BAREKENG: Journal of Mathematics and Its Applications
March 2025 Volume 19 Issue 1 Page 0465–0476
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss1pp0465-0476

IMPLEMENTATION OF K-MEDOIDS AND K-PROTOTYPES CLUSTERING FOR EARLY DETECTION OF HYPERTENSION DISEASE

Hardianti Hafid[1], Selvi Annisa[2*]

[1]*Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Negeri Makassar*
Jln. Daeng Tata, Makassar, 90224, Indonesia

[2]*Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Lambung Mangkurat*
Jln. A. Yani Km 36, Kalimantan Selatan, 70714, Indonesia

*Corresponding author's e-mail: *selvi.annisa@ulm.ac.id*

Article History:
Received: 2nd July 2024
Revised: 29th November 2024
Accepted: 29th November 2024
Published: 13th January 2025

Keywords:
K-Medoids;
K-Prototypes;
Hypertension;
Clustering.

ABSTRACT

Hypertension is a serious concern because of its significant impact on public health, especially in the context of lifestyle changes and specific health conditions. One method for grouping patients based on complex clinical data is the Clustering method. This research type is quantitative, namely taking or collecting the necessary data and then analyzing it using the K-Medoids and K-Prototypes methods. The K-Medoids method is more resistant to outliers and noise than the K-Means method, which is more suitable for this research. The K-Prototypes method can handle mixed numerical and categorical data, effectively grouping hypertensive patients based on different variable categories. This research used the K-Medoids and K-Prototypes grouping methods to categorize patients into risk categories based on gender, age, family history of hypertension, smoking status, pulse rate, and increased systolic and diastolic blood pressure. The Elbow and Silhouette Coefficient methods were applied to evaluate the data and determine the optimal number of clusters for dividing patients into low-risk and high-risk hypertension groups. The analysis revealed that two clusters are the optimal solution. The clustering results show K-Medoids' superiority in grouping data with higher Silhouette Coefficient values compared to K-Prototypes. Overall, the K-Medoids and K-Prototypes algorithms can detect early hypertension risk by dividing patients into different risk groups. Although the clustering results are still weak, these two methods show potential in helping health institutions identify and treat hypertension risk in Indonesia.

How to cite this article:

H. Hafid and S. Annisa., "IMPLEMENTATION OF K-MEDOIDS AND K-PROTOTYPES CLUSTERING FOR EARLY DETECTION OF HYPERTENSION DISEASE," *BAREKENG: J. Math. & App.*, vol. 19, iss. 1, pp. 0465-0476, March, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Hypertension, or high blood pressure, is one of the leading health problems faced by people throughout the world. Hypertension is a condition with an abnormal increase in systolic and diastolic pressure. A person is categorized as suffering from hypertension if their systolic or diastolic blood pressure exceeds 140/90 mmHg, while normal blood pressure is 120/80 mmHg. The International Society of Hypertension (ISH) categorizes blood pressure (BP) into four classifications: normal (<130/85 mmHg), high-normal (130–139/85–89 mmHg), grade 1 hypertension (140–159/90–99 mmHg), and grade 2 hypertension (≥160/100 mmHg) [1]. Hypertension is closely related to lifestyle changes, consumption of foods high in fat, cholesterol levels, lack of physical activity, and stress [2]. Handling hypertension is a national priority in the non-communicable disease category. Several health institutions have conducted research to identify risk factors for non-communicable diseases, including hypertension. Hypertension requires serious attention because the death rate associated with this disease continues to increase. From 1975 to 2015, the number of adults with high blood pressure more than doubled, rising from 594 million to 1.13 billion, primarily in low- and middle-income countries. This increase is mainly due to population growth and aging [3].

Early detection through an Early Alert System (EAS) is a strategy for managing hypertension. However, the current EAS may face challenges in accurately categorizing patients due to the diverse risk factors for hypertension. While EAS primarily relies on quantitative data such as blood pressure, hypertension risk factors also include categorical data like family history, smoking habits, and dietary patterns [4]. Thus, a method is required to perform an early classification of hypertension risk based on the patient's condition, which includes a combination of categorical and numerical health indicators.

The healthcare sector is increasingly dependent on advances in data-driven information technology. These advancements play an important role in finding patterns in patient data, which in turn can help with disease diagnosis [5]. Clustering methods are one of the tools that can be used to analyze and group this data into meaningful categories. Clustering is a multivariate analysis technique. According to [6], multivariate analysis is a joint analysis with more than two variables in the observations. Clustering algorithms are usually applied only to categorical or numerical data. However, since patient data often contains both types (mixed data), traditional clustering algorithms such as K-Means and Hierarchical clustering face limitations. The K-Medoids and K-Prototypes algorithms are specifically designed to address this issue by clustering mixed categorical and numerical data.

The K-Medoids and K-Prototypes algorithms can be applied to classify the risk of hypertension by grouping patients into different risk categories based on their clinical data. The K-Medoids method is a development of the K-Means method where the mean is very susceptible to outliers. Outliers with extreme values can change the average of most of the data, causing data imbalance. According to [7], the K-means method will be more sensitive to data containing outliers because it uses the mean to measure the middle value. K-Medoids is a development of the K-Means algorithm, which uses median or medoid values as the cluster center, so it is more resistant to outliers [8]. K-Prototypes is a clustering algorithm that is an extension of K-Means and K-Modes, designed to handle clustering on data with both numerical and categorical attributes [9]. Therefore, this algorithm is suitable for use with the data in this research.

