International Journal of Science and Technology Research Archive

ISSN: 0799-6632 (Online)

Journal homepage: https://sciresjournals.com/ijstra/

(RESEARCH ARTICLE)

Check for updates

Comparing K-Means and K-medoids algorithms for clustering hamlet regions by tax liabilities in tax determination documents

Firsta Rahmania Sucahyo [1,*], Indyah Hartami Santi [1], Mohammad Faried Rahmat [1] and Diki Fahrizal [2]

[1] *Information Technology, Informatics Engineering, Islamic University Balitar, Blitar, Indonesia.*
[2] *Magister of Electrical Engineering, School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung, Indonesia.*

International Journal of Science and Technology Research Archive, 2025, 08(01), 069-078

Publication history: Received on 25 December 2024; revised on 01 February 2025; accepted on 04 February 2025

Article DOI: https://doi.org/10.53771/ijstra.2025.8.1.0023

Abstract

The application of data mining information technology in Village Offices, especially in village office administration services, is very important to ensure the efficiency and accuracy of information services. This research aims to compare the effectiveness of the K-Means and K-Medoids algorithms in clustering hamlet areas based on the tax owed in the tax assessment documents in Pandanarum village. Using quantitative descriptive methods, the two algorithms are applied to group hamlets based on tax payable data as the main variable. The clustering process is analyzed using evaluations such as Sum of Squared Errors (SSE) and Silhouette Score to determine the effectiveness of each algorithm. The research results show that the K-Medoids algorithm has lower performance compared to K-Means, especially in terms of cluster stability and a high Silhouette Score value with a value of 0.454615 and SSE 480.9462. Apart from that, the K-Medoids algorithm is more robust against outliers in the tax payable data, and produces a lower Silhoutte Score value with a value of 0.382616 and an SSE of 567.6125 which indicates weaker clustering. Thus, this research concludes that the K-Means algorithm is superior in clustering hamlet areas based on taxes owed compared to the K-Medoids algorithm.

Keywords: Clustering; K-Means; K-Medoids; Tax Payable; Tax Assessment Document

1 Introduction

The application of data mining technology in village offices, particularly in administrative services, is crucial for ensuring efficiency and accuracy in information management. This is particularly relevant for the management of Rural and Urban Land and Building Taxes (PBB), which are imposed on land and buildings owned, controlled, or utilized by individuals or organizations, except for areas used for plantation, forestry, or mining activities, as regulated by Law No. 28 of 2009 on Regional Taxes and Levies. According to this law, "land" refers to the earth's surface, including land, inland waters, and territorial seas, while "buildings" refer to structures permanently affixed to land or water. Taxes serve two primary functions: revenue collection and regulatory control. The government uses taxes to generate funds for various purposes, and the budgetary function necessitates that citizens comply with their tax obligations. The level of compliance reflects how well taxpayers adhere to the regulations in their respective areas. The volume of data stored is proportional to the number of taxpayers in the region [1], [2].

In Pandanarum Village, land and building tax data is recorded in the Tax Assessment Record Book (DHKP), which covers 4,576 taxpayers, classified into two categories: building tax and land tax. The DHKP contains records of every house and landowner to manage the tax obligations for each property. The data is updated annually. Additionally, the Village Land Register, known as "Letter C," provides details on land ownership in Indonesia. Issued by the National Land Agency and typically held by the village, the "Letter C" book identifies actual landowners, including property descriptions, location,

* Corresponding author: Firsta Rahmania Sucahyo

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

plot size, boundaries, ownership number, and tax registration number. Pandanarum Village covers 369 hectares and has a population of 8,900, with 2,561 households [3], [4].

Observations conducted in December 2023 revealed that the "Letter C" book and DHKP are still manually recorded, leading to disorganized and unsorted data on taxpayer regions, which causes delays in retrieving information on taxpayers with outstanding taxes. This highlights the need for a clustering method. Clustering is a data mining technique used to group data with similar characteristics into one cluster while separating those with different characteristics into other clusters [5], [6]. The purpose of clustering in this study is to identify patterns and group hamlet areas to facilitate the village office staff in managing taxpayer regions with outstanding taxes [7].

K-Means and K-Medoids algorithms are employed for clustering in this research. K-Means is faster and computationally more efficient, as it uses the mean calculation to determine the cluster centroid [8], while K-Medoids aims to minimize absolute error criteria and is more resistant to outliers by selecting a medoid [9], an actual member of the dataset, as the cluster center [10]. This makes it more representative, especially when variables such as "Outstanding Taxes" or "Age" have a wide or skewed distribution. The goal of this comparison is to determine the most suitable algorithm for clustering taxpayer regions, improving service efficiency for village office staff .

Both K-Means and K-Medoids are known for their accuracy and efficiency when dealing with large datasets [11], [12]. Mardalius notes that the flexibility of K-Means and K-Medoids allows users to choose the number of clusters to generate [13], [14]. The study evaluates clustering quality using the Silhouette Coefficient method and the Elbow method. The clustering results will classify taxpayer regions into three clusters based on outstanding taxes: high, medium, and low tax clusters [15].

