Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 13, No. 1, Juli 2025

ISSN 0216 – 0544
e-ISSN 2301– 6914

# PCA-counseled k-means and k-medoids with dimension reduction for improved in determining optimal aid clustering

**Achmad Jauhari [a,*], Ika Oktavia Suzanti [b], Devie Rosa Anamisa [c], Fadhila Tangguh Admojo [d]**

[a,b,c] *Informatics Engineering Department, University of Trunojoyo Madura, Bangkalan, Indonesia*
[d] *Malaysian Institute of Information Technology, Universiti kuala lumpur*
E-mail: jauhari@trunojoyo.ac.id*, iosuzanti@trunojoyo.ac.id, devros_gress@trunojoyo.ac.id, fadhila.tangguh@s.unikl.edu.my

***Abstract***

*Assuring effective allocation requires targeted distribution of aid, which makes aid clustering a crucial component. For the purpose of using data-driven segmentation based on important characteristics to determine effective help targeting, accuracy in clustering is essential. The study explores the combination of Principal Component Analysis (PCA), k-means, and k-medoids to enhance aid clusters, with the goal of increasing aid distribution accuracy and efficiency. The information gathered consists of 1600 records with 13 attributes. In order to standardized the data having two processes in it, preprocessing is first applied. When used with PCA, it makes measuring variance easier and preserves 80% of the variation by choosing five components. The number of clusters may be determined with the use of PCA, k-medoids, and the k-means approach. Greater PCA-k-means silhouette coefficients, which indicate better clustering ability, are highlighted by comparative analysis. This analysis shows that PCA-k-means is an effective technique for creating accurate and unique clusters within a data set's structure.The clustering results using the PCA-k-means algorithm have produced the greatest accuracy in the silhouette score of 0.49 and the DBI score is 0.84.*

*Key words: Aid data, Clustering, K-means, K-medoids, Principal Component Analysis.*

## INTRODUCTION

In an era where the need for targeted distribution of aid is increasingly urgent [1], clustering data on aid recipients is a crucial element in ensuring the efficiency and effectiveness of such distribution. In this context, boosting the efficiency and precision with which clusters are determined is critical to ensuring that aid beneficiaries are effectively targeted. With a data-driven approach, segmentation based on highly relevant and significant characteristics ensures that help beneficiaries are included in groups that truly require it. Income, work position, health issues, and social vulnerabilities are examples of such factors. In order to offer a more accurate answer, advanced data analysis techniques such as machine learning or intelligent algorithms are used to investigate many criteria that are significant in establishing aid recipient clusters [2], [3], [4]. An approach using the k-means and K-Medoids methods, which is supported by the PCA (Principal Component Analysis) dimension reduction technique, is the main focus in this article.

Clustering is a data analysis method that defines homogenous groups in a data set and is used extensively in a variety of industries [5]. The k-means algorithms is now widely used clustering techniques same as k-medoids [6], [7], [8]. Although both try to arrange data into

clusters, each takes a different strategy. K-means emphasizes centroid representation, whereas k-medoids employs real data points as cluster centers. These distinctions allow for distinct techniques to interpreting data structures and studying cluster formation. However, the accuracy of the generated clusters is frequently hampered by the data's high dimensionality and complexity that surpasses processing capabilities.

Principal Component Analysis (PCA) has been shown to be successful in balancing the trade-off between accuracy and computational efficiency [9], [10] [11]. Using PCA to compress data dimensions into smaller but more meaningful dimensions is an appealing way to improve cluster quality. PCA attempts to summarize the information in the data by identifying the main components that are most informative or best represent the changes in the dataset [12], [13]. As a result, PCA minimizes data dimensionality without compromising crucial information. In [14], the PCA method is used to reduce variables in the k-means clustering algorithm. As a result, reducing the dimensionality of the dataset using the PCA method is proven to improve cluster quality. There is an increase in model quality in the application of PCA-k-means compared to traditional k-means. The increase in model quality is seen from the increase in clustering accuracy and the results of measuring cluster validity with the Davies-Bouldin Index show that PCA-k-means produces the smallest DB Index value compared to traditional k-means.

This study aims to explore how k-means and k-medoids can be integrated by leveraging PCA's dimensionality reduction techniques. Exploring the potential for major improvements in establishing appropriate clusters for assistance recipient data therefore provides a more precise and efficient approach for aid distribution. This article will go into detail regarding k-means, k-medoids, and the importance of PCA in data dimensionality reduction in the next part. Furthermore, case studies and experimental findings will be examined to demonstrate how the combination of the three might affect the improvement of establishing more ideal clusters of assistance recipient data.

