International Journal of Research Publication and Reviews, Vol 6, Issue 6, pp 850-857 June 2025

WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Application of K-Harmonic Means and K-Medoids Methods in Grouping Districts/Cities in Central Java Based on Indicators for Compiling the Community Literacy Development Index

***Rosita Maela Sari[1], Iut Tri Utami[2]*, Alan Prahutama[3]***

[1,2,3]Universitas Diponegoro, Jl. Prof. Soedarto No. 13, Tembalang 50275, Semarang
*triutami.iut@gmail.com

**A B S T R A C T**

Literacy levels in Indonesia are low and one of the reasons is the poor quality of literacy facilities. Grouping the six indicators that make up the Index of Community Literacy Development (IPLM) aims to identify patterns and gaps so that more effective improvement strategies can be created. Java province, which consists of 35 districts/cities, was used because its IPLM is still below Indonesia's IPLM in 2023. This clustering was done with 2 methods, namely K-Harmonic Means and K-Medoids, then validated using Silhouette Coefficients. The data used has met the detection of non-multicollinearity. The results show that both methods produce an optimal cluster count of 4, with Silhouette values of 0,347639 for K-Medoids and 0,3521074 for K-Harmonic Means, making K-Harmonic Means the better approach. Cluster profiling reveals that cluster 3 consists of 4 districts/cities, with most variables classified as very low, while cluster 4 includes 6 districts/cities, where most variables are superior. The government can create a more effective and targeted strategy for improving literacy facilities according to the characteristics of each region.

Keywords: Literacy; K-Harmonic; K-Medoids; Silhouette

## 1. Introduction

Indonesia has a very low reading interest. A survey conducted by UNESCO on global literacy stated that the reading interest of Indonesians is only 0.001%, meaning only one out of 1,000 Indonesians regularly reads (Perpusnas, 2021). According to Badan Bahasa (2023), one of the factors contributing to the low literacy rate in Indonesia is the poor quality of literacy facilities, such as libraries, book collections, and library personnel. Indonesia has a Community Literacy Development Index (IPLM) as a reference to assess whether the facilities, infrastructure, and literacy services are adequate.

The IPLM consists of seven components: adequacy of library personnel, equitable distribution of library services, community participation in promotional/socialization activities, adequacy of collections, library members, daily community attendance rate, and number of libraries that meet the National Education Standards (SNP) (Perpusnas, 2023). Grouping districts/cities based on IPLM indicators can be used by the government as a reference in allocating funds to improve literacy facilities in each region.

This research focuses on grouping districts/cities based on IPLM indicators in Central Java Province, as it ranks third in population density in Indonesia, with 35 districts/cities. Many educational and economic activities are also centered in Central Java. The IPLM value for Central Java in 2023 was 64.4, which is still below the national IPLM average of 68.19 that year. Therefore, further research is needed to identify areas that should be prioritized in improving literacy facilities. Grouping based on IPLM indicators can help the government allocate resources more effectively.

Clustering or cluster analysis aims to group certain data objects into clusters that share similar characteristics. The IPLM indicators will be clustered using two methods: K-Medoids (Partitioning Around Medoids) and K-Harmonic Means, as both are improved versions of earlier methods such as K-Means, and offer better performance compared to conventional clustering approaches. The K-Harmonic Means algorithm, proposed by Zhang and later modified by Hammerly, is a center-based clustering method similar to K-Means but with the advantage of handling the initialization sensitivity of K-Means (Yang et al., 2009). The PAM algorithm, also known as K-Medoids, was developed by Peter J. Rousseeuw and Leonard Kaufman. It is similar to K-Means in that both use a partitioning algorithm to divide datasets into clusters (Wira et al., 2019). However, K-Medoids is considered more stable because it uses actual data points (medoids) as centroids and includes an improvement process during iteration, which avoids getting trapped in random medoid initialization (Bintoro et al., 2024).