Several previous researches using K-Medoids include [10] conducting groupings based on toddler measles immunization data in Indonesia, [11] clustering ulcer disease in Karawang Regency, [12] clustering the spread of COVID-19 in Indonesia, [13] clustering crime patterns in Yogyakarta. Meanwhile, previous research that used K-Prototypes included [14] detecting mortality factors in heart failure patients and [15] conducting cluster analysis on various data sets using the K-Modes and K-Prototypes algorithms. Based on previous research and the characteristics of research data, which consists of mixed data, the author is interested in conducting scientific research titled “Implementation of K-Medoids and K-Prototypes Clustering for Early Detection of Hypertension Disease”.

2. RESEARCH METHODS

This research employs quantitative methods, gathering and analyzing data from the Haji Makassar Regional General Hospital. This data is provided with the permission of the hospital through the South Sulawesi Department of Investment and One-Stop Integrated Services (PTSP), starting from October 31,

2023. The data, encompassing 70 patient medical records from December 2022 to June 2023, focuses on patients diagnosed with hypertension in Makassar City. K-Medoids and K-Prototypes algorithms will then be applied to analyze this data.

2.1 K-Medoids Algorithm

The K-Medoids algorithm can overcome the problems of outliers, extreme data distances, and noise because the K-Medoids approach uses the median value of the data, so it remains stable even though there are outliers and noise. Therefore, the K-Medoids algorithm is more robust compared to K-Means [16]. The algorithm for clustering techniques using the K-Medoids method includes:

1. Initialize the number of centroids (number of clusters)
2. Allocate data to the nearest cluster by calculating the distance of each data using Euclidean distance

$$d\left(x_{ij}, y_{kj}\right) = \sqrt{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - y_{kj}\right)^2} \quad (1)$$

$d\left(x_{ij}, y_{kj}\right)$: Euclidean distance between the i^{th} observation of the j^{th} variable to the centroid of the k^{th} cluster on the j^{th}
x_{ij} : i^{th} observation, j^{th} variable
y_{kj} : centroid of the k^{th} cluster, j^{th} variable
p : number of variables
n : number of observations

3. Randomly select an observation object from each cluster as a new medoid
4. Calculate the distance of each observation object in each cluster to the new candidate medoids
5. Calculate the new total distance value and divide it by the old total distance to obtain the total deviation (S). If $S < 0$, swap the objects with the cluster data to form a new set of objects as medoids.
6. Repeat steps 3 to 5 until the medoid does not change again to get the cluster and its members.

Figure 1 below shows the flowchart of steps 1 through 6 of the K-Medoids algorithm

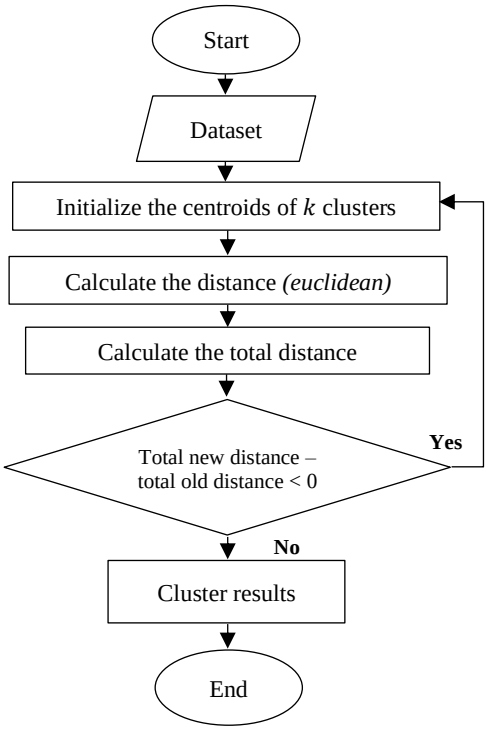


Figure 1. K-Medoids Flowchart

2.2 K-Prototypes Algorithm

K-Prototypes is a development of K-Means, a development carried out by [17] to maintain the efficiency of the K-means algorithm when handling large data. It can be applied to numerical and categorical data. The algorithm stages of the clustering technique using the K-Prototypes method include:

1. Determine the number of clusters (k)
2. Initialize prototypes by randomly assigning initial centroids, denoted as to represent the cluster centers
3. Compute the initial distances between all observations in the dataset and the initial cluster centroids. The core principle behind K-Prototypes is measuring the similarity between data objects and these cluster centroids (prototypes). The Euclidean distance metric is used for numerical data to compute this similarity. For categorical data, a different metric, the k-modes distance is employed [17]. The following is the distance calculation in the k-prototypes algorithm

$$d(X_i, Z_l) = X_i, Z_l \left(\sum_{j=1}^{m_r} \left(x_{ij}^r - z_{lj}^r\right)^2 \right) + \gamma l \sum_{j=l+1}^{m_c} \delta \left(x_{ij}^c - z_{lj}^c\right)^{\frac{1}{2}} \quad (2)$$

x_{ij}^r : The value of a specific numerical (r) attribute for observation object j
z_{lj}^r : The average (mean) or prototype value of the j^{th} numerical attribute (r) for all observation objects within cluster l
m_r : The total number of numerical (r) observation objects in the dataset
γl : A weight used for categorical observation of objects within the cluster l. This weight is calculated based on the standard deviation of the corresponding numerical attribute across all observation objects in the cluster.
x_{ij}^c : The value of the i^{th} observation on the j^{th} categorical attribute (c).
z_{lj}^c : The most frequent value (mode) of the jth categorical attribute (c) within cluster l
m_c : The total number of categorical (c) observation objects in the dataset