## MATERIAL AND METHODS

This methodology outlines crucial procedures for carrying out studies that make use of the suggested approach to identify the best help target data clusters. The whole procedure will assist in finding valuable patterns in the data and direct more sensible choices on the distribution of aid (refer to Fig.1). The stages in this research methodology include data preprocessing, which:

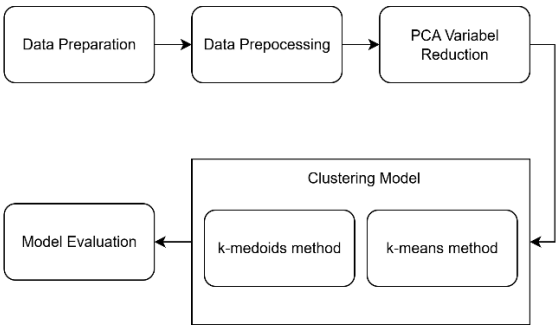



Fig. 1. Methods of research

### Dataset Preparation

Data from stunting support programs were used in this study. Gather pertinent data about aid recipients, making sure that the characteristics are diverse and include details about the recipients' location, needs, and other factors that are judged significant. A total of 1600 data records were collected from various Community Health Centers in the Madura region to acquire this information. Thirteen qualities in all, representing the deciding criteria, are contained in the data that was acquired.

- Gender (Gd)
- Occupation (Occ)
- Education (Ed)
- Home Ownership (HO)
- Having Money/ Jewelry/ Livestock/ Other Saving (HM)
- Roof Type (RT)
- Waal Type (WT)
- Floor Type (FT)
- Source of Lighting (SoL)
- Cooking Fuel (CF)
- Source of Dringking Water (SoDW)
- Have Defecation Facilities (HDF)
- Stunting Risk (SR)

## Data Preprocessing

A step in data analysis called data preprocessing gets the data ready for further processing. At this stage, two operations are carried out: data type conversion and data normalization. It is well known that data received in the form of categorical variables need additional processes to alter the ordinal numbers. The OrdinalEncoder function from scikit-learn is used in this procedure to transform categorical data into ordinal integers ranging from 0 to $k-1$ [15]. $k$ denotes the number of distinct categories.

The second procedure, if necessary, normalizes or standardizes the data to guarantee consistent scaling across all aspects. It is crucial to scale or normalize the data in order to ensure that each characteristic has an equal impact on the clustering outcomes since clustering algorithms are sensitive to different scales. The Min-max normalization computation (Equation 1) was selected for data scaling. The normalized value of the resultant computation is denoted by $x_{scaled}$. $x$ is original value, while max and min represent the feature's initial maximum and minimum values [16].

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

## Principal Component Analysis (PCA)

One method for reducing dimensionality is PCA, which involves projecting data from a high-dimensional environment to a lower-dimensional environment while retaining as much relevant information as is practical [17]. PCA helps reduce the number of dimensions in a data set with many characteristics or dimensions, which reduces computational complexity without sacrificing the significance of the data [18]. As a result, it reduces processing time and saves computer resources while retaining the essence or critical information in the processed data. This is accomplished by identifying a linear combination of data attributes (variables) that best characterizes the variance or dispersion in the data. The important of the critical information in the original data is kept throughout this procedure, but the data dimensionality is greatly decreased.

The PCA technique begins the process by computing the covariance matrix from standard data to highlight the relationship between features since the data is ready to be analyzed after the preprocessing step. This computation seeks to determine the relationship between variables (Equation 2) [19].

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \quad (2)$$

Where $X$ is a set of n vectors, each element $X_i$ is a data set instance, and $\bar{X}$ is the average vector of each feature. Using this covariance matrix, PCA determines the eigenvalues and eigenvectors. Eigenvectors define the converted coordinate system's new axes, whereas eigenvalues define the degree of variation along each of these axes. The next step is to use eigenvectors associated with the selected principal components to convert the data into a reduced dimensional space. Thus, PCA provides for a simplified representation of the data while preserving crucial information by identifying the main components that contribute the most to the data's variance.

## K-Means

A popular centroid-based clustering method in machine learning called K-Means divides a data set into $K$ distinct groups [20][21]. The cluster centers are then modified to minimize the total of the distances between each of the data points and their corresponding cluster centers via an iterative process in which each data point is allocated to the closest cluster based on similar qualities. The "$K$" in k-means denotes the number of clusters that the algorithm wishes to identify in a data set. The first step is to assign a random centroid to each cluster. The centroid is then repeatedly updated and samples are assigned to clusters based on the closest Euclidean distance between the sample and the centroid [22]. This procedure is divided into two steps: first, sample characteristics are assigned to clusters with the closest centroid; Secondly, each cluster's centroid is adjusted to represent the mean of all samples inside that cluster. Until convergence, when the maximum number of iterations is achieved or the change in centroid location becomes insignificant, these iterations continue. The method aims to minimize the within-cluster sum of squares (WCSS) in order to achieve tight clustering. The Elbow technique finds locations where the rate of decline in WCSS suddenly changes, creating a elbow on the plot, and is frequently used to

determine the right number of clusters in k-means.