Each resulting cluster will be validated to determine the optimal method and optimal number of clusters. The validation method used is the Silhouette Coefficient; a value closer to 1 indicates better cluster structure, and vice versa. This research aims to identify the optimal method and cluster configuration based on Silhouette Coefficient comparisons, so the results can be used by the government and local communities to improve literacy facilities and infrastructure, which in turn may lead to more evenly distributed literacy improvement across all regions.

## 2. Literature Review

Literacy can be defined as the ability to read and write, as well as textual literacy in writing and reading. In terms of its application, literacy refers to the integration of listening, reading, writing, speaking, and critical thinking skills (Purwati, 2017). According to the National Library (Perpusnas, 2023), the Community Literacy Development Index (IPLM) is a benchmark for efforts made by local governments (districts/cities and provinces) in building and improving the quality of libraries as lifelong learning centers to promote a culture of literacy in society. The IPLM consists of seven elements: adequacy of library staff, equitable distribution of library services, community involvement in promotional/socialization activities, adequacy of collections, library membership, daily community attendance rate, and the number of libraries that meet the National Education Standards (SNP) (Perpusnas, 2023). The measurement of IPLM follows the formula below (Dinas Arsip & Perpustakaan Kota Semarang, 2023):

$$\text{IPLM} = \sum_{q=1}^{7} \frac{UPLM_q}{AM} \times 100 \qquad (1)$$

$UPLM_q$: the value of the $q$ element of community literacy, $q$ : 1,2,3,…,7; dan $AM$ : the number of population in the local area

Cluster analysis is a data analysis method aimed at grouping objects or members into several clusters, each having its own characteristics, so that all members within the same cluster share similar characteristics (Talakua et al., 2017). In general, clustering methods are divided into two categories: non-hierarchical and hierarchical methods.

There are two assumptions in cluster analysis: the sample must be representative, and multicollinearity must not be present. Using a representative sample will produce optimal results and match the actual population conditions. When population data is used, it can be assumed that the representative sample test has already been satisfied (Hair et al., 1998). Multicollinearity is an issue that must be addressed because it can influence clustering results by indicating a linear relationship between independent variables, making it difficult to determine the influence of each variable. Multicollinearity can be detected by calculating the Variance Inflation Factor (VIF). If the VIF value is less than 10, the variable does not suffer from multicollinearity and meets the assumption.

$$\text{VI}F_j = \frac{1}{1 - R_j^2} \qquad (2)$$

$$R_j^2 = \frac{JKR}{JKT} = \frac{\sum_{i=1}^{n}\left(\widehat{x_{ij}} - \bar{x}_j\right)^2}{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2} \qquad (3)$$

where:

$VIF_j$: *Variance Inflation Factor* of the-$j$, $j$ : observed variabels $j = 1,2,3,\dots,m$ ($m =$ number of observed objects), $i$ : observed object $i = 1,2,3,\dots,n$ ($n =$ number of observed), $R_j^2$ : coefficient of determination of the jth variable $j$-th, JKR : Regression Sum of Squares, JKT: Total Sum of Squares, $x_{ij}$ *:* value of the $i$-$th$ observed object on the $j$-th variable, $\widehat{x_{ij}}$*:* estimated value of x for the i-th observed object data on the j-th variable, and $\bar{x_j}$ *:* average value of the $j$-th variable.

Measures of central tendency, also called averages, are values that represent a group or series of statistical data, often referred to as central or typical values. There are several types of averages in statistics, including arithmetic mean, weighted mean, geometric mean, and harmonic mean. The harmonic mean formula is as follows (Wirawan, 2016):

$$\bar{x}_{harmonik} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad (4)$$

$\bar{x}_{harmonik}$ : harmonic mean dan $x_i$ : data value for the $i$-th object.