4. Assign the observation to clusters based on the closest prototype distance from the measured observations
5. Calculate the updated centroid for each cluster once all objects have received assignments
6. Move all data points in the dataset to the new prototypes
7. Repeat steps 2 through 6 until reaching the maximum number of iterations, which serves as a termination condition, or until the cluster centroid positions no longer change (convergence).

Figure 2 below shows the flowchart of steps 1 through 7 of the K-Prototypes algorithm

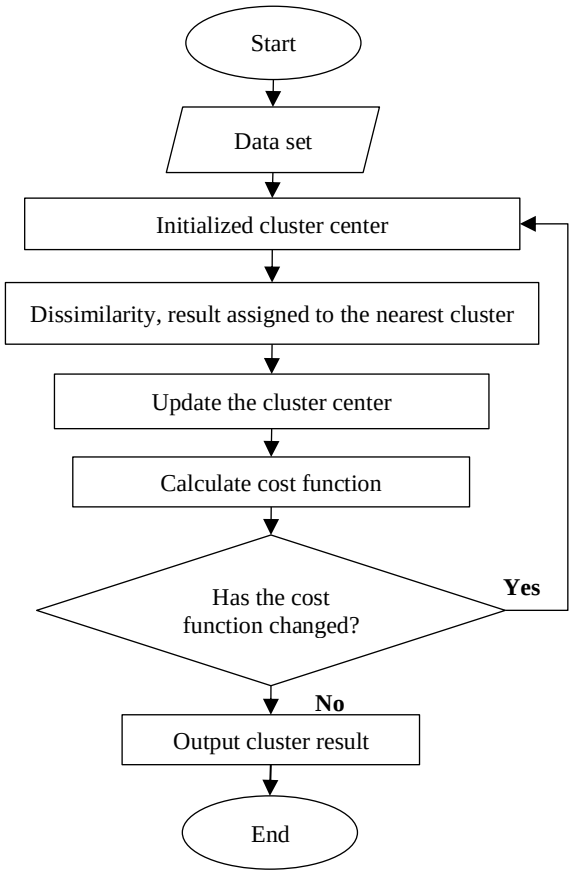


Figure 2. K-Prototypes Flowchart

2.3 Cluster Evaluation

1. Elbow Method

   When performing cluster analysis, the elbow method helps us identify the optimal number of clusters (k) to create. This method involves plotting a graph where the value increases on one axis and a measure of cluster quality (often explained as variance) increases on the other. The ideal value corresponds to an "elbow" shape in the graph. This elbow appears as a sharp decrease in the rate of improvement, indicating that adding more clusters provides diminishing returns. Alternatively, we can identify the elbow by comparing the percentage change in explained variance between adding additional clusters [18].

2. Silhouette Coefficient

   The average Silhouette value of each cluster formed will be used to evaluate the quality of the resulting cluster. The following is the Silhouette Coefficient calculation

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

   $\boldsymbol{s(i)}$ is the Silhouette Coefficient, $\boldsymbol{a(i)}$ is the average distance of the $\boldsymbol{i}^{th}$ object to all objects in A (the cluster where the $\boldsymbol{i}^{th}$ object is located), $\boldsymbol{b(i)}$ is the neighboring cluster of object ($\boldsymbol{i}$) which reaches the minimum value [19] [20]. To interpret the cluster validation results using the average value of the Silhouette Coefficient [21], the range of values and interpretations is in Table 1.

Table 1. Silhouette Coefficient Range and Interpretation

Silhouette Coefficient	Interpretation
0.71-1.0	Very Good
0.51-0.70	Good
0.26-0.50	Weak
< 0.25	No Applicable

According to [22], the Silhouette Coefficient is used to assess the level of confidence in the clustering process of an observation. The Silhouette Coefficient value is in the range -1 to 1, with the following interpretation:

a. Values greater than 0: These indicate well-grouped observations. The closer the coefficient gets to 1, the higher the confidence that the observation belongs to its assigned cluster.

b. Values less than 0: These suggest an observation might be misplaced and assigned to the wrong cluster.

c. Value of 0: This indicates the observation sits exactly on the border between two clusters, making its cluster assignment uncertain.

2.4 Research Variables

This research has six variables: (X_1) Gender, (X_2) Age, (X_3) Family History of Hypertension, (X_4) Smoking habits, (X_5) Pulse, and (X_6) The level of hypertension which can be seen in Table 2.

