RESEARCH ARTICLE • OPEN ACCESS

Cluster Analysis of BPJS Kesehatan Claim Data in Madiun City to Identify High Claim Patterns and Fraud Indications

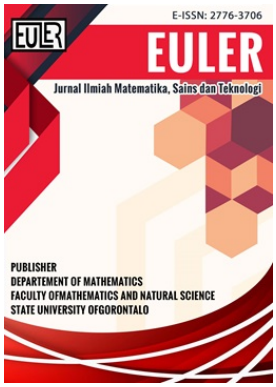
Muhammad Qolbi Shobri, Putri Balqis Al-Kubro, and Gabriella Vindy Kawuri

Volume 13, Issue 3, pp. 419–424, Dec. 2025

Received 25 October 2025, Revised 26 November 2025, Accepted 3 December 2025, Published 6 December 2025

To Cite this Article : M. Q. Shobri, P. B. Al-Kubro, and G. V. Kawuri,"Cluster Analysis of BPJS Kesehatan Claim Data in Madiun City to Identify High Claim Patterns and Fraud Indications", *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 13, no. 3, pp. 419–424, 2025, *https://doi.org/10.37905/euler.v13i3.35013*

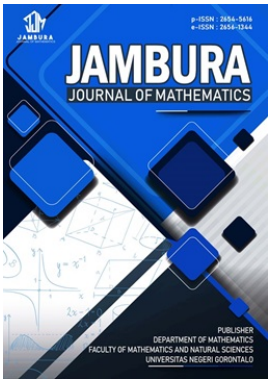
JOURNAL INFO • EULER : JURNAL ILMIAH MATEMATIKA, SAINS DAN TEKNOLOGI



Homepage	:	http://ejurnal.ung.ac.id/index.php/euler/index
Journal Abbreviation	:	Euler J. Ilm. Mat. Sains dan Teknol.
Frequency	:	Three times a year
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/euler
Online ISSN	:	2776-3706
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/euler/oai
Google Scholar ID	:	QF_r-_gAAAAJ
Email	:	euler@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics



Jambura Journal of Mathematics

Jambura Journal of Mathematics Education

Jambura Journal of Probability and Statistics

Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi, Volume 13, Issue 3, Pages 419–424, Dec. 2025 *https://doi.org/10.37905/euler.v13i3.35013*

Research Article

Check for updates

Cluster Analysis of BPJS Kesehatan Claim Data in Madiun City to Identify High Claim Patterns and Fraud Indications

Muhammad Qolbi Shobri[1,*], **Putri Balqis Al-Kubro**[1], **Gabriella Vindy Kawuri**[2]

[1]*Department of Actuarial Science, Muhammadiyah University of Madiun, Madiun 63137, Indonesia*
[2]*Department of Informatics, Muhammadiyah University of Madiun, Madiun 63137, Indonesia*

ARTICLE HISTORY
Received 25 October 2025
Revised 26 November 2025
Accepted 3 December 2025
Published 6 December 2025

KEYWORDS
BPJS Kesehatan (Health Social Security)
K-Means
Hierarchical Clustering
High Claim
Potential Fraud

ABSTRACT. *The increasing number of BPJS Kesehatan (Health Social Security) service claims in Madiun City poses significant challenges to financing efficiency and raises concerns about potential irregular or fraudulent claims. This study aims to identify high-claim patterns and detect indications of fraud using a data mining approach through the K-Means and Hierarchical Clustering methods. The research employed secondary data consisting of 309 hospital claim records from Madiun City in 2025. The primary variables were the number of claims and total claim costs, supported by additional variables such as age, gender, occupation, type of service, and disease diagnosis. Data analysis involved three main stages: preprocessing, clustering, and cluster quality evaluation using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Ths Study further compared the performance of both clustering methods, revealing that K-Means achieved superior validity scores across major evaluation metrics. The K-Means method produced the best performance, with a Silhouette Score of 0.617 and a Calinski-Harabasz Index of 419.581, reflecting well-separated and compact cluster structures. Three main clusters were identified-low, medium, and high. The high-claim cluster consisted of participants aged 55 years and above, with a claim frequencies of 2 to 7 claims and total claim costs exceeding IDR 20 million. This cluster was dominated by retirees, housewives, and private-sector employees utilizing inpatient services. Although categorized as a high-risk group, verification results revealed no signs of fraud but rather complex medical needs. These findings suggest that integrating clustering analysis into BPJS Kesehatan's claim monitoring system can support early anomaly detection and enhance both financing efficiency and claim management integrity.*

1. Introduction

The Social Security Agency for Health (BPJS Kesehatan) is the implementing institution of Indonesia's national health insurance system, which aims to provide equitable access to healthcare services for all citizens [1, 2]. In Madiun City, the BPJS Kesehatan Branch Office reported that by the end of 2023, the total healthcare expenditure had reached approximately IDR 1.5 trillion. This budget was allocated for both primary and advanced healthcare services, with around 4 million cases recorded at primary care facilities and 2 million cases at secondary care levels under the National Health Insurance (JKN) scheme [3].