K-Harmonic Means Clustering is a center-based clustering method that uses the harmonic mean of distances from data objects to the centroid. This method was developed to address the weaknesses of K-Means, whose clustering results are highly sensitive to the initialization of the starting centroids. The stages of cluster analysis using K-Harmonic Means (Widiartha, 2011) are:

1. Choose the number of clusters (K) to be created
2. Select the input parameter (p), tolerance value, and maximum number of iterations
3. Initialize the centroids randomly from each cluster, using random values within the minimum and maximum of the original data
4. Calculate the distance of each data point to each centroid using the Euclidean distance formula:

$$d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k}) = \sqrt{\sum\nolimits_{j=1}^{m} \left(x_{ij} - c_{kj}\right)^2} \qquad (5)$$

where:

$d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k})$ : *euclidean* distance of the $i$-th object with the $k$-th cluster center, $x_{ij}$ : object on the $i$-th on the $j$-th variable, $c_{kj}$ : the $k$ cluster center on the $j$-th variable, and $k$ : *cluster* order with $k$ : $1,2,3,\dots,K$ ($K =$ number of *cluster*).

5. Calculate the objective function value using the following formula:

$$\text{KH}M_t = \sum_{i=1}^{n} \left( \frac{K}{\sum_{k=1}^{K} \left( \frac{1}{(d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k}))^p} \right)} \right) \qquad (6)$$

where:

$KHM_t$ : objective function of the $t$-th iteration, $t$ : iteration generated with $t = 1,2,3,\dots,T$ ($T =$ number of iterations), and $p$ : parameters that have been determined at the beginning.

6. Calculate the membership value of each object in each cluster $m(\boldsymbol{x_i}, \boldsymbol{c_k})$. using the formula:

$$\text{m}(\boldsymbol{x_i}, \boldsymbol{c_k}) = \frac{d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k})^{-p-2}}{\sum_{k=1}^{K} \left(d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k})\right)^{-p-2}} \qquad (7)$$

where:

$m(\boldsymbol{x_i}, \boldsymbol{c_k})$ : membership value of the $i$-th object in the $k$ cluster.

7. Calculate the weight of each object using the formula:

$$\text{w}(\boldsymbol{x_i}) = \frac{\sum_{k=1}^{K} \left(d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k})\right)^{-p-2}}{\left(\sum_{k=1}^{K} \left(d_{euc}(\boldsymbol{x_i}, \boldsymbol{c_k})\right)^{-p}\right)^2} \qquad (8)$$

where:

$w(\boldsymbol{x_i})$ : weight value of object-$i$

8. Compute the new centroid based on the membership and weights of each object using:

$$\text{cb}(\boldsymbol{c_k}, \boldsymbol{x_j}) = \frac{\sum_{i=1}^{n} m(\boldsymbol{x_i}, \boldsymbol{c_k})\, w(\boldsymbol{x_i}) x_{ij}}{\sum_{i=1}^{n} m(\boldsymbol{x_i}, \boldsymbol{c_k})\, w(\boldsymbol{x_i})} \qquad (9)$$

where:

$cb(\boldsymbol{c_k}, \boldsymbol{x_j})$ : new *centroid* at the $k$-th cluster for the $j$-th variable

9. Repeat steps 4 to 8 until convergence, when the objective function no longer changes significantly or when:

$$|Objective\ Function_t - Objective\ Function_{t-1}| < \varepsilon \qquad (10)$$

where:

$\varepsilon$ : tolerance limit that has been determined at the beginning

10. Final cluster assignment is based on the highest membership value

K-Medoids or Partitioning Around Medoids (PAM) is a partitioning clustering method used to divide or group data into several clusters. Medoids represent the cluster in K-Medoids, and are specific data points that represent a cluster (Rahmah et al., 2022). K-Medoids is similar to K-Means, with the difference being that K-Means uses the mean as the centroid, while K-Medoids uses specific data points (medoids). The steps in K-Medoids clustering (Wira et al., 2019) are:

1. Determine the value of K, the number of clusters to be created
2. Randomly select K initial medoids from the set of n objects
3. Calculate the distance of each object to the medoids using the Euclidean formula (Equation 5), assign objects to the nearest medoid, and compute the total distance

4. Randomly select another set of K objects from the n objects as candidate new medoids
5. Recalculate the distance of each data object to the new candidate medoids using the Euclidean distance formula, then reassign each object to the nearest cluster and compute the total distance again
6. Calculate the total distance difference using:

$$\text{STJ} = TJM_{new} - TJM_{old} \quad (11)$$

Where:

$STJ$ : total distance difference, $TJM_{new}$ : total distance of new medoids, dan $TJM_{old}$: Total distance of old medoids

7. If $STJ < 0$, the iteration continues, and the new candidate medoids replace the old ones for the next calculation. If STJ > 0, the iteration stops (converges).
8. Repeat steps 4 to 7 until convergence, if $STJ > 0$

The Silhouette Method is a validation method useful for assessing the strength and quality of a cluster. The formula for Silhouette Coefficient (Gunawan, 2020) is:

$$\text{a}(i) = \frac{1}{A-1} \sum_{h \neq i; h \epsilon A}^{H} d_{euc}(\boldsymbol{x_i}, \boldsymbol{x_h}) \quad (12)$$

$$\bar{d}_{euc}(\boldsymbol{x_i}, \boldsymbol{B}) = \frac{1}{B} \sum_{u \neq i, u \epsilon B}^{U} d_{euc}(\boldsymbol{x_i}, \boldsymbol{x_u}) \quad (13)$$

$$b(i) = min\left(\bar{d}_{euc}(\boldsymbol{x_i}, \boldsymbol{B})\right) \quad (14)$$

$$\text{s}(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (15)$$

where:

$a(i)$ : average distance of object $i$ to other objects in the same cluster, $A$ : number of objects in cluster A, $d_{euc}(\boldsymbol{x_i}, \boldsymbol{x_h})$ : *euclidean* distance

between object $i$ and object $h$, $h$ : object other than $i$ in *cluster* A with $h = 1,2,3,\ldots,H$; $i \neq h$ ($H =$ number of objects other than $i$ in *cluster* A), $\bar{d}_{euc}(\boldsymbol{x_i}, \boldsymbol{B})$ : average distance of object $i$ from *cluster* A with all data in different cluster, namely *cluster* B, $B$ : number of objects in *cluster* B, $d_{euc}(\boldsymbol{x_i}, \boldsymbol{x_u})$ : *euclidean* distance between object $i$ and $u$, $u$ : objects in different *cluster*, namely cluster B with $u = 1,2,3,\ldots,U$; $i \neq u$ ($U =$ number of objects in different, namely *cluster* B), $b(i)$ : minimum value of average distance of object $i$ from *cluster* A with all data in different cluster, namely *cluster* B, and $s(i)$ : *Silhouette Coefficients* value

A Silhouette Coefficient closer to 1 indicates higher cluster quality; the lower the value, the poorer the cluster quality. According to Rousseeuw (1987), the interpretation of Silhouette Coefficients is as follows:

Tabel 1. Silhouette Coefficients Interpretation

| Silhouette Coefficient | Interpretation |
|---|---|
| $0,7 < SC \leq 1$ | Very Strong Structure |
| $0,5 < SC \leq 0,7$ | Strong Structure |
| $0,25 < SC \leq 0,5$ | Weak Structure |
| $SC \leq 0,25$ | Poor Structure |

## 3. Research Methodology

This study uses secondary data sourced from the Central Java Provincial Statistics Agency (BPS). The data collected consist of six indicators that compose the Community Literacy Development Index (IPLM), based on districts/cities in Central Java Province for the year 2023. This data was updated by BPS Central Java on February 26, 2024. The variables used are: Adequacy of Library Personnel, Equity of Library Services, Community Participation in Promotion/Socialization Activities, Adequacy of Collections, Daily Community Visit Rate, and Number of Libraries Meeting the National Education Standards (SNP). These variables are measured in the form of an index. Clustering analysis in this study is carried out using two methods: K-Harmonic Means and K-Medoids, and the optimal method and cluster count are validated using the Silhouette Coefficients. The data is processed manually using Microsoft Excel and computationally using R Studio. The steps of the data analysis are as follows:

1. Data Input

2. Assumptions in Cluster Analysis

The representativeness test is not conducted because the dataset represents the population, thus analysis continues directly to multicollinearity detection by calculating the VIF (Variance Inflation Factor). If all VIF values are less than 10, analysis may proceed. If any variable has a VIF value greater than 10, Principal Component Analysis should be performed to resolve multicollinearity.