Previous research has been conducted on similar topics, such as the implementation of K-Means and K-Medoids for clustering potential poultry production areas, and comparing the algorithms for mapping fruit production regions [16], [17]. This study aims to compare the effectiveness of the K-Means and K-Medoids algorithms in clustering taxpayer regions based on outstanding taxes in Pandanarum Village's tax assessment documents.

2 Methods

The research methodology comprises four main stages: Pre-Research, Data Mining Methodology (Knowledge Discovery in Databases), Data Mining Process, and Results. The tools utilized in this research include Microsoft Excel and Python for data processing and analysis [18].

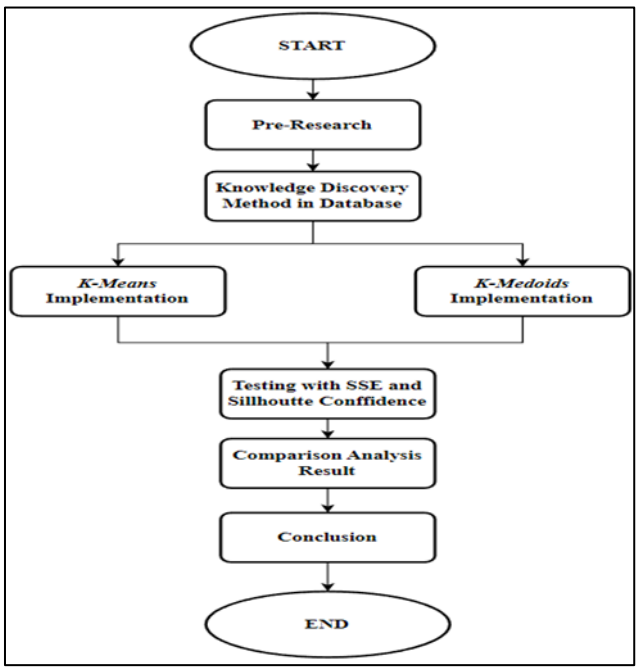


Figure 1 Research method flowchart

2.1 Pre-Research

- **Data Collection**: The data collection process was carried out through observation, interviews, and literature review. Observations and interviews were conducted at the Pandanarum Village Office to gather relevant information, while the literature review provided additional theoretical support for the study[14].
- **Data Selection**: The primary data source used in this research is the Tax Assessment Record Book (DHKP), which includes variables such as the tax identification number (NDP), taxpayer names, tax objects, taxable individuals, and outstanding taxes. Additionally, supporting demographic data from Pandanarum Village (including name, gender, age, religion, education, occupation, marital status, and address) was used. The relevant attributes were selected based on their importance to the clustering analysis.
- **Sample Data**: Using Slovin's formula, a sample size of 368 data points was derived from the initial DHKP population of 4,567 entries. These selected data points are used in the clustering process to ensure that the results are statistically representative of the population.
- **Elbow Method Application**: The elbow method was applied to determine the optimal number of clusters for the study. The silhouette score was subsequently used to evaluate the cohesion and separation of the clusters, helping to define the most appropriate number of clusters [19].

2.2 Data Mining Methodology (Knowledge Discovery in Databases)

- **Data Selection**: In this phase, variables from the DHKP and demographic data of Pandanarum Village were selected based on their relevance to the study. The variables used in the clustering of hamlet areas by outstanding taxes included tax owed, tax object, age, and education level.
- **Pre-Processing (Data Cleaning)**: The data cleaning process was conducted to remove errors, missing values, and duplicate records from both the DHKP and demographic datasets. This pre-processing step was crucial to ensuring the accuracy of the subsequent data mining steps and minimizing potential distortions during clustering.
- **Data Transformation**: The selected data were transformed from categorical to numerical formats. Variables that were originally in text or categorical form (such as education level and tax object) were converted into numerical values to facilitate the computational requirements of the K-Means and K-Medoids algorithms.

2.3 Data Mining Process

- **Data Mining (Clustering)**: After pre-processing and transforming the data, 368 samples from the DHKP and demographic data were clustered using both the K-Means and K-Medoids algorithms. These algorithms are applied to identify patterns in the data and group hamlet areas based on the outstanding taxes, tax objects, and other relevant demographic factors [20]. The K-Means algorithm uses the mean value to determine cluster centroids, while K-Medoids uses actual data points (medoids) to represent the clusters, making it more robust to outliers.
- **Clustering Evaluation**: The clustering results were evaluated using the Silhouette Coefficient to assess the quality and cohesion of the clusters. This step compared the performance of the K-Means and K-Medoids algorithms to identify which algorithm produced the most coherent clusters. The evaluation also helped determine the high, medium, and low outstanding tax clusters in the Pandanarum hamlet areas [21].