Along with the increasing utilization of healthcare services, the number of insurance claims has also risen significantly. This situation has led to various challenges, including the emergence of fraud indications in the claims system [4]. The President Director of BPJS Kesehatan stated that by the end of 2024, the institution was projected to face annual budget deficits, partly due to irregular claims and potential fraud [5, 6]. Furthermore, the Deputy Chairperson of the Corruption Eradication Commission (KPK) revealed that potential fraud in the healthcare sector could reach up to IDR 20 trillion annually [7]. Such fraudulent practices may include inflated medical costs, fictitious claims, and diagnosis manipulation aimed at increasing claim values [8]. If left unaddressed, this issue may threaten the long-term sustainability of the National Health Insurance (JKN) program.

Manual fraud detection has inherent limitations due to the vast volume of claim data. Therefore, data-driven approaches using data mining and machine learning techniques are needed to automatically and accurately identify irregular claim patterns. One relevant approach is clustering analysis, an unsupervised learning technique designed to group data based on shared characteristics [9]. In this study, two main clustering methods are employed: *K-Means* and *Hierarchical Clustering*. *K-Means* is known for its efficiency in handling large datasets, while *Hierarchical Clustering* provides a more informative representation of relationships between data points through its hierarchical structure [10].

Previous studies have demonstrated the effectiveness of clustering methods in various contexts. Research by [11–13] confirmed that *K-Means* often outperforms other clustering techniques, while studies by [14, 15] indicated that *Hierarchical Clustering* can yield better results based on the *Silhouette Score*. Meanwhile, studies on fraud detection within the JKN program [16–18] emphasized that weaknesses in internal audits, insufficient monitoring systems, and behavioral factors such as oppor-

*Corresponding Author.

Email : *mqs151@ummad.ac.id* (M. Q. Shobri)
Homepage : http://ejurnal.ung.ac.id/index.php/euler/index / E-ISSN : 2776-3706

Table 1. Research variables

No.	Variable	Description	Unit/Category
1	Age	Represents the age of the patient at the time of the claim	Years
2	Gender	Describes the gender of the patient receiving healthcare services	Male / Female
3	Occupation	Employment status of the patient when submitting the claim	Civil Servant, Private Employee, Student, Unemployed, etc.
4	Healthcare Service	Type of healthcare service received by the patient	Inpatient, Outpatient, Emergency Care
5	Disease Diagnosis	Type or code of the main diagnosis based on hospital data	Based on ICD-10 or disease description
6	Claim Frequency	Number of healthcare claims submitted by the patient during the one-month period	Number of claims
7	Total Claim Cost	Total cost of claims covered by BPJS Kesehatan	Rupiah (IDR)

tunity and pressure are among the primary causes of fraud. Several cases, including upcoding practices in BPJS Depok and findings in Buton, highlight the need to strengthen human resources, reporting systems, and integrated fraud prevention policies.

Based on the above background, this study focuses on the clustering of BPJS Kesehatan insurance claim data in Madiun City to detect high claim patterns and fraud indications. This approach is important since no previous study has specifically combined and compared *K-Means* and *Hierarchical Clustering* methods on BPJS Kesehatan claim data in Madiun City. The innovation of this research lies in its object and methodology, namely the integration of two clustering algorithms to enhance the accuracy of claim pattern detection. Therefore, the results of this study are expected to provide both academic contributions and practical recommendations for improving the efficiency and integrity of the BPJS Kesehatan claim management system.

2. Methods

2.1. Data and Variables

This study utilized secondary data consisting of 309 health insurance claim records obtained from several hospitals in Madiun City. The dataset represents a one-month claim recap of patiens claims during 2025, selected because it is considered representative of the current condition of BPJS Kesehatan claim pattern. The research variables analyzed in this study are presented in Table 1.

