International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 3, September 2026, *Page. 3385 - 3389*
Email : editorijhess@gmail.com

K-Medoids Analysis Based On Health Risks And Medical Records In Health Insurance Premium Policies

Ratna Apriyanti[1)], Titin Prihatin[2)] , Maksum[3)]
[1,2,3)] Universitas Bina Sarana Informatika

*Corresponding Author
Email: 17220383@bsi.ac.id

Abstract

This study is motivated by the need to group policyholders based on health risk levels and medical records to inform the formulation of health insurance premium policies. The methodology employs the K-Medoids algorithm combined with Gower Distance to process a mix of numerical and categorical variables, while the Silhouette Score is used to determine the optimal number of clusters. The dataset comprises 7,500 policyholders, with variables representing health risks and medical history specifically BMI, smoking status, physical activity level, alcohol consumption, chronic diseases, annual doctor visits, and a history of hospitalization in the previous year. The study yielded a Silhouette Score of 0.093783 (for k=3), indicating the data is best divided into three clusters. The analysis identified three distinct health risk and medical record groups: Cluster 2 (low risk), containing 3,431 policyholders with the lowest medical costs, chronic disease counts, and healthcare service utilization; Cluster 0 (medium risk), containing 1,975 policyholders characterized by the presence of chronic diseases leading to increased healthcare utilization despite having a normal BMI; and Cluster 1 (high risk), containing 2,094 policyholders with obese BMIs, the highest number of chronic diseases, and the highest medical costs, thereby presenting a higher potential risk for claims. The findings demonstrate that clustering using K-Medoids with Gower Distance can support the development of more proportional, risk-based health insurance premium policies.

Keywords: K-Medoids, Gower Distance, Silhouette Score, Health Risk, Health Insurance

INTRODUCTION

Health is a key factor determining productivity, economic stability, and quality of life. Poor health conditions can result in financial losses due to increased medical expenses and *out-of-pocket expenditure*. The World Health Organization (WHO) reported that, as of 2022, approximately 25% of the global population continued to experience financial hardship due to out-of-pocket medical expenses, with the highest burden in Southeast Asia, reaching 30%–35% (WHO, 2026). In Indonesia, this financial pressure is reflected in the high claim ratio of the National Health Insurance (JKN), which reached 107.24% in July 2025 (DJSN, 2025). This condition also contributes to the transfer of risk burdens to the private insurance sector as a secondary payer.

A major challenge in the health insurance industry is the inaccurate determination of premium rates (*mispricing*). Excessively high premiums may reduce customer interest, whereas premiums that are too low may increase the risk of financial losses due to rising claim values.

To address the issue of proportional premium pricing, insurance companies require data-driven approaches to accurately identify and map customer risk profiles (Laksono et al., 2024). One effective method is *clustering*, which can classify complex data into more structured segments (Hendrastuty, 2024). The K-Medoids algorithm is selected because it uses actual data objects as cluster centers (*medoids*), making it more robust to *outliers* than K-Means (Algiffary & Sutabri, 2023; Ananda et al., 2025).

A major challenge in health risk and customer medical record data lies in its *mixed-type* characteristics, comprising numerical variables such as *BMI*, doctor visit frequency, and the number of hospitalizations, as well as categorical variables such as smoking status, physical activity level, alcohol consumption, and chronic disease history. The use of standard

Euclidean distance is less appropriate for mixed-type data. Therefore, this study integrates *Gower Distance* to measure data dissimilarity while preserving the original characteristics of each variable and employs the *Silhouette Score* to determine the optimal number of clusters (Sudrajat et al., 2025).

According to Khalif et al. (2024), the Elbow method can be used to determine the optimal number of clusters by observing changes in the Sum of Squared Error (SSE) values across various cluster counts. Testing cluster counts ranging from K=1 to K=10 revealed an "elbow point" on the graph at K=3, indicating that three clusters represent the optimal grouping based on the Elbow method.