2.4 K-Means Algorithm

The K-Means algorithm is a widely used clustering method in data mining. It aims to partition a dataset into k clusters, where each data point belongs to the cluster with the nearest mean, serving as the cluster's centroid [22]. The following steps describe how K-Means was applied in this research:

- **Initialization of Centroids**. The first step in K-Means involves selecting k initial centroids, where k is the number of clusters determined by the elbow method (in this case, 3). These centroids represent the center of each cluster, initially chosen randomly from the dataset.
- **Assigning Data Points to the Nearest Centroid**. Each data point is assigned to the cluster whose centroid is closest. The Euclidean distance is typically used to calculate the distance between data points and centroids. In this research, variables such as outstanding tax, taxpayer object, and age were converted into numeric values, making it suitable for distance calculations [23].
- **Updating Centroids.** Once all data points are assigned to clusters, the algorithm recalculates the centroids by taking the mean of all the points in each cluster. These new centroids represent the updated centers of the clusters and will likely shift from their initial positions.

- **Reiteration of the Process.** Steps 2 and 3 are repeated iteratively. Data points are reassigned to the nearest new centroid, and centroids are updated based on the new assignments.The algorithm continues to iterate until the centroids no longer move significantly, indicating that the clusters have stabilized [24].

2.5 K-Medoids Algorithm

The K-Medoids algorithm, also known as the Partitioning Around Medoids (PAM) algorithm, is another clustering method similar to K-Means but more robust to outliers [25], [26]. Instead of using the mean of the points to determine the centroid, K-Medoids selects medoids, which are actual data points within the dataset, as the cluster centers. Here's how K-Medoids was applied in this research:

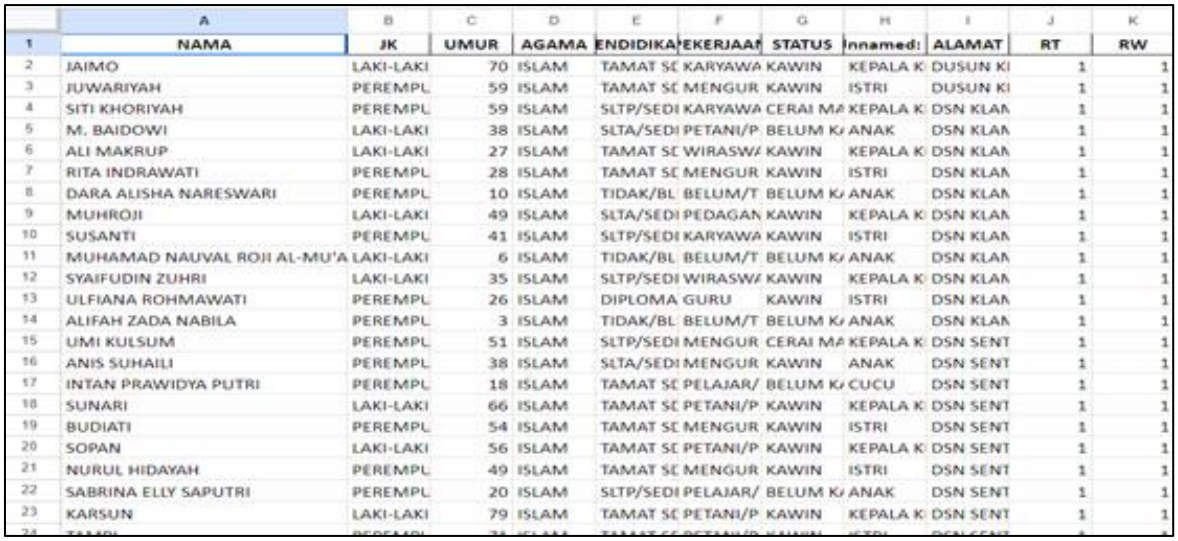
- Initialization of Medoids. Like K-Means, the K-Medoids algorithm begins by randomly selecting k medoids (in this case, 3) from the dataset. These medoids represent the most central points in each cluster, and each data point is initially assigned to the nearest medoid.
- Assigning Data Points to Medoids. Each data point is then assigned to the cluster corresponding to the nearest medoid, again based on the Euclidean distance or another distance metric. In this study, attributes such as outstanding tax, taxpayer object, and age are considered in assigning data points to clusters [27].
- Updating Medoids. Once all data points are assigned to clusters, the algorithm attempts to minimize the sum of absolute differences between the data points and their medoids by swapping the current medoids with other points in the dataset. If a swap reduces the total cost (i.e., the sum of distances between data points and medoids), the new point becomes the medoid of the cluster.
- Reiteration of the Process. This process is repeated iteratively. Medoids are recalculated, and data points are reassigned to the nearest medoids. The process continues until the algorithm converges, meaning that the medoids no longer change significantly [28].

3 Results and discussion

3.1 Data Collection

The data collected for this research was obtained from the Pandanarum Village Office, specifically from the Tax Assessment Record Book (DHKP), which includes taxpayer data. The dataset contains the following attributes: Tax Identification Number (NDP), taxpayer name, tax object, taxable individual, and outstanding taxes. In addition, supporting data was gathered from the demographic records of Pandanarum Village, including attributes such as name, gender, age, religion, education level, occupation, marital status, and address.