Table 2. Research Variables

Variable	Operational Definition	Variable Unit
X_1	The patient's gender consists of male and female.	Male / Female
X_2	Patient's age at the time of health examination.	Years
X_3	Whether or not there is a family history of hypertension can influence pulse frequency and increase blood pressure.	Present / Absent
X_4	A person who is an active smoker or not has a probability of developing hypertension.	Yes / No
X_5	Frequency or number of times the artery beats in one minute. The pulse occurs because of the heart pumping	(times/minute)
X_6	Increased blood pressure, both systolic and diastolic pressure. The level of hypertension consists of the risk classes Prehypertension, Grade 1 Hypertension, and Grade 2 Hypertension.	Prehypertension, Grade 1 Hypertension, and Grade 2 Hypertension.

2.5 Research Procedures

The stages of data analysis in this research are as follows:

1. Convert categorical variables into numerical representations. Converting these numbers allows one to use validation techniques such as the Elbow Method and Silhouette Coefficient when evaluating clusters.
2. Summarize and describe the pertinent features of the research data in terms of central tendency, dispersion, and shape of variables.
3. Apply the K-Medoids and K-Prototypes algorithms to partition the research data into meaningful clusters. Both techniques will group observations with similar characteristics together into clusters, exposing some latent patterns based on the underlying data structure.
4. Compute the average silhouette coefficients to evaluate the quality of the generated clusters. Higher average Silhouette Coefficient values indicate well-separated and well-defined clusters.
5. Derive findings and insights from the cluster analysis and validation output. In this final stage, we will analyze the nature of each cluster and infer insights about where specific patterns in our data may be located.

3. RESULTS AND DISCUSSION

3.1 Result

3.1.1 Exploratory Data Analysis

Table 3. Descriptive Statistics of Numerical Variables

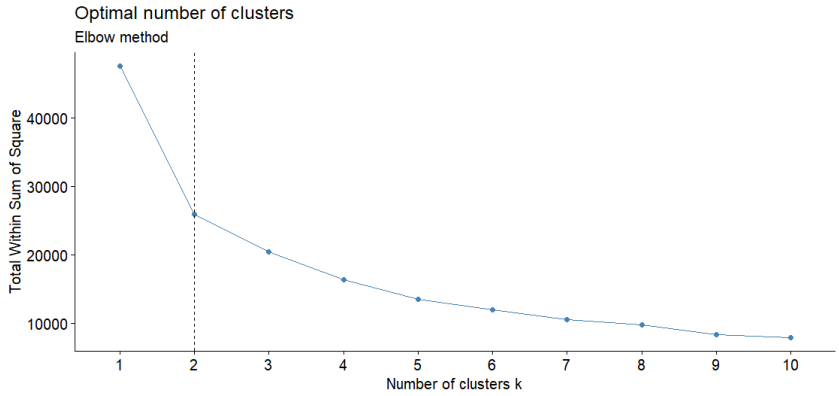
	Min	Median	Mean	Max
X_2	18	58	58.26	88
X_5	68	86	87.96	128

Table 4. Descriptive Statistics of Categorical Variables

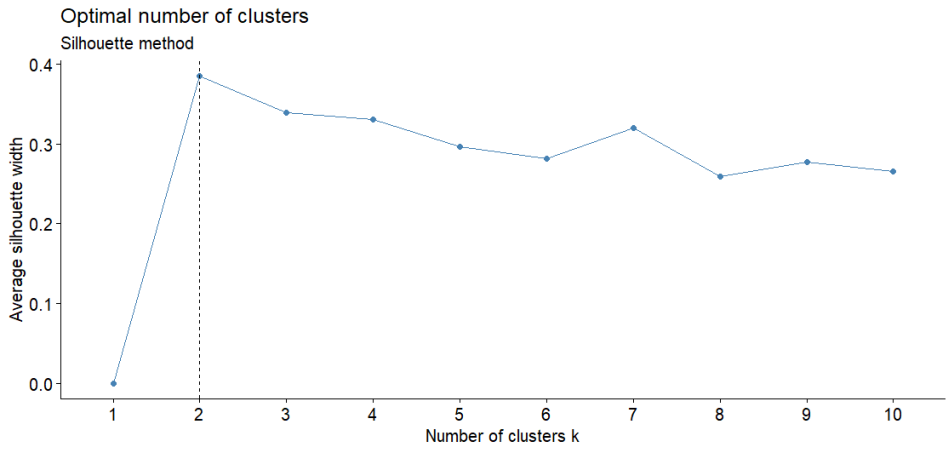
Frequency/ Number of Patients			
X_1	X_3	X_4	X_6
Male = 26	Present = 35	Yes = 25	Prehypertension = 21
Female = 44	Absent = 35	No = 45	Hypertension Stage 1 = 38
			Hypertension Stage 2 = 11

Table 3 indicates that individuals ranged in age from 18 to 88 years, with a mean age of 58.26. Pulse rates varied between a minimum of 68 times per minute and 128 times per minute. **Table 4** illustrates that 26 males and 44 females were in the study. Among these individuals, 35 had a family history of hypertension, while 35 did not. Additionally, 25 participants were smokers, while 45 were non-smokers. Of the total, 21 individuals had prehypertension, 38 had stage one hypertension, and 11 had stage two hypertension.