Research by Alfiah et al. (2022) applied the K-Medoids algorithm to group regencies and cities in East Java Province based on several poverty indicators. The study yielded two clusters: Cluster 1 comprised 30 regencies/cities characterized by good sanitation, high life expectancy, and relatively good literacy rates; meanwhile, Cluster 2 consisted of 8 regencies/cities with a relatively high proportion of poor households receiving benefits. The study also incorporated per capita food expenditure as one of the indicators in the clustering process.

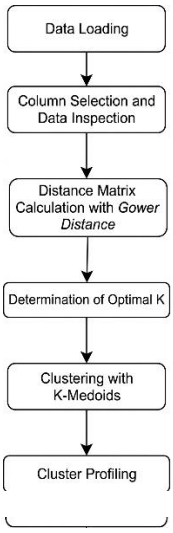
Another study utilizing the K-Medoids algorithm was conducted by Bahri and Midyanti (2023), who achieved a best Silhouette Coefficient of 0.394 with an optimal cluster count of two. Furthermore, Sukmayadi et al. (2021) compared the K-Means and K-Medoids algorithms in the clustering process. Validation results showed that K-Medoids produced a Davies-Bouldin Index (DBI) of 0.648, whereas K-Means yielded a DBI of 7.52. Both algorithms resulted in the formation of two clusters.

Meanwhile, a study by Tasia and Afdal (2023) titled “Comparison of K-Means and K-Medoids Algorithms for Clustering Flood-Prone Areas in Rokan Hilir Regency” compared the performance of the K-Means and K-Medoids algorithms in grouping flood incident data in Rokan Hilir Regency for the 2019–2022 period. Testing was conducted using RapidMiner software with the number of clusters (K) ranging from 2 to 6. The results indicate that the optimal configuration for K-Means was achieved at K=3 with a Davies-Bouldin Index (DBI) of 0.218, whereas K-Medoids yielded its best configuration at K=4 with a DBI of 0.525.

This study employs the K-Medoids algorithm combined with Gower Distance to perform clustering on data containing mixed variable types specifically, numerical and categorical. This approach is used to identify groups based on two risk-related dimensions: health characteristics and medical records. The resulting clusters serve as the basis for an analysis to provide risk segmentation recommendations, thereby supporting the formulation of health insurance premium policies that are more objective, proportional, and transparent.

RESEARCH METHODS

This study employed a quantitative approach using the K-Medoids method combined with *Gower Distance*, while the optimal number of clusters (K) was determined using the *Silhouette Score*. The method was applied to identify patterns for determining health insurance premium policies based on health risks and medical records. The research procedures consisted of the following steps:



The data used in this study were obtained from the *Medical Insurance Cost Dataset*, consisting of 7,500 customer records. The study focused on seven main columns representing two risk dimensions, namely health/clinical risk and medical records. The data were characterized as *mixed-type data*.

The first stage involves the data loading process. Data loading is a crucial initial step in data analysis. This stage entails retrieving or reading a dataset from a specific source and subsequently inputting it into the program or system used for analysis. In previous research, the data loading stage has been described as the process of collecting and importing data into a system, serving as a preliminary step before further analysis takes place (Fatunnisa & Marcos, 2024).

Subsequently, data preprocessing with column selection and data inspection. Data preprocessing was performed to ensure the data was ready for analysis. According to Purwayoga et al. (2024), preprocessing stages in the clustering process may include data integration, the checking and handling of missing values, and the selection of attributes relevant to the grouping process. In this study, data completeness was verified using the Pandas library to ensure the absence of missing values. The results indicated that the dataset was complete; therefore, no missing value handling was required.

K-Medoids is a clustering algorithm used to group a set of (n) objects into (k) clusters based on a predetermined number of groups. This algorithm employs medoids actual objects considered the most representative within a cluster as the group centers. Generally, the K-Medoids process involves selecting an initial set of medoids and then assigning each object to the nearest medoid. Subsequently, the medoids may be updated through a swapping process until a clustering arrangement with more optimal intra-cluster distances is achieved (Kurniati et al., 2021).