Figure 1 Hardcopy of Tax Document



	A	B	C	D	E	F	G	H	I	J	K
1	NAMA	JK	UMUR	AGAMA	ENDIDIKA	EKERJAAN	STATUS	Unamed:	ALAMAT	RT	RW
2	JAIMO	LAKI-LAKI	70	ISLAM	TAMAT SE	KARYAWA	KAWIN	KEPALA K	DUSUN KI	1	1
3	JUWARIYAH	PEREMPU	59	ISLAM	TAMAT SE	MENGUR	KAWIN	ISTRI	DUSUN KI	1	1
4	SITI KHORIYAH	PEREMPU	59	ISLAM	SLTP/SEDI	KARYAWA	CERAI MA	KEPALA K	DSN KLAN	1	1
5	M. BAIDOWI	LAKI-LAKI	38	ISLAM	SLTA/SEDI	PETANI/P	BELUM KA	ANAK	DSN KLAN	1	1
6	ALI MAKRUP	LAKI-LAKI	27	ISLAM	TAMAT SE	WIRASWA	KAWIN	KEPALA K	DSN KLAN	1	1
7	RITA INDRAWATI	PEREMPU	28	ISLAM	TAMAT SE	MENGUR	KAWIN	ISTRI	DSN KLAN	1	1
8	DARA ALISHA NARESWARI	PEREMPU	10	ISLAM	TIDAK/BL	BELUM/T	BELUM KA	ANAK	DSN KLAN	1	1
9	MUHROJI	LAKI-LAKI	49	ISLAM	SLTA/SEDI	PEDAGAN	KAWIN	KEPALA K	DSN KLAN	1	1
10	SUSANTI	PEREMPU	41	ISLAM	SLTP/SEDI	KARYAWA	KAWIN	ISTRI	DSN KLAN	1	1
11	MUHAMAD NAUVAL ROJI AL-MU'A	LAKI-LAKI	6	ISLAM	TIDAK/BL	BELUM/T	BELUM KA	ANAK	DSN KLAN	1	1
12	SYAIFUDIN ZUHRI	LAKI-LAKI	35	ISLAM	SLTP/SEDI	WIRASWA	KAWIN	KEPALA K	DSN KLAN	1	1
13	ULFIANA ROHMAWATI	PEREMPU	26	ISLAM	DIPLOMA	GURU	KAWIN	ISTRI	DSN KLAN	1	1
14	ALIFAH ZADA NABILA	PEREMPU	3	ISLAM	TIDAK/BL	BELUM/T	BELUM KA	ANAK	DSN KLAN	1	1
15	UMI KULSUM	PEREMPU	51	ISLAM	SLTP/SEDI	MENGUR	CERAI MA	KEPALA K	DSN SENT	1	1
16	ANIS SUHAILI	PEREMPU	38	ISLAM	SLTA/SEDI	MENGUR	KAWIN	ANAK	DSN SENT	1	1
17	INTAN PRAWIDYA PUTRI	PEREMPU	18	ISLAM	TAMAT SE	PELAJAR/	BELUM KA	CUCU	DSN SENT	1	1
18	SUNARI	LAKI-LAKI	66	ISLAM	TAMAT SE	PETANI/P	KAWIN	KEPALA K	DSN SENT	1	1
19	BUDIATI	PEREMPU	54	ISLAM	TAMAT SE	MENGUR	KAWIN	ISTRI	DSN SENT	1	1
20	SOPAN	LAKI-LAKI	56	ISLAM	TAMAT SE	PETANI/P	KAWIN	KEPALA K	DSN SENT	1	1
21	NURUL HIDAYAH	PEREMPU	49	ISLAM	TAMAT SE	MENGUR	KAWIN	ISTRI	DSN SENT	1	1
22	SABRINA ELLY SAPUTRI	PEREMPU	20	ISLAM	SLTP/SEDI	PELAJAR/	BELUM KA	ANAK	DSN SENT	1	1
23	KARSUN	LAKI-LAKI	79	ISLAM	TAMAT SE	PETANI/P	KAWIN	KEPALA K	DSN SENT	1	1

Figure 2 Softcopy of Tax Document

3.2 Data Sample

Based on the data obtained from the Tax Assessment Record Book (DHKP) containing 4,567 entries of taxpayer information, the sampling process was conducted to determine the dataset to be used for this research. The researcher applied Slovin's formula to draw a representative sample from the total population of DHKP data. By calculating the initial population of 4,567 entries using Slovin's formula, a sample size of 368 data points was determined to be sufficient for analysis in this study.

$$n = \frac{N}{1 + N \cdot e^2} \quad \text{.....} \quad (1)$$

$$n = \frac{4567}{1 + 4567 \cdot 0{,}05^2} = 368 \quad \text{.....} \quad (2)$$

3.3 Determining the Best Cluster Results

At this stage, the elbow method was used to determine the optimal number of clusters. The elbow method helps identify the most appropriate number of clusters by evaluating the total within-cluster sum of squared errors (SSE) and observing where the curve forms an "elbow."