## K-Medoids

In contrast to k-means, the k-medoids technique focuses on using real data points as cluster representations, which are referred to as medoids [23]. The first step is to randomize the center point of each cluster. Following that, the iterative procedure begins by allocating each sample to the cluster with the nearest center point based on distance metric. After the first determination, one of the samples in the cluster with the least change in total distance from the center point of all samples gets its center point updated for each cluster. Until the center point does not change significantly or until the maximum number of iterations is achieved, this iteration is repeated. Finding data points that represent the cluster's center is the primary objective of the k-medoids approach in order to reduce the total distance between each cluster sample and the center point. This enables the development of strong, concentrated clusters on real data points that represent the data set's structure.

## Evaluation

### *Silhouette Coefficient*

Silhouette Coefficient is an evaluation statistic used to assess the quality of the algorithm's clusters. This statistic compares how well each sample is put in its own cluster to other clusters. Equation 3 shows the Silhouette Coefficient formula for each sample $i$ in the cluster [24], [25].

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (3)$$

where $a(i)$ represents the average distance between sample $i$ and all other samples in the same cluster (intra-cluster distance), and $b(i)$ represents the average distance between sample $i$ and all samples in the nearest cluster other than the cluster it possesses (inter-cluster distance). The Silhouette Coefficient value ranges from -1 to 1. A score near to 1 suggests that the sample is well put in its own cluster and far away from other clusters, whereas a negative value indicates that the sample may be misplaced in the cluster.

### *Davies-Bouldin Index*

An evaluation metric called the Davies-Bouldin index and the Silhouette coefficient is employed to rate the quality of clusters produced by a clustering algorithm. The objective is to assess how well-defined, compact, and well-separated the clusters are within the dataset [26]. Equation 4 depicts the Davies-Bouldin exponential formula for n clusters.

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{j\neq i}\left(\frac{s_i+s_j}{d(c_i,c_j)}\right) \quad (4)$$

Where $s_i$ is a measure of cluster $i$'s compactness computed by the average distance between each point in the cluster and its center, $d(c_i, c_j)$ is the distance between the centers of clusters $i$ and $j$ calculated using the ecludian distance, and $c_i$ is the cluster center $i$. The Davies-Bouldin index determines how close one cluster is to another (lower values are preferable), with a low result indicating that the clusters are mutually exclusive and compact [27]. This index may be used to assess the clarity and separation of clusters, which is significant for assessing clustering algorithms' capacity to group data. The DBI evaluation aim is approaching zero since it demonstrates that the clusters are ideally distinct and compact.

# RESULT AND DISCUSSION

## Preparation of Data and PCA Variable Selection

In this study, 1600 records with 13 attributes were processed (Tabel 1). The data has been standardized at the preprocessing step with OrdinalEncoder funtion to transform categorical data into ordinal integers ranging and min max scaler for converts data values into a specified range, usually 0 to 1. The outcomes is standardized data are shown in Table 2.

The PCA algorithm is then used based on the preprocessing data. Starting, deciding on the number of features based on variance with the aim of considering how many features (main components) you want to retain based on variance in PCA. Fig. 2 depicts the results in the form of a cumulative variance plot. The graph depicts the amount of variation caught (on the y-axis) as a function of the number of components included (on the x-axis). This demonstrates how much information is kept when different amounts of major components are used. The proportion of variance explained

by each main component is computed by dividing the component's eigenvalue by the total of all eigenvalues and multiplying the result by 100. The variance proportion plot will have the same shape as the eigenvalue plot.

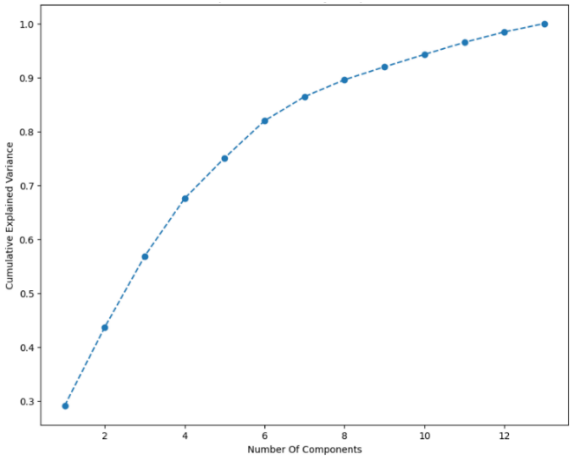



Fig. 2. Explained variance by components

Table 3. PCA summary

| Component | Eigenvalue | Variance (%) | Cummulative variance (%) |
|---|---|---|---|
| PC 1 | 0.330821 | 29.092417 | 29.092417 |
| PC 2 | 0.165172 | 14.525206 | 43.617623 |
| PC 3 | 0.150158 | 13.204888 | 56.822511 |
| PC 4 | 0.121881 | 10.718240 | 67.540751 |
| PC 5 | 0.085624 | 7.529752 | 75.070502 |