2.2. Research Procedure

2.2.1. Research Flow

The research began with the collection of BPJS Kesehatan claim data from hospitals in Madiun City. The data underwent preprocessing, including handling missing values, removing outliers, and standardizing variables. Subsequently, clustering analysis was performed using the K-Means and Hierarchical Clustering methods to identify grouping patterns based on claim frequency and cost characteristics. The clustering results were then interpreted to identify high-cost claim groups that may indicate potential fraud. The overall research flow is illustrated systematically in Figure 1.

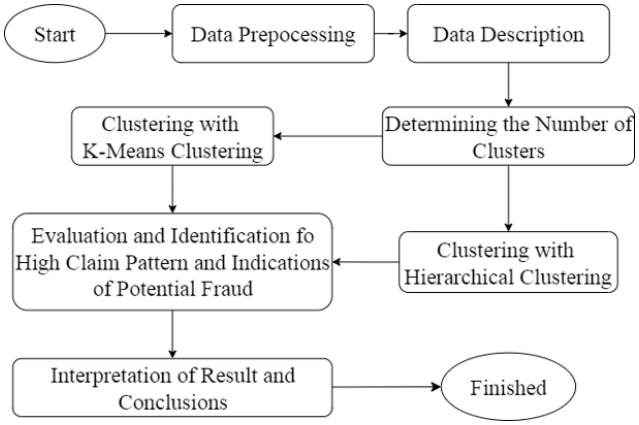


Figure 1. Rasearch flowchart

2.2.2. K-Means Clustering

The *K-Means Clustering* method was employed to identify high-claim patterns and potential fraud indications in *BPJS Kesehatan* insurance claim data from Madiun City. *K-Means* is a non-hierarchical clustering technique that groups data into several clusters based on the similarity of object characteristics. Each data point is assigned to the cluster with the nearest centroid, ensuring high similarity within clusters and clear separation between clusters.

The analysis began by determining the optimal number of clusters (k). Once the cluster number was set, centroids were initialized randomly. The distance between each data point and centroid was calculated using the Euclidean distance, as shown in eq. (1). The new cluster center was then computed using eq. (2) [15].

$$d(x_i, \mu_j) = \sqrt{\sum_{p=1}^{n} (x_{ip} - \mu_{jp})^2}, \tag{1}$$

$$\mu_j = \frac{1}{M_j} \sum_{i=1}^{M_j} x_i, \tag{2}$$

where $d(x_i, \mu_j)$ represents the Euclidean distance between the i-th data point and the j-th cluster centroid; x_i is the data vector; μ_j is the centroid of cluster j; n is the number of variables; p denotes the variable index; and M_j is the number of members in cluster j.

2.2.3. Hierarchical Clustering

The *Hierarchical Clustering* method was applied to form a hierarchy of data groupings based on the similarity levels among *BPJS Kesehatan* claim data in Madiun City. This method does not require a predetermined number of clusters and proceeds iteratively until all data points merge into a single hierarchical struc-

ture [19].

Initially, each claim record is treated as an individual cluster. Two clusters with the smallest distance are merged successively until a dendrogram is formed, representing the degree of similarity among claims. The distance between objects is computed using *Euclidean Distance* (eq. (3)), and cluster merging is performed using the *Average Linkage* method (eq. (4)). *Average Linkage* was selected because it produces stable and balanced cluster structures, making it effective for identifying high-claim patterns and detecting anomalous groups that may serve as early indicators of potential fraud.

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{p=1}^{n}(x_{ip} - x_{jp})^2}, \tag{3}$$

$$D(C_A, C_B) = \frac{1}{M_A M_B}\sum_{i=1}^{M_A}\sum_{j=1}^{M_B} d(x_i, x_j), \tag{4}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the Euclidean distance between two data points from different clusters; $\boldsymbol{x}_i$ is the i-th data vector in cluster C_A and $\boldsymbol{x}_j$ is the j-th data vector in cluster C_B; $D(C_A, C_B)$ is the average distance between clusters C_A and C_B; and M_A and M_B denote the number of members in clusters C_A and C_B, respectively [14].

2.2.4. Cluster Evaluation

Cluster evaluation was conducted to assess the quality and validity of the clustering results, ensuring that the objects within each cluster are highly similar (homogeneous) while significantly different from those in other clusters (heterogeneous). In this study, cluster validity was measured using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

a. Silhouette Score

The *Silhouette Score* $s(i)$ measures how well a data point i fits within its cluster, where $a(i)$ is the average distance to all other points in the same cluster, and $b(i)$ is the minimum average distance to points in the nearest neighboring cluster. The score is normalized using $\max(a(i), b(i))$, as shown in eq. (5).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \tag{5}$$

A value close to 1 indicates strong cohesion within the cluster, a value near 0 indicates that the point lies at the boundary between clusters, and a negative value indicates potential misclassification [11].

b. Davies–Bouldin Index

The *Davies–Bouldin Index (DBI)* evaluates overall cluster quality by comparing intra-cluster compactness and inter-cluster separation. A smaller DBI value indicates more compact and well-separated clusters, as defined in eq. (6) and eq. (7) [20].