3. Cluster Analysis using K-Harmonic Means

   a. Determine the number of clusters $(K)$ to be formed

This study applies a trial-and-error approach, testing values of $K = 2$ s.d $10$ for computational calculations, while for manual calculations, only $K = 3$ is used for time efficiency.

   b. Determine the input parameter value $(p)$

This study used $p = 2$ because K-Harmonic Means works better with $p \geq 2$ values (Gungor et al., 2008), but was still tested for $p = 3$ dan $p = 4$.

   c. Set the tolerance value $\varepsilon = 10^{-4}$ and maximum number of iterations to 100.

This study uses a tolerance of $\varepsilon = 10^{-4}$ because stable clustering results were obtained for $10^{-6}, 10^{-5}, 10^{-4}$, so the largest exponent was chosen for program running efficiency.

   d. Initialize the initial centroids randomly from each cluster, using random values within the range of the original dataset.
   e. Calculate the distance from each data point to each centroid using the Euclidean distance formula (Equation 5).
   f. Compute the objective function using Equation 6.
   g. Calculate the membership value of each object in each cluster using Equation 7.
   h. Compute the weight value of each object using Equation 8.
   i. Calculate new centroids based on membership values and object weights using Equation 9.
   j. Repeat steps "e" to "i" until the iteration converges, the objective function becomes stable, or the stopping condition based on Equation 10 is met.
   k. Final cluster assignment is based on the highest membership value.

4. Cluster Analysis using K-Medoids

   a. Determine the number of clusters $K$ to be formed
   b. Randomly select initial medoids from the data objects
   c. Calculate the distance from each object to the initial medoids using Euclidean distance (Equation 5), assign each object to the nearest medoid, and calculate the total distance
   d. Select again $K$of $n$ objects at random as new candidate medoids
   e. Recalculate the distance from each object to the new candidate medoids using Euclidean distance (Equation 5), assign objects to clusters, and compute the total distance again
   f. Calculate the difference in total distance using Equation 11
   g. If the $STJ\ value\ is < 0$, then the iteration continues or the new candidate medoids are transformed into new medoids for the next calculation, and if $STJ > 0$ then the iteration stops/converges
   h. Repeat steps d to g until the iteration stops, that is, when the convergence condition is met $STJ > 0$

5. Determine the Optimal Method and Number of Clusters using Silhouette Coefficients

   a. Calculate the Silhouette Coefficients for each resulting cluster
   b. Compare the Silhouette Coefficients across cluster count, and select the optimal clustering result based on the value closest to 1

## 4. Results and Discussion

The assumptions in cluster analysis consist of two stages. The representativeness test is not conducted because the data used represent the population. Therefore, the analysis proceeds directly to multicollinearity detection. This detection is conducted using the VIF (Variance Inflation Factor), calculated with Equation 2. Table 2 presents the VIF values calculated using R Studio syntax.

Table 1. VIF Values

| **Variable** | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| **VIF** | 1,421262 | 1,456876 | 1,118693 | 1,185677 | 1,095662 | 1,231982 |

The decision based on Table 2 is that the VIF values for all variables are less than 10. It can be concluded that there is no multicollinearity among the variables used. The clustering results obtained using K-Harmonic Means and K-Medoids must then be compared using a validation test. The purpose is to determine the best number of clusters and the optimal method used in this study. Table 3 shows the validation results for the K-Harmonic Means and K-Medoids clusters.