From the sample data, the elbow method indicated that 3 clusters were the most suitable. This is evident from the elbow-like shape in the graph, where adding more clusters beyond three results in diminishing improvements to the SSE, suggesting that three clusters provide the best balance between complexity and performance.

The cluster labels were divided into three groups: Cluster 1, Cluster 2, and Cluster 3, based on the proximity of each data point to the nearest centroid (k). Each cluster represents a distinct grouping of data points with similar characteristics, as determined by the clustering algorithm.

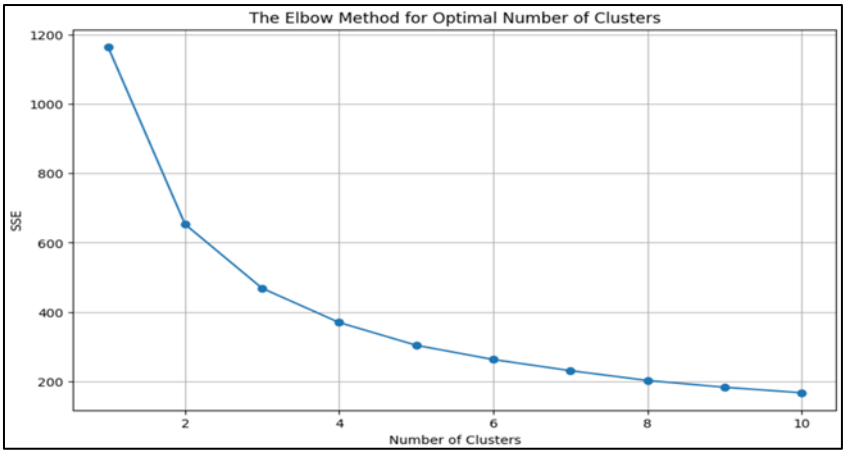


Figure 3 Elbow Method for Optimal Number

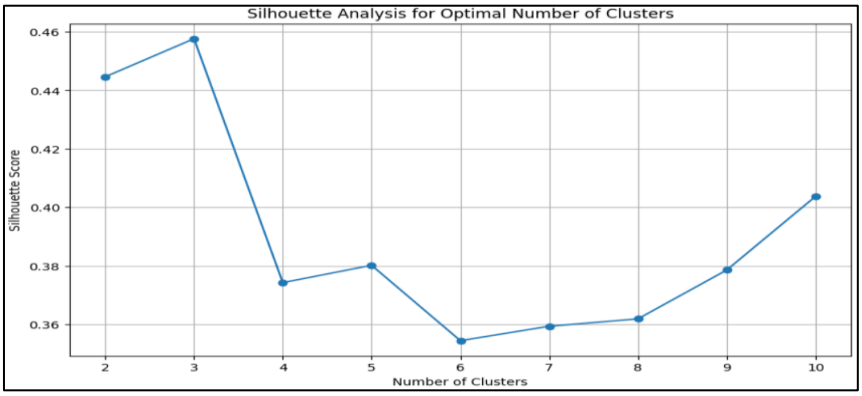


Figure 4 Silhouette Analysis

After determining the initial clusters using the elbow method, a silhouette score analysis was conducted to further evaluate the optimal number of clusters. The silhouette score measures how similar an object is to its own cluster compared to other clusters, with values ranging from -1 to 1. A higher silhouette score indicates better-defined clusters.

As seen in Figure 4.5, the silhouette analysis confirmed that the optimal number of clusters is 3, with a silhouette score of 0.45. This score indicates a moderate level of cohesion and separation between clusters, meaning that the data points within each cluster are fairly like one another, while the clusters themselves are sufficiently distinct from each other. Thus, three clusters were confirmed as the most appropriate grouping for this dataset.

3.4 K-Means Algorithm Evaluation

In this stage, the evaluation of the K-Means algorithm involves measuring two important metrics: Sum of Squared Errors (SSE) and the Silhouette Coefficient. These metrics are used to determine how well the data points are clustered and how distinct each cluster is from the others.

- **Sum of Squared Errors (SSE)** is a metric that calculates the squared distance between each data point and its corresponding cluster centroid. A lower SSE value indicates that the points within each cluster are closer to their centroid, reflecting a better fit.
- **Silhouette Coefficient** measures how similar a data point is to its own cluster compared to other clusters. The coefficient ranges from -1 to 1, with higher values indicating better-defined clusters.

From the results, we observe the following progression across the iterations:

- In the **first iteration**, K-Means produced an SSE of **480.9462** and a Silhouette Score of **0.4546**. This iteration reflects the initial grouping of data points into clusters, where the SSE is relatively high, and the Silhouette Score suggests moderate cohesion and separation between clusters.
- In the **second iteration**, the SSE decreases to **469.0378**, showing that the clusters become tighter and data points are closer to their centroids. Simultaneously, the Silhouette Score increases slightly to **0.4553**, which suggests an improvement in the separation between clusters.
- In the **third iteration**, the SSE stabilizes at **468.8070**, and the Silhouette Score reaches **0.4640**, which indicates that the algorithm has converged. At this point, further iterations do not significantly change the clustering structure, suggesting that the clusters have been optimized.