The graphs in Fig. 2 display the cumulative variance explained by the major component and all previous components, while the proportion of variance plot indicates the specific percentage of the total variation explained by each principal component. The reference point is the point at which the plot reaches the required amount of variance. The general norm is to keep the variance under 80%. As a result, this study opted to keep five components. Table 3 displays the overall outcomes of PCA reduction.

## K-Means and PCA-K-Means

Twenty clusters were assigned at random for selecting the clusters. The k-means and PCA-k-means algorithms will be tested with up to 20 clusters in this study. Next, find the optimal number of clusters. The Elbow technique was adopted to accomplish this purpose. The method entails inspecting the within-cluster sum of squares (WCSS) graph for kinks or elbows (Fig. 3). The portion of the graph that appears before the elbow usually dips abruptly, but the one that appears after is considerably smoother.

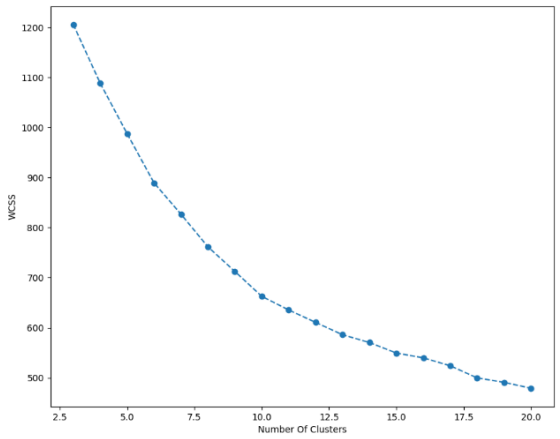
The kink appears at the 14th cluster mark in the K-Means method, whereas PCA-k-means appears at the 17th mark. Finally, create a cluster visualization (Fig. 4) to help you design the cluster. To better comprehend the structure of clustered data, view cluster findings using the k-means and PCA-k-means algorithms. To assess the efficacy of the two methods discussed above, the silhouette coefficient and DBI were calculated (Fig. 5). According to the k-means algorithm findings, the silhouette score: 0.2997831266686746 and the DBI score: 1.2838431441725384. Meanwhile, the greatest accuracy in the PCA-k-means method was obtained in the silhouette score: 0.4900544270778146 and DBI score: 0.8485498784977883. Fig. 5 shows that PCA-k-means performs well in that each sample is assigned to its own cluster when compared to other clusters. This is demonstrated by a silhouette coefficient value of (almost 1) and a DBI value of (nearly 0).

Table 1. Original dataset

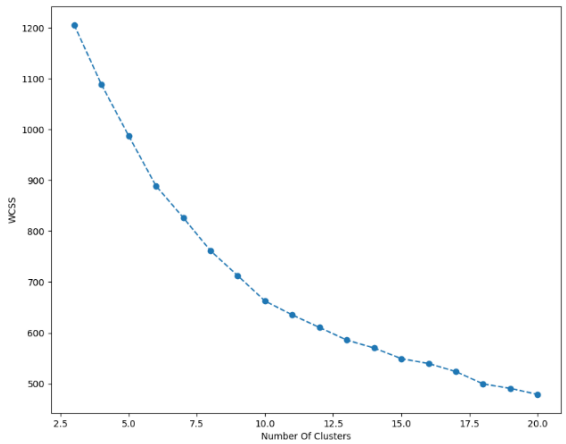
| Gd | Occ | Ed | HO | HM | RT | WT | FT | SoL | CF | SoDW | HDF | SR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Man | Self Employed | elementary school | One's own | Yes | Wood | Wood | Cement | No electricity | Gas | wells | No | 1 |
| Man | Farmer | elementary school | Hitchhiking | No | Straw | Bamboo | Cement | No electricity | Gas | wells | No | 0 |
| Woman | Farmer | No School | Hitchhiking | No | Straw | Bamboo | Land | No electricity | Wood | wells | No | 1 |
| Man | Not Woking | No School | Hitchhiking | Yes | Roofile | Walls | Ceramic | electricity | Gas | Drilling wells | Yes | 0 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| … | … | … | … | … | … | … | … | … | … | … | … | … |