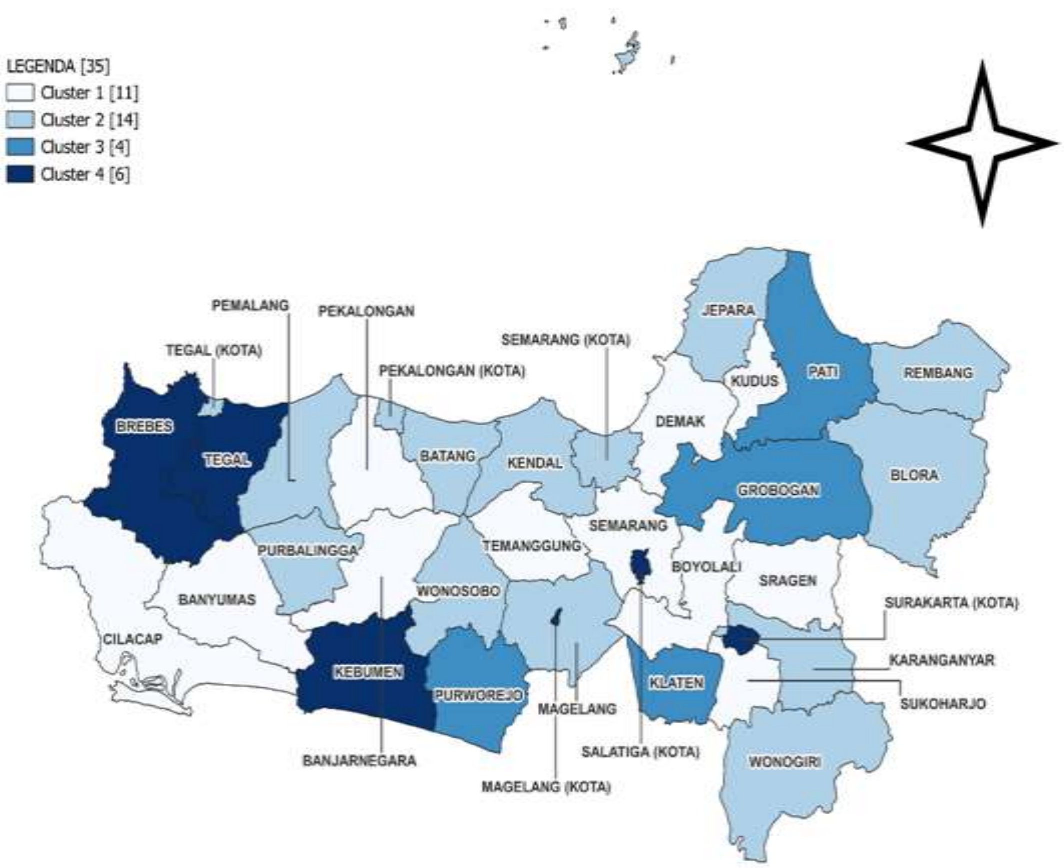
Table 1. Best *Cluster* Validation

| **Method** | ***Cluster*** | ***Silhouette Coefficients*** |
|---|---|---|
| ***K-Harmonic Means*** | 2 | 0,2489092 |
| | 3 | 0,2930317 |
| | 4 | **0,3521074** |
| | 5 | 0,2826711 |
| | 6 | 0,3136869 |
| | 7 | 0,2923313 |
| | 8 | 0,2400378 |
| | 9 | 0,307633 |
| | 10 | 0,3162918 |
| ***K-Medoids*** | 2 | 0,2335065 |
| | 3 | 0,3038852 |
| | 4 | **0,347639** |
| | 5 | 0,34489 |
| | 6 | 0,2936645 |
| | 7 | 0,3116497 |
| | 8 | 0,312141 |
| | 9 | 0,3186361 |
| | 10 | 0,3228949 |

Based on Table 3, the optimal cluster count for both K-Harmonic Means and K-Medoids is 4 clusters, with the highest Silhouette Coefficient values. K-Harmonic Means yields a coefficient of 0.3521074, while K-Medoids yields 0.347639, indicating that K-Harmonic Means performs slightly better. Cluster profiling is conducted to examine the characteristics of each variable within each cluster. These characteristics are analyzed based on the average values. Table 4 presents the average values of each variable in each cluster.