3.1.2 Determination of Optimal Clusters using the Elbow and Silhouette Coefficient Methods



(a)



(b)

Figure 3. Plot of Optimal Cluster Validation Results with (a) Elbow Method and (b) Silhouette Coefficient

Figure 3 illustrates the application of the elbow method to determine the optimal number of clusters. This method identifies the point on the graph where the rate of decrease in a cost function (often explained as variance) slows down significantly. This "elbow" is interpreted as the optimal number of clusters. In this study, the elbow appears at , indicating two clusters as the optimal solution. The findings from the silhouette method further reinforce this conclusion. The silhouette method also suggests two clusters as the optimal number, demonstrating consistency between the two approaches.

3.1.3 K-Medoids Cluster Results

The initial step involved identifying the best cluster number using the elbow and silhouette method. **Table 5** displays the data distribution resulting from the outcomes of the K-Medoids algorithm, as indicated by the plots from the elbow and silhouette methods.

Table 5. Distribution of Clustering Results Using the K-Medoids Algorithm

Number of Clusters	Number of Patients in Each Cluster				
	1	2	3	4	5
2	36	34			
3	32	11	27		
4	27	10	16	17	
5	11	18	15	9	17

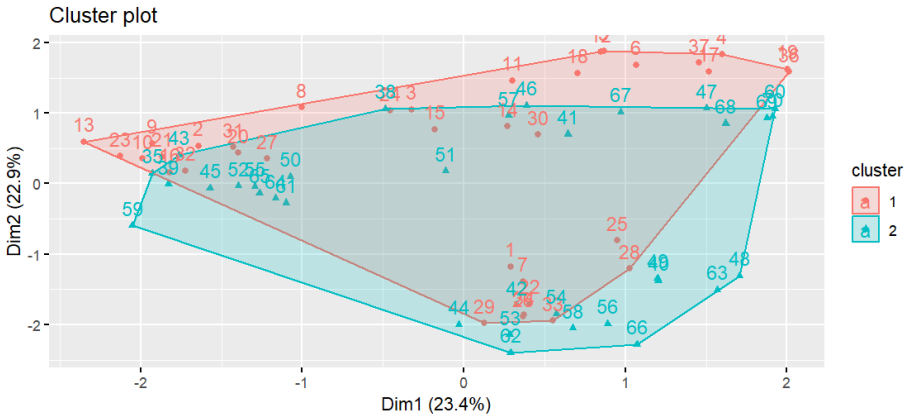


Figure 4. K-Medoids Cluster Results

Figure 4 is a cluster plot illustrating the distribution of patients, with each axis explaining a percentage of variability. In this plot, the two clusters are represented by different colors and shapes: cluster 1 is marked in red, and cluster 2 is marked in green. Each point on the graph represents a patient, and the lines surrounding each group of points indicate the cluster coverage area. Based on the clustering process in **Table 5** and **Figure 4**, using the K-Medoids clustering method, the number of patients in Cluster 1 (low risk) is 36, and Cluster 2 (high risk) is 34. Overall, the distribution of points in Figure 4 shows that the K-Medoids algorithm successfully separated the data into two main clusters with different patient distributions.

3.1.4 K-Prototypes Cluster Results

The K-Prototypes clustering approach requires specifying the initial number of clusters (k) before the clustering begins. This study opted to start with $k = 2$. The data used for the analysis, detailed in **Table 2**, comprises four categorical and two numerical variables. It is important to note that the choice of k can influence the clustering outcome. The algorithm iteratively refines the cluster centers until they stabilize, signifying the completion of the clustering process. **Table 6** presents the data distribution based on the final K-Prototypes clustering results.

Table 6. Distribution of Clustering Results Using the K-Prototypes Algorithm

Number of Clusters	Number of Patients in Each Cluster				
	1	2	3	4	5
2	35	35			
3	21	26	23		
4	15	18	21	16	
5	14	19	15	10	12

Identifying the most suitable number of clusters is crucial to ensure the results accurately represent real-world conditions. This study leverages the elbow and silhouette coefficient methods (presented in **Figure 5**) to determine the optimal number of clusters, which is two in this case.

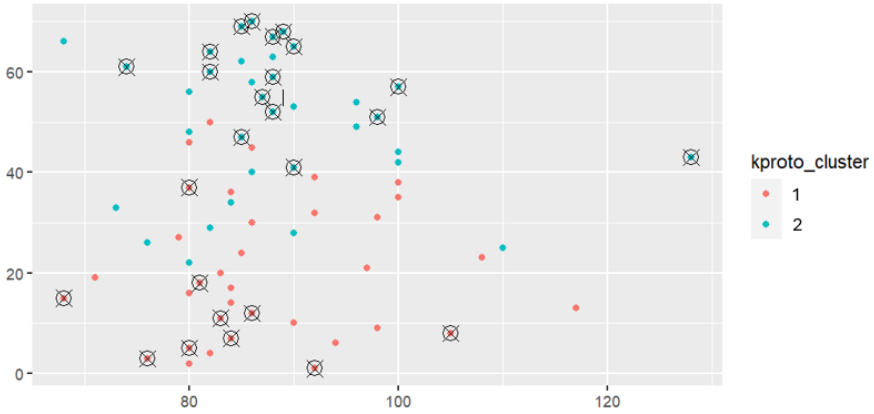


Figure 5. Cluster K-Prototypes Results with $k = 2$.