Overall, the K-Means algorithm effectively minimizes the SSE with each iteration while slightly improving the Silhouette Score, showing that the algorithm performs well in creating compact and well-separated clusters.

Table 1 K-Means SSE

Iterasi	Silhoutte	SEE
1	0.454615	480.9462
2	0.455291	469.0378
3	0.464036	468.8070

3.5 K-Medoids Algorithm Evaluation

The **K-Medoids algorithm** differs from K-Means in that it selects actual data points as cluster centers (medoids) rather than calculating centroids. This makes K-Medoids more robust to outliers, as it minimizes the absolute error rather than the squared error. However, this algorithm can sometimes struggle with large datasets or clusters that vary significantly in size.

- In the **first iteration**, the K-Medoids algorithm produced an SSE of **567.6125** and a Silhouette Score of **0.3826**. The relatively high SSE indicates that the clusters are not as tight as those produced by K-Means, and the lower Silhouette Score suggests that the separation between clusters is weaker.
- Across the **second** and **third iterations**, the SSE and Silhouette Score remained unchanged, indicating that K-Medoids converged after the first iteration. The inability to reduce the SSE or improve the Silhouette Score suggests that the algorithm was unable to find a more optimal clustering configuration.

This stability in SSE and Silhouette Score across iterations points to a limitation of K-Medoids in this particular case: while it is resistant to outliers, it may not always produce the most compact or well-separated clusters, especially in datasets with a large number of points or varied distribution patterns.

Table 2 K-Medoid SSE

Iterasi	SSE	Silhoutte
1	567,6125	0.382616
2	567,6125	0.382616
3	567,6125	0.382616

3.6 Performance of K-Means Algorithm

From the analysis, the K-Means algorithm outperforms K-Medoids in this study. The final SSE value of 468.8070 and the Silhouette Score of 0.4640 indicate that the clusters produced by K-Means are compact and well-separated. K-Means excels in this scenario for several reasons:

- Iterative improvement: The algorithm continuously refines the position of the centroids, reducing SSE with each iteration, which results in tighter clusters.
- Efficiency: K-Means is computationally efficient and works well with large datasets, making it ideal for clustering the tax data in this study.

- Cluster structure: The higher Silhouette Score indicates that the clusters have clear boundaries, with data points being closer to their assigned centroids than to points in other clusters.

These results show that K-Means is well-suited for tasks involving large datasets with continuous numeric attributes, like tax assessments, as it creates well-separated clusters that facilitate more accurate grouping of hamlet areas based on tax owed.

3.7 Performance of K-Medoid Algorithm

In contrast, the **K-Medoids algorithm** performed less effectively in this study. The SSE of **567.6125** and the **Silhouette Score of 0.3826** across all iterations suggest that the clusters are less compact and the boundaries between clusters are less distinct. Several factors contribute to the lower performance:

- **Robustness to outliers**: While K-Medoids is more resistant to outliers by selecting medoids, in this dataset, which does not have significant outliers, this feature did not provide a noticeable advantage.
- **Lack of iterative improvement**: Unlike K-Means, which improves cluster quality with each iteration, K-Medoids remained static after the first iteration. This shows that K-Medoids was not able to optimize the clusters for this dataset.
- **Slower convergence**: K-Medoids generally requires more computational resources, and while it converged quickly in this case, it did not result in better clusters.

This suggests that while K-Medoids can be useful for datasets with outliers or non-numeric attributes, it may not be the best choice for datasets like the tax data in this study, where continuous numeric variables dominate, and cluster compactness is crucial.

3.8 Comparative Analysis

When comparing the two algorithms based on the results:

- K-Means consistently achieved lower SSE values and higher Silhouette Scores, reflecting tighter, better-defined clusters.
- K-Medoids produced higher SSE values and lower Silhouette Scores, which indicate weaker clustering performance.

The results demonstrate that K-Means is the superior algorithm for clustering hamlet areas based on tax owed in this study, providing more reliable and interpretable groupings. This finding aligns with previous research that highlights K-Means' efficiency and accuracy for clustering large datasets with continuous numeric variables. K-Medoids, while offering robustness to outliers, falls short in terms of optimizing cluster quality for this specific task.