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}(R_{ij}), \tag{6}$$

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}, \tag{7}$$

where S_i and S_j denote the within-cluster scatter of clusters i and j; M_{ij} is the distance between the centroids of clusters i and j; and k is the total number of clusters.

c. Calinski–Harabasz Index

The *Calinski–Harabasz Index (CHI)* measures clustering quality based on the ratio of between-cluster variance to within-cluster variance. A higher CHI value indicates compact intra-cluster cohesion and distinct inter-cluster separation, signifying better clustering performance, as expressed in eq. (8).

$$CHI = \frac{\mathrm{Tr}(B_k)/(k-1)}{\mathrm{Tr}(W_k)/(n-k)}, \tag{8}$$

where $\mathrm{Tr}(B_k)$ is the sum of squared distances between cluster centroids and the overall centroid (between-cluster variance), $\mathrm{Tr}(W_k)$ is the sum of squared distances within each cluster (within-cluster variance), k is the number of clusters, and n is the total number of observations [21].

3. Results and Discussion

3.1. Descriptive Statistics

Descriptive analysis was conducted to provide a general overview of the characteristics of BPJS Kesehatan claim data in Madiun City. This descriptive statistic includes both numerical and categorical variables used in the study. The numerical variables consist of age, number of claims per period (one month), and total claim costs, while the categorical variables include gender, type of occupation, and type of healthcare service received. This analysis aims to understand the data distribution patterns prior to the clustering process, which is used to detect potential high claims and indications of fraud.

Table 2. Descriptive statistics of numerical data

Variable	Mean	Std. Dev.	Min	Max
Age (Years)	47.8	20.8	0.5	88
Claim Frequency	2.26	1.5	1	7
Total Claim Cost (IDR)	3,018,367	4,614,705	179,100	23,365,400

Based on Table 2, the average age of claim participants is 47.8 years, with a relatively wide age range (standard deviation of 20.8), indicating that claimants come from diverse age groups, ranging from infants (0.5 years) to elderly participants (88 years). The average number of claims is 2.26 per month, with a maximum of seven claims, suggesting the presence of participants with high utilization intensity. Meanwhile, total claim costs vary substantially, with an average of IDR 3,018,367 and a maximum of IDR 23,365,400, indicating significant differences in healthcare service usage and medical expenses among participants.

As shown in Table 3, the majority of claimants are female (54.40%), indicating that women tend to utilize *BPJS Kesehatan* services more frequently than men. In terms of occupation, private employees represent the largest proportion (20.70%), followed by students (15.50%) and housewives (15.20%). This suggests that healthcare claims are relatively high among both productive-age and non-working groups. Regarding the type of healthcare service, inpatient care accounts for the highest proportion of claims

Table 3. Descriptive statistics of categorical data

Variable	Category	Percentage
Gender	Male	45.60%
	Female	54.40%
Occupation	Unemployed	5.80%
	Daily Laborer	3.90%
	Housewife	15.20%
	Private Employee	20.70%
	Trader	4.20%
	State-Owned Enterprise Employee	1.60%
	Student	15.50%
	Retiree	10.40%
	Farmer	9.10%
	Civil Servant	5.50%
	Entrepreneur	8.10%
Healthcare Service	Inpatient	43.70%
	Outpatient	24.90%
	Emergency (ER)	31.40%

(43.70%), followed by emergency care (31.40%) and outpatient care (24.90%), indicating that most claims originate from services with relatively higher cost intensity.