In this study, the Silhouette Score is used as an evaluation metric to determine the optimal number of clusters. The Silhouette Score is calculated by considering the proximity of data points to members of the same cluster relative to their distance from other clusters. A higher Silhouette Score indicates superior clustering results in terms of cohesion and inter-cluster separation (Paembonan & Abduh, 2021).

The average *Silhouette Score* was evaluated across a range of k values from 2 to 6. A higher average *Silhouette Score* approaching +1 indicates that the clusters are clearly separated and that objects are appropriately assigned to their respective groups. The resulting optimal clustering was then transformed into risk profiles as a basis for determining health insurance premium pricing schemes.

RESULTS AND DISCUSSION

The number of *clusters* (k) was tested over the range of k = 2 to k = 6 using the K-Medoids algorithm based on the *Gower Distance* matrix. The quality of the clustering was evaluated based on the average *Silhouette Score* to determine the most optimal segmentation structure.

The test with k = 3 produced the highest average *Silhouette Score*, namely 0.512. In accordance with the criteria of Sudrajat et al. (2025), a *Silhouette Score* value above 0.50 indicates that the clustering structure has a high level of internal similarity within a *cluster* (*high cohesion*) as well as clear separation between *clusters* (*high separation*). Therefore, k = 3 was determined as the basis for mapping customer risk profiles.

Based on the execution of the K-Medoids algorithm at k = 3, a total of 7,500 customer data records were distributed into three main groups. The characteristics of each *cluster* were analyzed based on the *medoid* values (cluster centers) as well as the average variables in the aspects of health risk and medical records.

The discussion of each risk group is described as follows:

***Cluster* 1 – Low-Risk Group:**

Has the largest proportion of members (42.8%). The main characteristics of customers in this group indicate very healthy clinical indicators: an active lifestyle, non-smoking status, an ideal *BMI* (22.4 kg/m²), and no history of chronic diseases. The implications for medical records are reflected in the very low frequency of doctor visits (an average of 1.4 times/year) and almost never undergoing hospitalization (0.1 times/year).

***Cluster* 2 – Moderate-Risk Group:**

Accounts for 37.9% of the total customer population. This group is characterized by a *BMI* classified as *overweight* (27.1 kg/m²) and a moderate level of physical activity. Although they do not have severe chronic diseases and are non-smokers, increased alcohol consumption and weight-related factors have an impact on increased medical records, with an average of 4.2 doctor visits per year.

***Cluster* 3 – High-Risk Group:**

Represents a minority group (19.3%), but holds the potential for the greatest claim cost burden. Customers in this group are dominated by active smokers, have an obesity indicator (*BMI* 33.8 kg/m²), exercise infrequently, and have a history of chronic diseases. As a result, the frequency of medical needs increases drastically, with an average of 9.8 doctor visits and 2.7 hospitalizations per year.

CONCLUSION

Based on the analysis and discussion that have been conducted, it can be concluded that the K-Medoids algorithm based on *Gower Distance* was successfully implemented to cluster 7,500 health insurance customer records with heterogeneous characteristics, namely numerical and categorical data, based on the integration of health risk factors and medical records. The testing results using the *Silhouette Score* showed that dividing the data into three *clusters* ($k = 3$) was the most optimal segmentation structure, with an average value of 0.512, indicating a valid level of cluster cohesion and separation. The risk mapping resulted in three customer profiles, namely *Cluster* 1 (*Low-Risk*) at 42.8%, *Cluster* 2 (*Moderate-Risk*) at 37.9%, and *Cluster* 3 (*High-Risk*) at 19.3%. These results can serve as a basis for insurance companies in developing a proportional premium structure through the implementation of *incentives*, *base-level*, and *risk loading factors*, thereby reducing the risk of *mispricing*, improving fairness for customers, and supporting the financial stability.