Table 2. Standardized data

| Gd | Occ | Ed | HO | HM | RT | WT | FT | SoL | CF | SoDW | HDF | SR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 1.000000 | 0.555556 | 1.000000 | 1.0 | 1.0 | 0.333333 | 0.75 | 1.000000 | 0.5 | 0.833333 | 0.0 | 0.5 |
| 0.0 | 0.714286 | 0.555556 | 0.666667 | 0.0 | 0.8 | 0.000000 | 0.75 | 1.000000 | 0.5 | 0.833333 | 0.0 | 0.0 |
| 1.0 | 0.714286 | 1.000000 | 0.666667 | 0.0 | 0.8 | 0.000000 | 1.00 | 1.000000 | 0.0 | 0.833333 | 0.0 | 0.5 |
| 0.0 | 0.857143 | 1.000000 | 0.666667 | 1.0 | 0.6 | 1.000000 | 0.50 | 0.666667 | 0.5 | 0.666667 | 0.5 | 0.0 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| … | … | … | … | … | … | … | … | … | … | … | … | … |

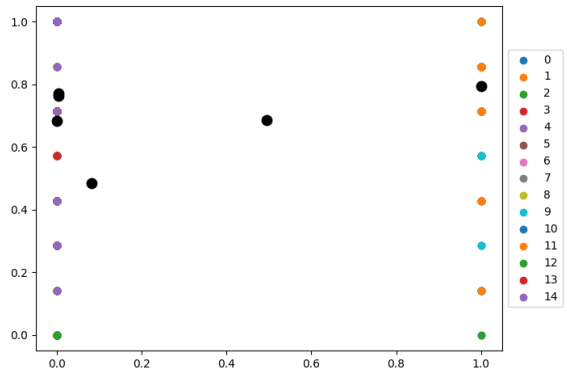



(a)

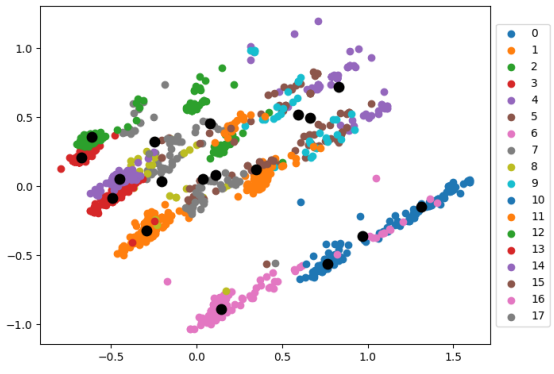



(b)

Fig. 3. WCSS plot: (a) K-means; (b) PCA-k-means

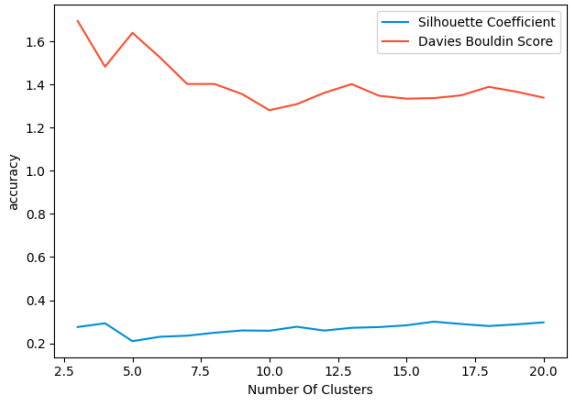


(a)

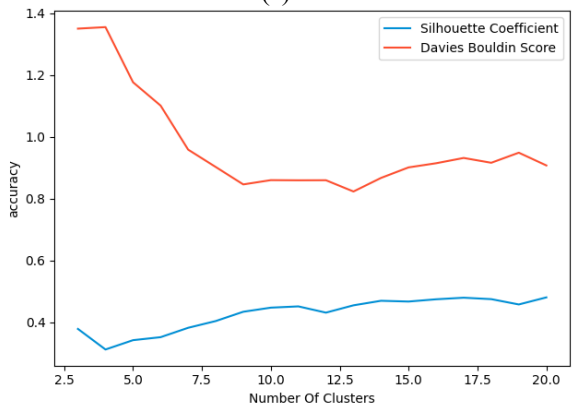


(b)

Fig. 4. Visualization of clustered data: (a) K-means; (b) PCA-k-means




(a)