3.2. *Claim Data Clustering*

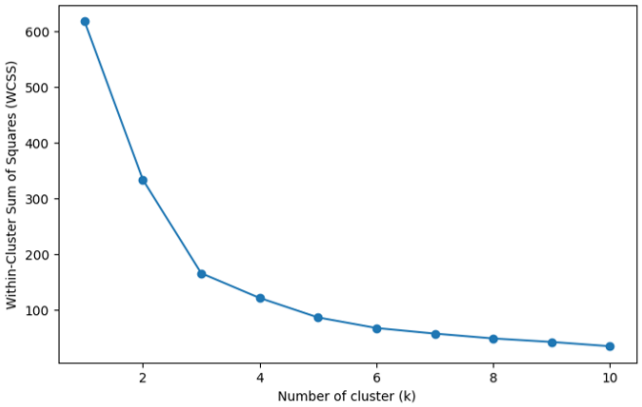
Clustering analysis of claim data was conducted using two main variables claim frequency and total claim costs as these are considered the most representative indicators of claim intensity and the financial burden borne by participants. Other variables, such as age, gender, occupation, healthcare service, and disease diagnosis, were included in the interpretation stage to enrich the characterization of each resulting group. The clustering process employed two methods: K-Means Clustering and Hierarchical Clustering.

3.2.1. K-Mean Clustering

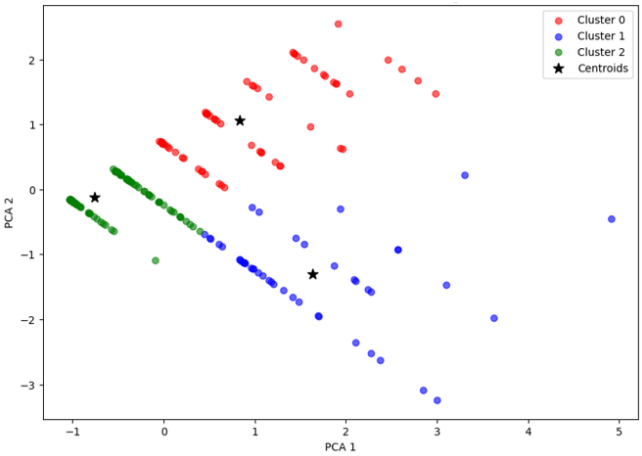
The optimal number of clusters in K-Means was determined using the Elbow Method, which is based on the change in the Within-Cluster Sum of Squares (WCSS). This method identifies the point of balance between the number of clusters and the degree of heterogeneity within groups. The results of the Elbow analysis are presented in Figure 2.

As shown in Figure 2a, the WCSS value sharply decreases as the number of clusters increases from $k = 1$ to $k = 3$, after which the decline becomes more gradual. The point where the reduction ceases to be significant, known as the elbow point, indicates the optimal number of clusters. Therefore, the best clustering configuration in this study consists of three clusters, categorized as low, medium, and high. This configuration provides a balance between model simplicity and grouping accuracy. Figure 2b illustrates the distribution pattern of the clusters based on the two main variables: claim frequency and total claim cost. Each point represents an individual claim record, and the colors indicate cluster membership. Cluster 0 (High) is represnted in red, Cluster 1 (Medium) is shown in blue, and Cluster 2 (Low) is displayed in green.

Cluster 0 (High): This cluster includes participants with moderate to high claim frequencies (2 to 7 claims) and relatively large total claim costs, exceeding IDR 20 million in several cases.



(a)



(b)

Figure 2. K-Mean Clustering:(a) Elbow method, (b) Visualization of K-Means clustering

Most members of this cluster are elderly participants (aged over 55 years) who are retirees, housewives, or private employees, with the majority utilizing inpatient services. This pattern represents a group with high claim frequency and substantial financial burden, warranting closer attention in risk control and potential fraud detection. Cluster 1 (Medium): This cluster comprises participants with low to moderate claim frequencies and moderately varied claim costs, generally below IDR 10 million. The group is dominated by younger participants, particularly students and private employees, who primarily use emergency and outpatient services. This cluster reflects non-chronic claim patterns with moderate costs, indicating normal utilization without significant anomalies.

Cluster 2 (Low): This cluster consists of participants with low claim frequencies (1 to 2 claims) and relatively small total claim costs (around IDR 200,000 to IDR 2 million). It is mainly composed of individuals in the productive age range (35 to 60 years), such as employees, entrepreneurs, and farmers, who predominantly utilize emergency and outpatient services. This cluster reflects routine and low-risk claim patterns, suggesting reasonable and efficient use of healthcare services.

Table 4. Clustering method evaluation results

Method	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Means Clustering	0.617*	0.694	419.581*
Hierarchical Clustering	0.499	0.509*	56.050

**The best performing method*

3.2.2. Hierarchical Clustering

Further clustering analysis was conducted using the Hierarchical Clustering method with the Average Linkage approach. This method was chosen because it can effectively represent hierarchical relationships between data points and visualize grouping structures through a dendrogram. The clustering result is shown in Figure 3.

