids Algorithms For Clustering The Spread Of The Covid-19 Outbreak in Indonesia. *ILKOM Jurnal Ilmiah*, *13*(1), 31–35. https://doi.org/10.33096/ilkom.v13i1.763.31-35

Vebyanti YPontoh, N., Rahman, F., Yunus, R., Yunus, S., & Ferdiana Paskual, M. (2024). Analysis of the Impact of Modern Retail Markets on the Income of Traditional Retailers in Wuasa Village, North Lore District. *EKOMA : Jurnal Ekonomi*, *3*(4).

Zou, H. (2020). Clustering Algorithm and Its Application in Data Mining. *Wireless Personal Communications*, *110*(1), 21–30. https://doi.org/10.1007/s11277-019-06709-z