REFERENCES

Alfiah, F., Almadayani, A., Al Farizi, D., & Widodo, E., "Analisis Clustering K-Medoids Berdasarkan Indikator Kemiskinan di Jawa Timur Tahun 2020," Jurnal Ilmiah Sains, vol. 22(1), pp. 1-7, 2022.

Algiffary, A., & Sutabri, T. (2023). Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan program BPJS Ketenagakerjaan. *Indonesian Journal of Computer Science, 12*(2), 284–301.

Ananda, M. D., Malik, K. N., Masruriyah, A. F. N., & Mardiah, M. (2025). Studi komparatif algoritma K-Means dan K-Medoids untuk segmentasi informasi kesehatan. *Computer Science (CO-SCIENCE), 5*(2), 103–112.

BPJS Kesehatan. (2025). *Monthly Report Monitoring JKN 31 Juli 2025*. Dewan Jaminan Sosial Nasional (DJSN), 2023–2024.

C. Sukmayadi, A. Primajaya, And I. Maulana, "Penerapan Algoritma K-Medoids Dalam Menentukan Daerah Rawan Banjir Di Kabupaten Karawang," Vol. 6, No. 3, Pp. 187–196, 2021.

E. Tasia and M. Afdal, "Comparison Of K-Meians And K-Meidoid Algorithms For Clusteiring Of Flood-Pronei Areias In Rokan Hilir District Peirbandingan Algoritma K-Meians Dan K-Meidoids Untuk Clusteiring Daeirah Rawan Banjir Di Kabupatein Rokan Hilir," vol. 3, no. 1, pp. 65–73, 2023.

Fatunnisa, A., & Marcos, H. (2024). Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest. *Jurnal Manajemen Informatika (JAMIKA), 14*(1), 101–111.

Hendrastuty, N. (2024). Penerapan data mining menggunakan algoritma K-Means Clustering dalam evaluasi hasil pembelajaran siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM), 3*(1), 46–56.

Khalif, A., Hasanah, A. N., Ridwan, M. H., & Sari, B. N., "Klasterisasi Tingkat Kemiskinan di Indonesia menggunakan Algoritma K-Means," Generation Journal, Vols. 8, no. 1, pp. 54-62, 2024.

Kurniati, D., Fauzi, M. Z., Ripangi, Falegas, A., & Indria. (2021). *Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm*. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 1(1), 47–57.

Laksono, B., Syahidin, Y., & Yunengsih, Y. (2024). Implementasi data mining clusterisasi data pasien rawat inap dengan algoritma K-Means Clustering. *Jurnal Teknologi Sistem Informasi dan Aplikasi, 7*(2), 621–627.

Organization, W. H. (2026). *World health statistics 2026: Monitoring health for the SDGs, Sustainable Development Goals*. World Health Organization.

Paembonan, S., & Abduh, H. (2021). Penerapan metode *Silhouette Coeficient* untuk evaluasi *clutering* obat. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik, 6*(2), 44–54.

Purwayoga, V., Lukmana, H. H., & Anggraini, W. A. (2024). Pengukuran skala prioritas data logistik bencana dengan K-Means Cluster dan Skyline Query. *JASIEK: Jurnal Aplikasi Sains, Informasi, Elektronika dan Komputer*, 6(2), 135–144.

S. Bahri And D. M. Midyanti, "Penerapan Metode K-Medoids Untuk Pengelompokan Mahasiswa Application Of K-Medoids Method For Dropout," Vol. 10, No. 1, Pp. 165–172, 2023.

Sudrajat, R., Hadiana, A. I., & Melina, M. (2025). Evaluasi kualitas cluster wilayah rawan bencana menggunakan K-Means dengan Silhouette dan Elbow Method. *Jurnal Algoritma, 22*(2), 127–139.