Table 3 Comparative Result

Algoritm	Silhoutte	SEE
K Means	0.454615	480.9462
K Medoids	0.382616	567,6125

4 Conclusion

Based on the research findings comparing the K-Means and K-Medoids algorithms using 368 data points, several conclusions can be drawn. Both algorithms successfully identified groups of hamlets in Sutojayan District based on tax debt indicators. The K-Means algorithm demonstrated faster performance but was more sensitive to outliers, producing the following cluster results: in Cluster 1, there were 0 low-tax areas, 0 medium-tax areas, and 28 high-tax areas; in Cluster 2, there were 91 low-tax areas, 79 medium-tax areas, and 66 high-tax areas; and in Cluster 3, there were 33 low-tax areas, 43 medium-tax areas, and 28 high-tax areas. In contrast, the K-Medoids algorithm proved more robust to outliers, with the following cluster results: in Cluster 1, there were 91 low-tax areas, 78 medium-tax areas, and 20 high-tax areas; in Cluster 2, there were 33 low-tax areas, 43 medium-tax areas, and 31 high-tax areas; and in Cluster 3, there were 0 low-tax areas, 1 medium-tax area, and 71 high-tax areas.

This comparison shows that K-Means is more efficient in terms of computational time and produces more compact clusters. However, K-Medoids is more robust against outliers and noise in the data, making it a better choice when the

data contains extreme values that could influence clustering results. In this study, K-Medoids exhibited lower performance in terms of evaluation metrics.

When comparing the two algorithms, K-Means had a lower SSE (480.9462) compared to K-Medoids (567.6125), indicating that K-Means is more effective in minimizing the distance between data points within a cluster. Additionally, K-Means also outperformed K-Medoids in terms of the Silhouette Coefficient, with a score of 0.4546 versus 0.3826 for K-Medoids. This suggests that the clusters formed by K-Means are more clearly separated than those formed by K-Medoids, resulting in better-defined groupings.

Compliance with ethical standards

Acknowledgments

A big thank you is conveyed to all those who have helped and been involved in this research, without the support and assistance of all of them, this paper will not be completed and the objectives of this research will not be achieved.

Disclosure of conflict of interest

The authors state that there are no personal, financial, or organizational conflicts of interest that may affect the output of this research.

References

[1] M. A. Putri, Rahaning Nining, F. M. Basysyar, and O. Nurdiawan, "Penerapan Data Mining Menggunakan Metode Clustering Untuk Mengetahui Kelompok Kepatuhan Wajib Pajak Bumi dan Bangunan," Jurnal Accounting Information System (AIMS, vol. 5, no. 2, pp. 145–156, 2022, [Online]. Available: https://jurnal.masoemuniversity.ac.id/index.php/aims

[2] D. Fahrizal, J. Kustija, and M. A. H. Akbar, "Development Tourism Destination Recommendation Systems using Collaborative and Content-Based Filtering Optimized with Neural Networks," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 7, no. 2, p. 285, Apr. 2024, doi: 10.24014/ijaidm.v7i2.28713.

[3] I. H. Santi, F. Febrinita, and W. D. Puspitasari, "Engineering Design Business Process Modelling Letter C Land Data Archiving System with Software Requirement Specifications Approach," vol. 6, no. 4, pp. 231–240, 2023.

[4] M. A. H. Akbar, D. Fahrizal, J. Kustija, and I. Surya, "Digital Technology Integration in TVET for Tourism: A Case Study for an Android-Based Application Development and Implementation," in 2024 9th International STEM Education Conference (iSTEM-Ed), 2024, pp. 1–6. doi: 10.1109/iSTEM-Ed62750.2024.10663108.

[5] D. E. Cahyani, L. M. T. Utami, and H. Setiadi, "Clustering of Javanese News in Krama Alus Level with Javanese Stemming," in 2019 International Conference on Information and Communications Technology (ICOIACT), 2019, pp. 462–467. doi: 10.1109/ICOIACT46704.2019.8938438.

[6] J. Oyelade et al., "Data Clustering: Algorithms and Its Applications," in 2019 19th International Conference on Computational Science and Its Applications (ICCSA), 2019, pp. 71–81. doi: 10.1109/ICCSA.2019.000-1.

[7] I. M. Karo Karo, S. Dewi, M. Mardiana, F. Ramadhani, and P. Harliana, "K-Means and K-Medoids Algorithm Comparison for Clustering Forest Fire Location in Indonesia," Jurnal Ecotipe (Electronic, Control, Telecommunication, Information, and Power Engineering), vol. 10, no. 1, pp. 86–94, 2023, doi: 10.33019/jurnalecotipe.v10i1.3896.

[8] S. Ghaida Muthmainah and A. Id Hadiana, "Comparative Analysis of K-Means and K-Medoids Clustering in Retail Store Product Grouping," International Journal of Quantitative Research and Modeling, vol. 5, no. 3, pp. 280–294, 2024.

[9] Ö. N. Kenger, Z. D. Kenger, E. Özceylan, and B. Mrugalska, "Clustering of Cities Based on Their Smart Performances: A Comparative Approach of Fuzzy C-Means, K-Means, and K-Medoids," IEEE Access, vol. 11, pp. 134446–134459, 2023, doi: 10.1109/ACCESS.2023.3333753.

[10] N. Sureja, B. Chawda, and A. Vasant, "An improved K-medoids clustering approach based on the crow search algorithm," Journal of Computational Mathematics and Data Science, vol. 3, p. 100034, 2022, doi: https://doi.org/10.1016/j.jcmds.2022.100034.