Based on the clustering process in **Table 5**, **Table 6**, and **Figure 5**, using the K-Prototypes clustering method where $k = 2$, the number of patients included in cluster 1 (low risk) is 35, and cluster 2 (high risk) is 35. **Figure 5** shows the clustering results using the K-Prototypes method with two main clusters ($\boldsymbol{k = 2}$). In this plot, patient data is grouped into two clusters, represented by different colors and symbols, where cluster 1 represents the low-risk group and cluster 2 represents the high-risk group. Each cluster contains 35 patients, indicating a balanced distribution between the two risk groups. Each point on the graph represents a single patient, with its position determined by two numerical attributes that have been reduced to two dimensions to facilitate visualization. The distribution of points within each cluster shows the variation in characteristics among patients within each group, with closer points representing patients who have more similar characteristics. The choice of two clusters was validated using the elbow and silhouette coefficient methods to ensure that the selected number of clusters best fits the existing data patterns.

3.1.5 Cluster Characteristics

The K-Medoids and K-Prototypes algorithms were used to cluster data, and the results are shown in **Table 7**, with 1 being cluster 1 (low risk) and 2 being cluster 2 (high risk).

Table 7. K-Medoids and K-Prototypes Cluster Results

Patient	X_1	X_2	X_3	X_4	X_5	X_6	K-Medoids	K-Prototypes
1	Female	78	Absent	No	92	Prehypertension	1	1
2	Female	53	Present	No	80	Hypertension Stage 1	1	1
3	Female	58	Present	Yes	76	Hypertension Stage 2	1	1
4	Male	68	Absent	Yes	82	Hypertension Stage 1	1	1
5	Male	68	Present	Yes	80	Hypertension Stage 2	2	1
6	Male	55	Absent	Yes	94	Hypertension Stage 1	1	1
7	Female	68	Absent	No	84	Prehypertension	1	1
8	Female	61	Present	Yes	105	Hypertension Stage 2	2	1
9	Female	61	Present	No	98	Hypertension Stage 1	1	1
10	Female	45	Present	No	90	Hypertension Stage 1	1	1
⋮								
30	Female	61	Absent	Yes	86	Hypertension Stage 1	1	2
31	Female	80	Present	No	98	Hypertension Stage 1	2	2
32	Female	54	Present	No	92	Hypertension Stage 1	1	2
33	Female	51	Absent	No	73	Prehypertension	1	1
34	Female	57	Absent	No	84	Prehypertension	1	1
35	Female	53	Present	No	100	Hypertension Stage 1	1	2
⋮								
66	Female	59	Absent	No	68	Prehypertension	2	1
67	Male	60	Present	Yes	88	Hypertension Stage 2	2	2
68	Male	53	Absent	Yes	89	Hypertension Stage 1	2	2
69	Male	61	Absent	Yes	85	Hypertension Stage 1	2	2
70	Male	64	Absent	Yes	86	Hypertension Stage 1	2	2

Based on the cluster results in **Table 7** using K-Medoids and K-Prorotypes, patients in the low-risk cluster generally have lower pulse rates and blood pressure and fewer hypertension risk factors. Patients in the high-risk cluster tended to have higher pulse rates and blood pressure as well as more hypertension risk factors.

3.1.6 Evaluation of Cluster Results (Optimal Cluster Validation)

(a) **(b)**

Figure 6. Cluster Evaluation Based on the Average Silhouette Coefficient for (a) K-Medoids (b) K-Prototypes

Table 8. Silhouette Coefficient of K-Medoids and K-Prototypes

Cluster	K-Medoids		K- Prototypes	
	Number of Patients	Average Silhouette Coefficient	Number of Patients	Average Silhouette Coefficient
1 = Low Risk	36	0.38	35	0.26
2 = High Risk	34	0.39	35	0.27

Figure 6 and **Table 8** reveal the average Silhouette Coefficient for each clustering method. The K-Medoids method separates patients into low-risk (cluster 1) and high-risk (cluster 2) groups with Silhouette Coefficients of 0.38 and 0.39, respectively. While the K-Prototypes method achieves similar cluster separation (0.26 for low-risk and 0.27 for high-risk), its average Silhouette Coefficient is lower. Although both methods have an average coefficient greater than 0, indicating some grouping level, the values are closer to 0 than 1. It suggests weak cluster separation, meaning observations within each cluster may differ. By comparing the average Silhouette Coefficients, we can infer that the K-Medoids method performs better in this study. It yields a higher average coefficient, suggesting a more apparent distinction between low-risk and high-risk patient groups.

3.2 Discussion

Hospitals frequently diagnose hypertension by stages, primarily based on blood pressure measurements **[1]**. Although this method works quite well, it is a one-dimensional approach, as it considers only one risk indicator.

The clustering method in this study overcomes this limitation and considers several attributes (e.g., gender, age, family history of hypertension, smoking habits, pulse rate), thus providing a better overview of hypertension risk. As the clustering approach simultaneously assesses categorical and numerical variables, this multidimensionality of data could enable different hospital systems to prioritize interventions on a more tailored patient group that is low-risk or high-risk based on clusters rather than scores. This is in accordance with **[2]** which highlights the utility of a multifactorial risk assessment approach when screening hypertension to improve both early detection and treatment capabilities.