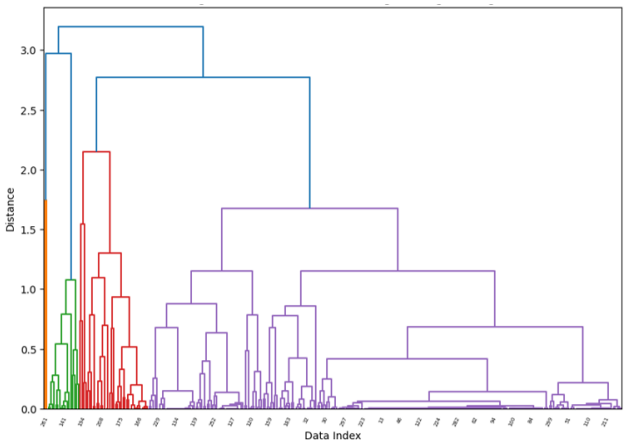


Figure 3. Dendogram of hierarchical clustering

Based on Figure 3, three main clusters were identified, representing variations in claim behavior among *BPJS Kesehatan* participants in Madiun City: Cluster 0 (High), Cluster 1 (Medium), and Cluster 2 (Low). These clusters demonstrate distinct differences in claim frequency, total claim costs, and demographic characteristics.

Cluster 0 (High): This cluster consists of two participants with high claim frequencies (6 to 7 claims) and very large total claim costs, ranging from IDR 13,056,400 to IDR 20,470,000. All members are female housewives (IRT) who primarily use inpatient services. Cluster 1 (Medium): This cluster includes participants with 6 to 7 claims and medium claim costs (IDR 3,000,000–8,000,000). The group is dominated by housewives, private employees, and students who mainly utilize inpatient and emergency services. The high frequency of claims within the same period suggests potential overutilization, warranting further verification of medical records. Cluster 2 (Low): This cluster contains most participants, with 1 to 5 claims and low to moderate total costs (generally below IDR 5,000,000). It includes informal workers, housewives, and students using various services, including inpatient, outpatient, and emergency care. The pattern reflects normal claim behavior consistent with general medical needs, without signs of irregularities.

3.2.3. Clustering Evaluation

Evaluation of clustering performance was conducted to assess the quality and validity of the formed data groups using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The results are presented in Table 4.

Based on the evaluation of both clustering methods in Table 4, K-Means demonstrated superior performance in producing well-defined and homogeneous data groups. The Silhouette Score of 0.617 for K-Means, compared to 0.499 for Hierarchical Clustering, indicates clearer and more stable cluster structures. Although the Davies-Bouldin Index of Hierarchical Clustering 0.509 was slightly lower implying slightly better separation between clusters the K-Means method achieved a significantly higher Calinski-Harabasz Index (419.581 versus 56.050), suggesting that K-Means generated more compact and distinctly separated clusters. Therefore, based on all three evaluation metrics, K-Means was deemed the most optimal and appropriate method for claim data grouping in this study.

3.3. Identification of High Claim Patterns and Potential Fraud

Based on the clustering analysis using both K-Means and Hierarchical Clustering methods, Cluster 0 (High) was identified as having the highest claim frequency and total claim costs. This cluster represents participants whose claim patterns significantly differ from the others and serves as the primary focus for identifying potential irregular or fraudulent claims. Participants in Cluster 0 have claim frequencies ranging from 2 to 7 claims, with total claim costs exceeding IDR 20 million. The group is dominated by elderly participants (aged over 55 years), particularly retirees, housewives, and private employees, most of whom use inpatient services. This pattern indicates a very high utilization intensity within a single financing period.

From a risk management perspective, such characteristics may initially suggest potential anomalies in claim behavior, such as overutilization or repeated claims for the same diagnosis. These patterns are commonly used as early indicators in health insurance fraud detection systems, as they may signal disproportionate financial burdens compared to other participant groups. However, upon further examination of medical records and case histories, the high claim costs were found to result from complex medical conditions and intensive treatment requirements consistent with clinical indications. Therefore, all claims in this cluster were deemed valid and procedurally appropriate, showing no signs of administrative or medical fraud.

This finding emphasizes that statistical cluster based analysis must be complemented by clinical and administrative verification to ensure accurate interpretation. Although Cluster 0 exhibits the highest claim values, comprehensive evaluation confirms that these claims reflect genuine medical needs rather than irregular claim behavior. Practically, this group should remain under regular monitoring by BPJS Kesehatan, as its high-risk characteristics could indicate potential cost inefficiencies if left unchecked. The implementation of data-driven claim monitoring systems, supported by medical validation, is essential to

maintain financing efficiency, claim accuracy, and the integrity of the national health insurance system.