Table 1. Average Variable Values in Each Cluster

| **Variable** | ***Cluster* 1** | ***Cluster* 2** | ***Cluster* 3** | ***Cluster* 4** |
|---|---|---|---|---|
| Equity of Library Services | 0,6045182 | 0,6971929 | 0,4956750 | 0,7279000 |
| Adequacy of Library Collections ($X_2$) | 0,3633182 | 0,3036500 | 0,1835750 | 0,5934667 |
| Adequacy of Library Personnel ($X_3$) | 0,4048727 | 0,1464643 | 0,1720000 | 0,9982500 |
| Daily Community Visit Rate ($X_4$) | 0,1460000 | 0,1749571 | 0,0416750 | 0,1782000 |
| Libraries Meeting National Education Standards ($X_5$) | 0,2370727 | 0,9558929 | 0,8098500 | 0,8722167 |
| Community Participation in Library Promotion Activities ($X_6$) | 0,9719727 | 0,9511786 | 0,1868250 | 0,9659833 |




The characteristics of the variables in each cluster are classified into four categories: very low, low, moderate, and high. The profiling based on the averages in Table 4 is as follows:

**Cluster 1** is high only on the Community Participation in Library Promotion indicator ($X_6$), but very low in Libraries Meeting SNP ($X_5$). It is low in Equity of Library Services ($X_1$), and Daily Community Visit Rate ($X_4$), while moderate in Adequacy of Collections ($X_2$), and Adequacy of Personnel Perpustakaan ($X_3$). Districts/Cities in this cluster include Cilacap, Banyumas, Banjarnegara, Boyolali, Sukoharjo, Sragen, Kudus, Demak, Semarang, Temanggung, and Pekalongan.

**Cluster 2** is high in Libraries Meeting SNP ($X_5$), but very low in Adequacy of Library Personnel $X_3$). It is moderate in Equity of Library Services ($X_1$) and Daily Community Visit Rate ($X_4$), and low in Adequacy of Collections ($X_2$) and Community Promotion Activities ($X_6$). Districts/Cities include Purbalingga, Wonosobo, Magelang, Wonogiri, Karanganyar, Blora, Rembang, Jepara, Kendal, Batang, Pemalang, Semarang City, Pekalongan City, and Tegal City.

**Cluster 3** is very low in Equity of Library Services ($X_1$), Adequacy of Collections ($X_2$), Daily Community Visit Rate ($X_4$), and Community Promotion Activities ($X_6$). It is low in Adequacy of Library Personnel ($X_3$) and Libraries Meeting SNP ($X_5$). Districts/Cities in this cluster include Purworejo, Klaten, Grobogan, and Pati.

**Cluster 4** is high in Equity of Library Services ($X_1$), Adequacy of Collections ($X_2$), Adequacy of Personnel ($X_3$), and Daily Visit Rate ($X_4$). It is moderate in Libraries Meeting SNP ($X_5$) and Community Promotion Activities $X_6$). Districts/Cities include Kebumen, Tegal, Brebes, Magelang City, Surakarta City, and Salatiga City.

## 5. Conclusion

Based on the analysis and explanation in the previous chapter, it can be concluded that the optimal cluster in the K-Harmonic Means method is at $K = 4$ with a Silhouette Coefficient of 0,3521074. Similarly, the optimal cluster in the K-Medoids method is also at $K = 4$ with Silhouette Coefficient of 0,347639. Therefore, in this study, the optimal method chosen is K-Harmonic Means.

Based on the clustering profile results, it can be concluded that Cluster 3 consists of 4 districts/cities with most variables classified in the very low category. Cluster 4 consists of 6 districts/cities with most variables in the high category. Meanwhile, the characteristics of variables in Cluster 1 and Cluster 2 are more evenly distributed across the four categories (high, moderate, low, and very low). Cluster 1 includes 11 districts/cities, and Cluster 2 consists of 14 districts/cities. The government and local communities can identify patterns and gaps and determine more effective and targeted strategies for improving literacy facilities according to the characteristics of each region

**References**


Badan Pengembangan & Pembinaan Bahasa. (2023). *Kondisi literasi indonesia yang sedang tidak baik-baik saja*. Retrieved from: https://badanbahasa.kemdikbud.go.id/berita-detail/3917/kondisi-literasi-indonesia-yang-sedang-tidak-baik-baik-saja (accessed October 23, 2024).