The results from both clustering methods reveal their specific strengths in handling mixed data types, which are typical in clinical datasets. The K-Medoids algorithm showed better clustering performance, with higher Silhouette Coefficients (0.38 for low-risk and 0.39 for high-risk clusters) than K-Prototypes (0.26 and 0.27, respectively). This in line with the findings by **[23]**, which highlighted K-Medoids' robustness to outliers, a common occurrence in medical data due to extreme variations in patient health indicators. Using medoids as cluster centers in K-Medoids ensures robustness against such data irregularities, which can distort traditional cluster means, as seen in K-Means or K-Prototypes when dealing with continuous variables.

K-Prototypes is proficient at managing mixed categorical and numerical data types, consistent with its initial design to handle extensive datasets with varied properties [17]. Clinical settings with both data types require this capacity to cluster based on thorough health profiles. K-Prototypes balanced patient categorization using mixed data features, resulting in 35 equal low-risk and high-risk people. This outcome corroborates the conclusions of [15], which highlighted the efficacy of K-Prototypes in handling datasets comprising both categorical and numerical variables for thorough pattern recognition.

Previous studies have also applied clustering methods to health data for risk assessment. For instance, [2] utilized a Neighbor Weighted K-Nearest Neighbor (NWKNN) approach to classify hypertension risk but found limitations in accommodating mixed data types effectively. In comparison, using K-Medoids and K-Prototypes in this study allowed for more flexible and robust handling of patient data, particularly where categorical (e.g., smoking habits, family history) and numerical (e.g., pulse rate, blood pressure) variables are present. This approach captures more dimensions of risk and enhances the accuracy and relevance of the clusters generated, offering a solution to the limitations observed in single-type clustering methods.

Another related study by [24] explored clustering in health data using Density-Based Spatial Clustering of Applications with Noise (DBSCAN), focusing on outlier detection in patient groups. Although DBSCAN proficiently detects outliers, it lacks direct support for mixed data clustering, a constraint remedied by K-Medoids and K-Prototypes in this study. K-Medoids and K-Prototypes methods may improve hypertension risk categorization by ensuring that clusters include critical patient characteristics and are robust against outliers, especially in varied patient populations.

The findings of this study suggest that K-Medoids and K-Prototypes clustering methods could significantly enhance hypertension risk stratification. Unlike traditional hospital classifications solely based on blood pressure readings, the clusters produced here provide a more comprehensive risk assessment that considers multiple health factors. More tailored patient management strategies may assist healthcare practitioners in identifying high-risk patients beyond blood pressure indicators like lifestyle and family history.

4. CONCLUSIONS

Based on the findings from research employing the K-Medoids and K-Prototypes algorithms for early hypertension detection, several conclusions were drawn as follows:

a. The evaluation results of the elbow and Silhouette Coefficient methods show that the optimal number of clusters is two. These two methods are consistent in determining the correct number of clusters. K-Medoids clustering groups patients into two clusters: cluster 1 with low risk consisting of 36 patients and cluster 2 with high risk consisting of 34 patients. K-Prototypes clustering also groups patients into two clusters: cluster 1 with low risk consisting of 35 patients and cluster 2 with high risk consisting of 35 patients.

b. The average Silhouette Coefficient value for the K-Medoids method is 0.38 for the low-risk cluster and 0.39 for the high-risk cluster. The average Silhouette Coefficient value for the K-Prototypes method is 0.26 for low-risk and 0.27 for high-risk clusters. An average Silhouette Coefficient exceeding 0 suggests that grouping the observations is possible; however, the resulting clusters remain weak. Further implementation and improvement of clustering methods can provide more accurate and valuable results.

ACKNOWLEDGMENT

The authors would like to thank the Makassar City Haji Regional General Hospital for providing data to conduct this research.

REFERENCES

[1] T. Unger, *et al.*, “International Society of Hypertension Global Hypertension Practice Guidelines,” *Hypertension*, vol. 75,

