Determining Dengue Hemorrhagic Fever Severity Levels Using K-Means Clustering: A Data-Driven Approach for Clinical Risk Stratification

Veronica Agesta Adyastuti, Warsi Maryati*, Yunita Wisda Tumarta Arif

Department of Medical Record and Health Information, Faculty of Health Sciences, Universitas Duta Bangsa Surakarta, Indonesia
*Email: warsi_maryati@udb.ac.id

ABSTRACT

Background: Dengue Hemorrhagic Fever (DHF) remains a persistent public health burden in Indonesia, with 12,994 cases reported in Central Java Province in 2023. At a regional general hospital in Surakarta, DHF inpatient cases surged from 48 in 2023 to 610 in 2025, underscoring the urgent need for systematic severity classification. **Objective:** this study aimed to determine DHF severity levels using K-Means Clustering as a data-driven approach for clinical risk stratification. **Methods:** A quantitative descriptive design with a retrospective approach was applied to all 610 DHF inpatient cases, yielding 592 valid records after data cleaning. Five clinical variables were analyzed: age, platelet count, hematocrit, hemoglobin, and leukocyte count. **Results:** Data processing followed the CRISP-DM framework in RapidMiner Studio with K=3. The Davies-Bouldin Index (DBI) value of 0.745 confirmed good clustering quality. Three clusters were identified: Cluster 0 (Dengue Without Warning Signs, n=251, 42.40%), Cluster 1 (Dengue With Warning Signs, n=72, 12.16%), and Cluster
2 (Severe Dengue, n=269, 45.44%). **Conclusion:** These findings demonstrate that K-Means Clustering effectively supports clinical risk stratification of DHF severity levels, offering a reliable data-driven strategy for timely clinical decision-making in tropical endemic settings..

Keywords *Clustering; Dengue Hemorrhagic Fever; K-Means; RapidMiner; Severity Level*

INTRODUCTION

Dengue Hemorrhagic Fever (DHF) is an acute infectious disease caused by the dengue virus of the family *Flaviviridae*, transmitted through the bite of the *Aedes aegypti* mosquito. The disease is clinically characterized by sudden high fever, hemorrhagic manifestations, thrombocytopenia, and hemoconcentration, and may be fatal if not managed promptly [1]. The urgency of DHF control is underscored by its potential to trigger extraordinary outbreak events with high mortality rates, particularly in densely populated urban areas with tropical climate characteristics such as Indonesia [2].

According to the Indonesian Ministry of Health, a total of 114,720 DHF cases were recorded nationally in 2023, resulting in 894 deaths [3]. In Central Java Province, 12,994 DHF cases were reported in the same year, demonstrating significant regional disease burden [4]. At General Hospital of Surakarta, DHF inpatient cases increased dramatically from 48 cases in 2023, to 560 in 2024, and further to 610 cases in 2025, placing DHF as the second-highest diagnosis in the hospital's top-10 disease list. This trend highlights the urgent need for systematic patient stratification tools.

The Indonesian Ministry of Health's national clinical guideline (Keputusan Menteri Kesehatan RI No. HK.01.07/MENKES/4636/2021) defines three DHF severity levels [5]: (1) *Dengue Without Warning Signs* - patients maintain adequate oral intake and urinary output with stable hematocrit; (2) *Dengue With Warning Signs* - characterized by plasma leakage, elevated hematocrit, and increasing capillary permeability, typically occurring around days 3–7 of illness; and (3) *Severe*

Dengue — marked by severe plasma leakage leading to dengue shock, massive hemorrhage, and critical organ damage. Early severity identification is paramount to facilitate timely interventions and prevent clinical deterioration [5].

Clinically relevant parameters for DHF severity include patient age (children aged 0–14 years show higher hospitalization risk), body temperature (defervescence at 37.5–38°C signals the critical phase), platelet count (< 100,000 cells/mm³ indicates thrombocytopenia), hematocrit elevation reflecting hemoconcentration, and progressive leukopenia indicating the immune response to dengue infection [5]. Systematic processing of these parameters through data mining offers an objective and reproducible classification approach.