4. Conclusion

This study demonstrates that the K-Means and Hierarchical Clustering methods are effective in identifying claim patterns within the BPJS Health system in Madiun City. The analysis produced three main clusters: low, medium, and high. The High Cluster consists of participants with a claim frequencies of 2 to 7 claims and total claim costs exceeding IDR 20 million. This cluster is dominated by older participants (over 55 years old), including retirees, housewives, and private-sector employees, with most claims originating from inpatient services. Performance evaluation indicates that K-Means is the superior method, achieving a Silhouette Score of 0.617 and a Calinski–Harabasz Index of 419.581. Although the High Cluster shows high-risk claim patterns in terms of financial burden, medical verification confirmed that all claims were appropriate and did not indicate any fraudulent activity. Clustering methods can only identify patterns and anomalies in service utilization; therefore, the results should be interpreted as early indicators rather than direct evidence of fraud. Fraud confirmation still requires complementary clinical assessments and administrative audits. These findings indicate that K-Means produces more compact and well-separated clusters compared to Hierarchical Clustering. As a recommendation, BPJS Health should integrate clustering analysis into its claim monitoring system to enhance early anomaly detection and improve audit efficiency. In addition, training for claim verification officers is essential to ensure that data analysis results can be utilized optimally.

Author Contributions. **Muhammad Qolbi Shobri**: Conceptualization, methodology, formal analysis, validation, literature review writing, final editing, project administration, and funding acquisition. **Putri Balqis Al-Kubro**: Data collection and curation, investigation, initial draft preparation, visualization, and provided support in analysis implementation and technical supervision. **Gabriella Vindy Kawuri**: Software development, implementation and validation of clustering models, visualization, initial draft preparation, and provided technical support during data analysis. All authors discussed the results and contributed to the final manuscript.

Acknowledgement. The authors would like to express their sincere gratitude to all parties who contributed to this research and the preparation of the manuscript. We deeply appreciate the editors and reviewers for their valuable feedback and support in improving the quality of this work.

Funding. This research was funded by the Ministry of Higher Education, Science, and Technology through the Directorate of Research and Community Service (DPPM 2025).

Conflict of interest. The authors declare no conflicts of interest related to this article.

Data availability. Not available.

References

[1] OJK, *Statistik Jaminan Sosial Indonesia 2022*. Jakarta: Direktorat Statistik dan Informasi IKNB, 2023.

[2] Indonesia AIDS Coalition, *Buku Panduan Jaminan Kesehatan Nasional (JKN) bagi Populasi Kunci*, 2016. [Online]. Available: https://siha.kemkes.go.id/portal/files. [Accessed: 2025].

[3] Jatimpos, "Sepanjang Tahun 2023 BPJS Kesehatan Kantor Cabang Madiun Keluarkan Biaya Kesehatan Rp1,5 Triliun," *Jatimpos.co*, 2024. [Online]. Available: https://www.tempo.co/ekonomi. [Accessed: 2025].

[4] M. Q. Shobri, R. A. Andyani, and M. Jeksen, "Regresi Logistik Bayesian dan Algoritma C4.5 dalam Klasifikasi Risiko Penggunaan BPJS Kesehatan Kota Madiun," *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 3, pp. 2064–2078, 2024. doi: 10.46306/lb.v5i3.814.

[5] Tempo, "Dirut Sebut BPJS Kesehatan Alami Defisit Sekitar Rp20 Triliun Tahun Ini," *Tempo.co*, Nov. 11, 2024. [Online]. Available: https://www.tempo.co/ekonomi. [Accessed: 2025].

[6] V. Singgih, "BPJS Kesehatan Terancam Tekor Rp20 Triliun dan Gagal Bayar Klaim, Kenaikan Iuran Jadi 'Keniscayaan'," *BBC News Indonesia*, Nov. 15, 2024. [Online]. Available: https://www.bbc.com/indonesia. [Accessed: 2025].

[7] CNN Indonesia, "KPK: Kerugian dari Fraud di Bidang Kesehatan Sekitar Rp20 Triliun," *CNN Indonesia*, Sep. 20, 2024. [Online]. Available: https://www.cnnindonesia.com/nasional/20240920023933. [Accessed: 2025].