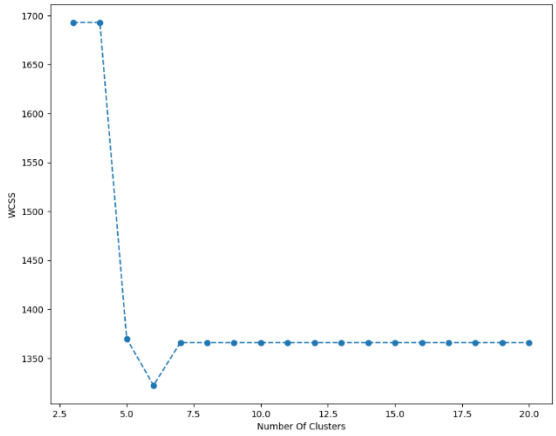



(b)

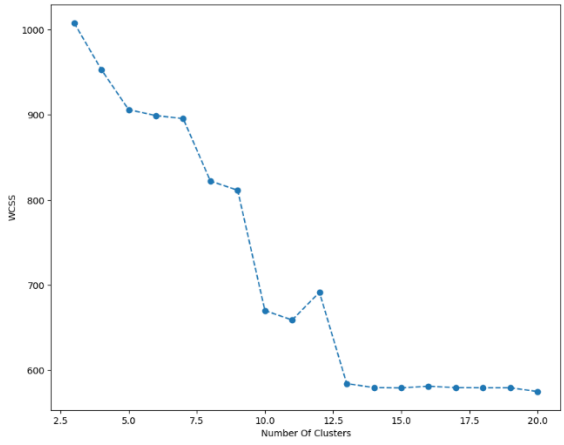
Fig. 5. Evaluation measurement: (a) K-means; (b) PCA-k-means

## K-Medoids and PCA-K-Medoids

When identifying clusters, it works in the same way as the k-means method. With a total of 20 clusters, this study intends to examine additional clustering techniques, particularly k-medoids and PCA-k-medoids. The elbow approach is then used to identify the ideal number of clusters. This approach entails inspecting the WCSS diagram for kinks or bends.



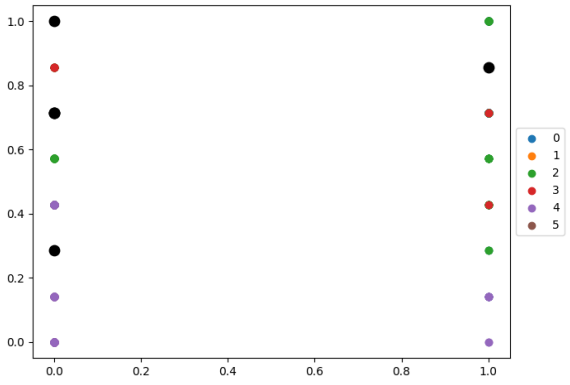

(a)



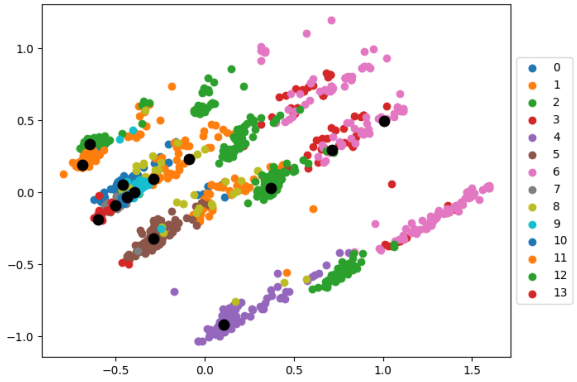

(b)

Fig. 6. WCSS plot: (a) K-means; (b) PCA-k-means

Fig. 6 shows that the kink occurs at mark 5 of the cluster in the K-Medoids algorithm, while PCA-K-Medoids is encountered at mark 13. Finally, build the cluster by creating a cluster visualization (Fig. 7). visualize cluster results to understand the structure of clustered data better. To see the performance of the two algorithms above, an evaluation was carried out by calculating the silhouette coefficient and DBI (Fig. 8). The results of the K-Medoids algorithm, the best accuracy was obtained in silhouette score: 0.13489339899249003 and DBI score: 2.6661737967151846. Meanwhile, in the PCA-k-medoids algorithm, the best accuracy was obtained in the silhouette score: 0.39968012202622377 and DBI score: 0.8848653366104022. In Fig. 7 it can be seen that PCA-k-medoids has good performance in that each sample is placed in its own cluster compared to other clusters. This is indicated by a silhouette coefficient value of (close to 1) and a DBI value of (close to 0)

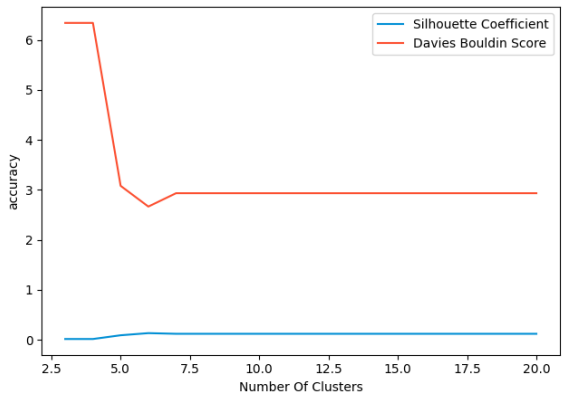


(a)