Badan Pusat Statistik Provinsi Jawa Tengah. (2024). *Indeks Pembangunan Literasi Masyarakat dan Unsur Penyusunnya Menurut Kabupaten/Kota di Provinsi Jawa Tengah*. Retrieved from https://jateng.bps.go.id/id/statistics-table/2/MjA0NSMy/indeks-pembangunan-literasi-masyarakat-dan-unsur-penyusunnya-menurut-kabupaten-kota-di-provinsi-jawa-tengah.html (accessed October 23, 2024).

Bintoro, Panji et al. (2024). *Pengantar machine learning*. Sumatera Barat: PT Mafy Media Literasi Indonesia.

Dinas Arsip & Perpustakaan Kota Semarang. (2023). *Kajian pengukuran indeks pembangunan literasi masyarakat (iplm) kota Semarang*.

Gunawan, Kevin Chandra. (2020). Analisis peserta online test dengan pembentukan cluster dengan algoritma k-means dan analisis tabulasi. *Jurnal Titra*, Vol. 8, No. 2, pp. 353–360.

Gungor, Z., & Alper, U. (2008). K-Harmonic Means Data Clustering with Tabu-search Method, Vol. 32, No. 6, pp. 1115–1125.

Hair, J.F., et al. (1998). *Multivariate Data Analysis* (5th ed.). New Jersey: Prentice-Hall.

Perpustakaan Nasional. (2021). *Kepala Perpusnas: Kondisi Literasi Harus Dipahami dari Hulu hingga Hilir*. Retrieved from https://www.perpusnas.go.id/berita/kepala-perpusnas:-kondisi-literasi-harus-dipahami-dari-hulu-hingga-hilir (accessed February 3, 2025).

Perpustakaan Nasional. (2023). *Indeks Pembangunan Literasi Masyarakat (IPLM) 2023*. Retrieved from https://dev-satudata.perpusnas.go.id/dataset/indeks-pembangunan-literasi-masyarakat-iplm-2023 (accessed October 23, 2024).

Purwati, S. (2017). Program literasi membaca 15 menit sebelum pelajaran dimulai untuk meningkatkan hasil belajar membaca dan menghafal surah pendek. *Jurnal Ilmu Pendidikan Sosial, Sains, dan Humaniora*, 3(4), 663–670.

Rahmah, E., Elin H., Alwis, N., & Sitti, R. (2022). *Penerapan algoritma k-medoids clustering untuk menentukan strategi promosi pada data mahasiswa (studi kasus: STIKES Perintis Padang)*. *Jurnal Nasional Komputasi dan Teknologi Informasi*, Vol. 5, No. 3, pp. 556–564.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65.

Talakua, M. W., Leleury, Z. A., & Talluta, A.W. (2017). Analisis cluster dengan menggunakan metode k-means untuk pengelompokkan kabupaten/kota di provinsi Maluku berdasarkan indikator indeks pembangunan manusia tahun 2014. *Jurnal Ilmu Matematika dan Terapan*, Vol. 11, No. 2, pp. 119–128.

Widiartha, I. Made. (2011). Studi komparasi metode klasterisasi data k-means dan k-harmonic means. *Jurnal Ilmu Komputer*, Vol. 4, No. 1.

Wira, Bagus, Alexius Endy Budianto, & Anggri Sartika Wiguna. (2019). Implementasi metode k-medoids clustering untuk mengetahui pola pemilihan program studi mahasiswa baru tahun 2018 di Universitas Kanjuruhan Malang. *Jurnal Terapan Sains & Teknologi*, Vol. 1, No. 3, pp. 54–69.

Wirawan, Nata. (2016). *Cara Mudah Memahami Statistika Ekonomi dan Bisnis (Statistika Deskriptif)*. Denpasar: Keraras Emas.

Yang, Fengqin, Sun, Tieli, & Zhang, Changhai. (2009). An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization, Vol. 36, No. 6, pp. 9847–9852.