[8] S. Tito, J. Julius, and K. N. Siregar, "Faktor Pemicu dan Penghambat Fraud dalam Program Jaminan Kesehatan Nasional dan Strategi Pencegahannya: Sebuah Scoping Review," *Jurnal Ekonomi Kesehatan Indonesia*, vol. 9, no. 2, Art. 5, 2024, doi: 10.7454/eki.v9i2.1124.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. USA: Morgan Kaufmann, 2012.

[10] J. W. G. Putra, "Pengenalan Konsep Pembelajaran Mesin dan Deep Learning," Aug. 17, 2020. [Online]. Available: https://wiragotama.github.io/. [Accessed: 2025].

[11] K. S. Pranata, A. A. S. Gunawan, and F. L. Gaol, "Development clustering system IDX company with k-means algorithm and DBSCAN based on fundamental indicator and ESG," *Procedia Computer Science*, vol. 216, pp. 319–327, 2023, doi: 10.1016/j.procs.2022.12.142.

[12] E. M. S. Rochman, Miswanto, and H. Suprajitno, "Comparison of clustering in tuberculosis using fuzzy c-means and k-means methods," *Communications in Mathematical Biology and Neuroscience*, 2022, Article ID 41, doi: 10.28919/cmbn/7335.

[13] M. S. Hasibuan, A. H. Lubis, and M. N. Sari, "Perbandingan Algoritma Clustering DBSCAN dan K-Means dalam Pengelompokan Siswa Terbaik," *INFOTECH: Jurnal Informatika & Teknologi*, vol. 5, no. 2, pp. 301–309, 2024. doi: 10.37373/infotech.v5i2.1457.

[14] I. Surairoh, A. C. Rani, K. Amalia, and D. Rolliawati, "Perbandingan Hasil Analisis Clustering Metode K-Means, DBSCAN dan Hierarchical pada Data Marketplace Electronic Phone," *Joins: Journal Information System*, vol. 8, no. 1, pp. 95–105, 2023. doi: 10.33633/joins.v8i1.8016.

[15] F. D. Wahyuningtyas, A. Arafat, A. Stiawan, and D. Rolliawati, "Komparasi Algoritma Hierarchical, K-Means, dan DBSCAN pada Analisis Data Penjualan Melalui Facebook," *EXPLORE: Jurnal Sistem Informasi dan Telematika*, vol. 14, no. 1, pp. 7–16, 2023. doi: 10.36448/jsit.v14i1.2931.

[16] A. Y. B. R. Thaifur, "Exploratory Study of Factors Influencing Fraud in the National Health Service in Buton Islands from a Hexagon Model Perspective," *Healthcare in Low-resource Setting*, vol. 13, no. 1, pp. 32–35, 2025. doi: 10.4081/hls.2024.12773.

[17] N. Sariunita and R. A. Syakurah, "Analisis Kejadian Upcoding Biaya Pelayanan Kesehatan di Wilayah Kerja BPJS Kesehatan Cabang Depok," *BIGES JUKES*, vol. 14, no. 2, pp. 1–6, 2023. doi: 10.35907/bgjk.v14i2.220.

[18] K. N. Aprilia and R. Nurhayati, "Analisis Kompetensi Auditor Internal terhadap Kemampuan Mencegah dan Mendeteksi Fraud dalam Program Jaminan Kesehatan Nasional (Studi Kasus di Rumah Sakit Bethesda Yogyakarta)," *ABIS: Accounting and Business Information Systems Journal*, vol. 9, no. 2, pp. 227–246, 2021. doi: 10.22146/abis.v9i2.65895.

[19] R. A. Andyani, M. Q. Shobri, M. Baihaqi, P. B. Al-Kubro, and M. S. Adhantoro, "Aplikasi Metode Ward dengan Berbagai Pengukuran Jarak (Studi Kasus: Klasifikasi Tingkat Perekonomian)," *JIKM: Jurnal Ilmiah Kampus Mengajar*, vol. 4, no. 2, pp. 177–190, 2024. doi: 10.56972/jikm.v4i2.208.

[20] D. A. T. Devanta, "Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 4, no. 1, pp. 545–552, 2023. doi: 10.30865/klik.v4i1.964.

[21] L. A. Sari, A. R. Hakim, and A. Rusgiyono, "Penggunaan Index Calinski–Harabasz pada Clustering K-Medoids Algorithm untuk Penggolongan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Karakteristik Penduduk Miskin," *Jurnal Gaussian*, vol. 14, no. 1, pp. 179–187, 2025. doi: 10.14710/j.gauss.14.1.179-187.