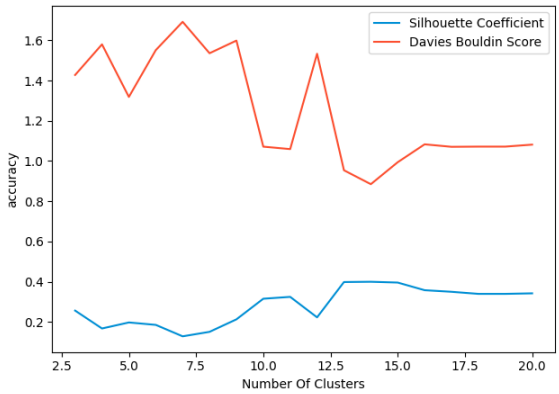


(b)

Fig. 7. Visualization of clustered data: (a) K-means; (b) PCA-k-means




(a)




(b)

Fig. 8. Evaluation measurement: (a) K-medoids; (b) PCA-k-medoids

## Discussion

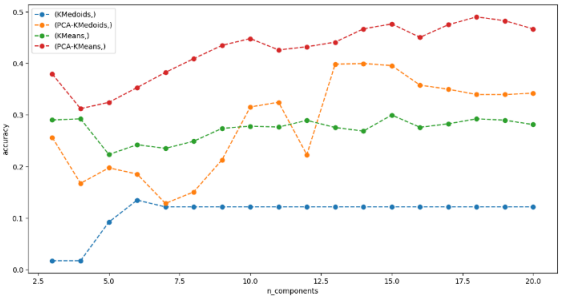



Fig. 9. Accuracy performance comparison

All computations were performed to identify the best number of clusters using methods k-means, k-medoids, and PCA for dimension reduction. Fig. 5 and Fig. 8 illustrate the results of applying each approach using silhouette coefficients and DBI. Fig. 9 was conveniently made to compare the performance of each strategy. The silhouette assessment coefficient is used to construct this graph, and a value near to 1 implies that the sample is well situated in its own cluster and far from other clusters. According to the silhouette accuracy points in Fig. 9, PCA-k-means has a silhouette coefficient value close to 1.

## CONCLUSION

The research involved a complete procedure that included data preparation, feature reduction by PCA, and testing of several clustering techniques. The dataset, which had 1600 records and 13 characteristics, was preprocessed in which categorical data was converted into ordinal integers using the OrdinalEncoder function and then normalized to a defined range using the min-max scaler. The study used PCA to measure variance and identify the ideal number of components. The cumulative variance plot shows the information maintained as a function of component inclusion. Maintaining 80% of variance by picking five components. For cluster selection, twenty clusters were selected at random. In this study, k-means, k-medoids, and PCA for reduce dimension methods will be examined with up to 20 clusters. Next, determine the optimal number of clusters. To achieve this goal, the Elbow method was used. The approach comprises looking for bends or elbows in the within-cluster sum of squares (WCSS) graph. The comparison analysis combines the performance of k-means, k-medoids, PCA-k-means, and PCA-k-medoids, highlighting the larger silhouette coefficient seen in PCA-k-means, which indicates greater clustering performance. Overall, these calculations and assessments aided in finding the most successful clustering algorithms for the dataset, emphasizing the power of PCA-k-means in creating accurate and unique clusters within the data structure.

## REFERENCES

[1] C. Apodaca, "Foreign Aid as Foreign Policy Tool," 2017, Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.332

[2] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," IEEE Access, vol. 7, pp. 60134-60149, 2019, https://doi.org/10.1109/ACCESS.2019.2914999

[3] H. Sun, H. V Burton, and H. Huang, "Machine learning applications for building structural design and performance assessment: State-of-the-art review," Journal of Building Engineering, vol. 33, p. 101816, 2021,

https://doi.org/10.1016/j.jobe.2020.101816

[4] A. E. Ezugwu et al., "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," Eng Appl Artif Intell, vol. 110, p. 104743, 2022. https://doi.org/10.1016/j.engappai.2022.104743

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

[6] B. Lund and J. Ma, "A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering," Performance Measurement and Metrics, vol. 22, no. 3, pp. 161-173, 2021. https://doi.org/10.1108/PMM-05-2021-0026

[7] W. Utomo, "The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia," ILKOM Jurnal Ilmiah, vol. 13, no. 1, pp. 31-35, 2021. https://doi.org/10.33096/ilkom.v13i1.763.31-35

[8] Öm. 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. https://doi.org/10.1109/ACCESS.2023.3333753

[9] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Inform Med Unlocked, vol. 17, Jan. 2019, https://doi.org/10.1016/j.imu.2019.100179

[10] C, Ding., X. He. Cluster Structure of K-means Clustering via Principal Component Analysis. In: Dai, H., Srikant, R., Zhang, C. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2004. Lecture Notes in Computer Science(), vol 3056. Springer, Berlin, Heidelberg. 2024. https://doi.org/10.1007/978-3-540-24775-3_50 .