[11] Jaja Kustija, Irgi Surya, and Diki Fahrizal, "Design of automated power factor monitoring and repair tool for industry in real time based on Internet of Things," International Journal of Science and Technology Research Archive, vol. 3, no. 2, pp. 001–008, Oct. 2022, doi: 10.53771/ijstra.2022.3.2.0106.

[12] J. Kuang, G. Xu, A. Jian, H. Jatnika, and H. Jamaludin, "Comparison K-Medoids Algorithm and K-Means Algorithm for Clustering Fish Cooking Menu from Fish Dataset Comparison K-Medoids Algorithm and K-Means Algorithm for Clustering Fish Cooking Menu from Fish Dataset," 2021, doi: 10.1088/1757-899X/1088/1/012034.

[13] Qomariyah and M. U. Siregar, "Comparative Study of K-Means Clustering Algorithm and K-Medoids Clustering in Student Data Clustering," JISKA (Jurnal Informatika Sunan Kalijaga), vol. 7, no. 2, pp. 91–99, 2022, doi: 10.14421/jiska.2022.7.2.91-99.

[14] T. Akbar, G. M. Tinungki, and S. Siswanto, "Performance Comparison of K-Medoids and Density Based Spatial Clustering of Application With Noise Using Silhouette Coefficient Test," BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 17, no. 3, pp. 1605–1616, 2023, doi: 10.30598/barekengvol17iss3pp1605-1616.

[15] T. Sucita, D. L. Hakim, R. H. Hidayahtulloh, and D. Fahrizal, "Solar irradiation intensity forecasting for solar panel power output analyze," Indonesian Journal of Electrical Engineering and Computer Science, vol. 36, no. 1, pp. 74–85, Oct. 2024, doi: 10.11591/ijeecs.v36.i1.pp74-85.

[16] J. Kustija, D. Fahrizal, and M. Nasir, "Revitalizing IoT-based air quality monitoring system for major cities in Indonesia," Sinergi (Indonesia), vol. 28, no. 3, pp. 605–616, 2024, doi: 10.22441/sinergi.2024.3.016.

[17] M. Wahyudi and L. Pujiastuti, "Comparison of K-Means Clustering and K-Medoids in Clustering Fresh Milk Production in Indonesia," Jurnal Bumigora Information Technology (BITe), vol. 4, no. 2, pp. 243–254, 2022, doi: 10.30812/bite.v4i2.2104.

[18] J. Kustija, D. Fahrizal, M. Nasir, D. Setiawan, and I. Surya, "Design and development of coastal marine water quality monitoring based on IoT in achieving implementation of SDGs," Indonesian Journal of Electrical Engineering and Computer Science, vol. 36, no. 3, pp. 1470–1484, Dec. 2024, doi: 10.11591/ijeecs.v36.i3.pp1470-1484.

[19] R. Sammouda and A. El-Zaart, "An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method," Comput Intell Neurosci, vol. 2021, no. 1, p. 4553832, Jan. 2021, doi: https://doi.org/10.1155/2021/4553832.

[20] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," Artif Intell Rev, vol. 56, no. 7, pp. 6439–6475, 2023, doi: 10.1007/s10462-022-10325-y.

[21] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," Entropy, vol. 23, no. 6, Jun. 2021, doi: 10.3390/e23060759.

[22] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," International Journal of Advances in Scientific Research and Engineering, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/ijasre.2021.34050.

[23] H. Hu, J. Liu, X. Zhang, and M. Fang, "An Effective and Adaptable K-means Algorithm for Big Data Cluster Analysis," Pattern Recognit, vol. 139, p. 109404, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109404.

[24] A. Fahim, "K and starting means for k-means algorithm," J Comput Sci, vol. 55, p. 101445, 2021, doi: https://doi.org/10.1016/j.jocs.2021.101445.

[25] A. V Ushakov and I. Vasilyev, "Near-optimal large-scale k-medoids clustering," Inf Sci (N Y), vol. 545, pp. 344–362, 2021, doi: https://doi.org/10.1016/j.ins.2020.08.121.

[26] N. Sureja, B. Chawda, and A. Vasant, "An improved K-medoids clustering approach based on the crow search algorithm," Journal of Computational Mathematics and Data Science, vol. 3, p. 100034, 2022, doi: https://doi.org/10.1016/j.jcmds.2022.100034.

[27] Z. Wu, L. Jin, J. Zhao, L. Jing, and L. Chen, "Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm," Comput Intell Neurosci, vol. 2022, no. 1, p. 9930613, Jan. 2022, doi: https://doi.org/10.1155/2022/9930613.

[28] Y. Shen, D. Zhang, R. Wang, J. Li, and Z. Huang, "SBD-K-medoids-based long-term settlement analysis of shield tunnel," Transportation Geotechnics, vol. 42, p. 101053, 2023, doi: https://doi.org/10.1016/j.trgeo.2023.101053.