no. 6, pp. 1334–1357, 2020, doi: https://doi.org/10.1161/hypertensionaha.120.15026.
[2] B. L. Yudha, L. Muflikhah, and R. C. Wihandika, "Klasifikasi Risiko Hipertensi Menggunakan Metode Neighbor Weighted K- Nearest Neighbor (NWKNN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 2, pp. 897–904, 2018.
[3] K. T. Mills, A. Stefanescu, and J. Hea, "The Global Epidemiology of Hypertension," *Nat. Rev. Nephrol.*, vol. 16, no. 4, pp. 223–237, 2020, doi: 10.1038/s41581-019-0244-2.
[4] D. O. Ondimu, G. M. Kikuvi, and W. N. Otieno, "Risk Factors for Hypertension among Young Adults (18-35) Years Attending in Tenwek Mission Hospital, Bomet County, Kenya in 2018," *Pan African Med. Journal*, vol. 33, no. 210, pp. 1–8, 2019, doi: 10.11604/pamj.2019.33.210.18407.
[5] B. Setiaji and P. A. K. Pramudho, "Pemanfaatan Teknologi Informasi Berbasis Data Dan Jurnal Untuk Rekomendasi Kebijakan Bidang Kesehatan," *Heal. J. Inov. Ris. Ilmu Kesehat.*, vol. 1, no. 3, pp. 166–175, 2022, doi: 10.51878/healthy.v1i3.1649.
[6] R. A. Johnson and D. W. Wicheren, *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.
[7] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley, 1990.
[8] H. Nabila, D. Retno, and S. Saputro, "Clustering Data Campuran Numerik dan Kategorik Menggunakan Algoritme Ensemble Quick RObust Clustering using linKs (QROCK)," *Prism. Pros. Semin. Nas. Mat.*, vol. 5, no. 1, pp. 716–720, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/article/view/54590
[9] Z. R. Fadilah and A. W. Wijayanto, "Perbandingan Metode Klasterisasi Data Bertipe Campuran: One-Hot-Encoding, Gower Distance, dan K-Prototype Berdasarkan Akurasi (Studi Kasus: Chronic Kidney Disease Dataset)," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 57–67, 2023, doi: 10.30871/jaic.v7i1.5857.
[10] S. Sundari, I. S. Damanik, A. P. Windarto, H. S. Tambunan, J. Jalaluddin, and A. Wanto, "Analisis K-Medoids Clustering Dalam Pengelompokkan Data Imunisasi Campak Balita di Indonesia," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 687, 2019, doi: 10.30645/senaris.v1i0.75.
[11] S. Nurlaela, A. Primajaya, and T. N. Padilah, "Algoritma K-Medoids Untuk Clustering Penyakit Maag Di Kabupaten Karawang," *I N F O R M a T I K a*, vol. 12, no. 2, p. 56, 2020, doi: 10.36723/juri.v12i2.234.
[12] S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. I. R.H.Zer, and D. Hartama, "Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia," *J. Teknol. Inf.*, vol. 4, no. 1, pp. 166–173, 2020, doi: 10.36294/jurti.v4i1.1296.
[13] E. H. S. Atmaja, "Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta," *Int. J. Appl. Sci. Smart Technol.*, vol. 1, no. 1, pp. 33–44, 2019, doi: 10.24071/ijasst.v1i1.1859.
[14] R. Novidianto and K. Fithriasari, "Algoritma ClusterMix K-Prototypes Untuk Menangkap Karakteristik Pasien Berdasarkan Variabel Penciri Mortalitas Pasien Dengan Gagal Jantung," *Inferensi*, vol. 4, no. 1, p. 37, 2021, doi: 10.12962/j27213862.v4i1.8479.
[15] R. Madhuri, M. R. Murty, J. V. R. Murthy, P. V. G. D. P. Reddy, and S. C. Satapathy, "Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms," *Adv. Intell. Syst. Comput.*, vol. 249 VOLUME, pp. 137–144, 2014, doi: 10.1007/978-3-319-03095-1_15.
[16] B. Prihasto, D. Darmansyah, D. P. Yuda, F. M. Alwafi, H. N. Ekawati, and Y. P. Sari, "Comparative Analysis of K-Means and K-Medoids Clustering Methods on Weather Data of Denpasar City," *J. Pendidik. Multimed.*, vol. 5, no. 2, pp. 91–114, 2023, doi: 10.17509/edsence.v5i2.65925.
[17] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283-304," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998, [Online]. Available: https://www.researchgate.net/publication/220451944_Huang_Z_Extensions_to_the_k-Means_Algorithm_for_Clustering_Large_Data_Sets_with_Categorical_Values_Data_Mining_and_Knowledge_Discovery_2_283-304
[18] P. M. Hasugian, B. Sinaga, J. Manurung, and S. A. Al Hashim, "Best Cluster Optimization with Combination of K-Means Algorithm And Elbow Method Towards Rice Production Status Determination," *Int. J. Artif. Intell. Res.*, vol. 5, no. 1, pp. 102–110, 2021, doi: 10.29099/ijair.v6i1.232.
[19] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[20] H. Hafid and A. Arisandi, "Klasifikasi Penggunaan Teknologi Pada Petani Milenial di Sulawesi Selatan Menggunakan Density Based Spatial Clustering Algorithm With Noise," *J. Math. Theory Appl.*, vol. 6, no. 1, pp. 104–113, 2024, doi: 10.31605/jomta.v6i1.3623.
[21] B. N. Sari and A. Primajaya, "Penerapan Clustering Dbscan Untuk Pertanian Padi Di Kabupaten Karawang," *J. Inform. dan Komput.*, vol. 4, no. 1, pp. 28–34, 2019, [Online]. Available: https://ejournal.akakom.ac.id/index.php/jiko/article/view/178%0Awww.mapcoordinates.net/en.
[22] G. Brock, V. Pihur, S. Datta, and S. Datta, "ClValid: An R package for cluster validation," *J. Stat. Softw.*, vol. 25, no. 4, pp. 1–22, 2008, doi: 10.18637/jss.v025.i04.
[23] N. Rizqia and P. Ratnasari, "Comparative Study of k-Mean , k-Medoid , and Hierarchical Clustering," vol. 5, no. 2, pp. 9–20, 2023.
[24] H. Santoso, "Case Base Reasoning Untuk Mendiagnosis Penyakit Hipertensi Menggunakan Metode Indexing Density Based Spatial Clustering Application With Noise (DBSCAN)," *ETHOS (Jurnal Penelit. dan Pengabdian)*, vol. 7, no. 1, pp. 88–100, 2019, doi: 10.29313/ethos.v7i1.4206.