[11] O. Khouili, M. Hanine, and M. Louzazni, "Harnessing Principal Component Analysis and Artificial Neural Networks for Accurate Solar Radiation Prediction," Int J Energy Res, vol. 2025, no. 1, p. 5846114, 2025. https://doi.org/10.1155/er/5846114

[12] F. L. Gewers et al., "Principal component analysis: A natural approach to data exploration," ACM Computing Surveys (CSUR), vol. 54, no. 4, pp. 1-34, 2021. https://doi.org/10.1145/3447755

[13] E. Elhaik,"Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," Sci Rep, vol. 12, no. 1, p. 14683, 2022. https://doi.org/10.1038/s41598-022-14395-4

[14] Istina Alya Rosyada, Dina Tri Utari,"Penerapan Principal Component Analys is untuk Reduksi Variabel pada Algoritma K-Means Clustering", Jambura Journal of Probability and Statistics, vol. 5, no. 1, p. 6-13, 2024. https://doi.org/10.37905/jjps.v5i1.18733

[15] S. Galli, Python feature engineering cookbook. Packt Publishing Ltd, 2024.

[16] M. Shantal, Z. Othman, and A. A. Bakar, "A novel approach for data feature weighting using correlation coefficients and min-max normalization," Symmetry (Basel), vol. 15, no. 12, p. 2185, 2023. https://doi.org/10.3390/sym15122185

[17] A. Eka Haryati, R. Desia Arindra Putri, and A. Dahlan, "Comparison of fuzzy subtractive clustering and fuzzy c-means," Jurnal Ilmiah KURSOR, vol. 11, no. 1, pp. 1-8, 2021. https://doi.org/10.21107/kursor.v11i1.254

[18] H. Honda, P. Dinh, P. T. Thao, Y. Tabata, and B. D. Anh, "Dimensionality reduction as a non-cooperative game," in 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2023, pp. 72-77 https://doi.org/10.1109/ICAIIC57133.2023.10067075

[19] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for

diabetes prediction by integrating PCA and K-means techniques," Inform Med Unlocked, vol. 17, Jan. 2019, https://doi.org/10.1016/j.imu.2019.100179

[20] I. M. S. Bimantara and I. M. Widiartha, "Optimization of K-Means Clustering Using Particle Swarm Optimization Algorithm for Grouping Traveler Reviews Data on Tripadvisor Sites," Jurnal Ilmiah Kursor, vol. 12, no. 1, pp. 1-10, 2023. https://doi.org/10.21107/kursor.v12i01.269

[21] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-means and alternative clustering methods in modern power systems," Ieee Access, vol. 11, pp. 119596-119633, 2023. https://doi.org/10.1109/ACCESS.2023.3327640

[22] I. B. Djordjevic, "Chapter 14 - Quantum Machine Learning," in Quantum Information Processing, Quantum Computing, and Quantum Error Correction (Second Edition), I. B. Djordjevic, Ed., Academic Press, 2021, pp. 619-701. https://doi.org/10.1016/B978-0-12-821982-9.00007-1

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

[24] R. O. Sinnott, H. Duan, and Y. Sun, "Chapter 15 - A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather," in Big Data, R. Buyya, R. N. Calheiros, and A. V. Dastjerdi, Eds., Morgan Kaufmann, 2016, pp. 357-388. https://doi.org/10.1016/B978-0-12-805394-2.00015-5

[25] S. Zheng and J. Zhao, "States Identification of Complex Chemical Process Based on Unsupervised Learning," in Computer Aided Chemical Engineering, vol. 44, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds., Elsevier, 2018, pp. 2239-2244. https://doi.org/10.1016/B978-0-444-64241-7.50368-2

[26] S. Dahiya, H. Nanda, J. Artwani, and J. Varshney, "Using clustering techniques and classification mechanisms for fault diagnosis," International Journal, vol. 9, no. 2, 2020. https://doi.org/10.30534/ijatcse/2020/188922020

[27] B. Surarso and R. Gernowo , "Implementation of k-medoids clustering for high education accreditation data," jurnal ilmiah kursor, vol. 10, no. 3, pp. 119-128, 2020.