Prior studies on DHF clustering include Khusaeri (2024) [7], who applied K-Means to cluster DHF geographic distributions, yielding two clusters (SSE=385.10); Mohsa *et al.* (2023) [8], who identified three optimal spatial clusters of DHF in West Java; and Adiputra (2021) [9], who clustered DHF patients by age group. However, none of these studies applied clustering to determine the clinical *severity level* of individual patients using laboratory parameters, which constitutes the primary novelty of the present study. A key strength of this study lies in its use of variables that directly reflect clinical outcomes - namely platelet count, hematocrit, hemoglobin, and leukocyte count - which are the core laboratory indicators stipulated in the national DHF management guideline for severity assessment. Prior studies in this domain did not specifically orient their variable selection toward clinical outcome-based parameters, nor did they address the determination of individual patient severity level. The integration of clinical outcome variables with severity level determination in a single clustering framework represents a meaningful methodological advancement that enhances both the clinical applicability and interpretability of the results.

The objective of this study was to apply K-Means Clustering with the CRISP-DM framework to classify DHF patients at General Hospital of Surakarta in 2025 into three WHO-defined severity levels based on clinical and laboratory data extracted from the hospital's electronic medical records.

METHODS

Data processing in this study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which provides a structured approach to data mining and analysis [6]. The first stage, Business Understanding, defined the research problem as the high number of DHF cases at General Hospital of Surakarta in 2025 (610 cases) and the need for a systematic, objective method to determine patient severity levels through clustering. The second stage, Data Understanding, involved exploring secondary data from the electronic medical records (EMR) of DHF patients recorded in the hospital's SIMRS during 2025, identifying the relevant clinical variables (age, body temperature, platelet count, hematocrit, and leukocyte count), and assessing data completeness and characteristics. The third stage, Data Preparation, encompassed editing (checking data completeness and consistency, with incomplete or inconsistent records excluded), coding (converting SIMRS data into .csv/.xlsx format compatible with RapidMiner), cleaning (removing duplicate records, missing values, input-error outliers, and standardizing measurement units), data transformation (adjusting the data format to meet the requirements of the K-Means algorithm), and tabulating the final dataset of the five clinical variables for analysis.

The fourth stage, Modeling, involved selecting K-Means Clustering as the analytical method, setting the number of clusters at K=3 based on the three DHF severity categories defined in the Indonesian Ministry of Health Decree No. HK.01.07/MENKES/4636/2021 (Dengue Without Warning Signs, Dengue With Warning Signs, and Severe Dengue), importing the prepared data into RapidMiner Studio, configuring the clustering parameters, executing the clustering process, and recording the resulting centroids and cluster memberships. The fifth stage, Evaluation, assessed clustering quality using the Davies-Bouldin Index (DBI), where lower values indicate better intra-cluster cohesion and inter-cluster separation. The final stage, Deployment, involved interpreting the

clustering results to determine the severity level of each cluster based on its centroid characteristics, mapping Cluster 0 to Dengue Without Warning Signs, Cluster 1 to Dengue With Warning Signs, and Cluster 2 to Severe Dengue, and presenting the findings through tables, visualizations, and a written research report.

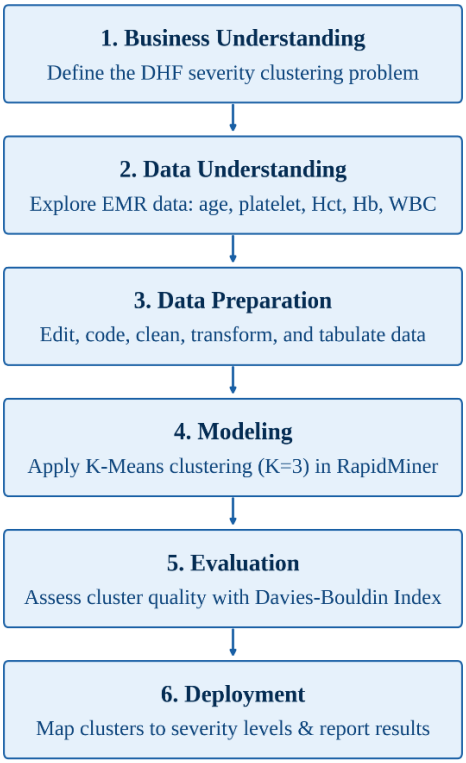

Figure 1. Research Stages Flowchart

RESULTS

1. Distribution of DHF Cases in 2025

A total of 610 DHF inpatient cases were recorded at General Hospital of Surakarta in 2025. The monthly distribution is presented in Table 1.

The peak incidence occurred in May 2025 with 97 cases (15.90%), followed by April (81 cases, 13.28%) and July (78 cases, 12.79%). The second quarter (April–June) accounted for 236 cases (38.69%) of the annual total, consistent with the rainy season in Surakarta and with established DHF seasonal patterns in tropical regions. The lowest monthly incidence was in November (18 cases, 2.95%).

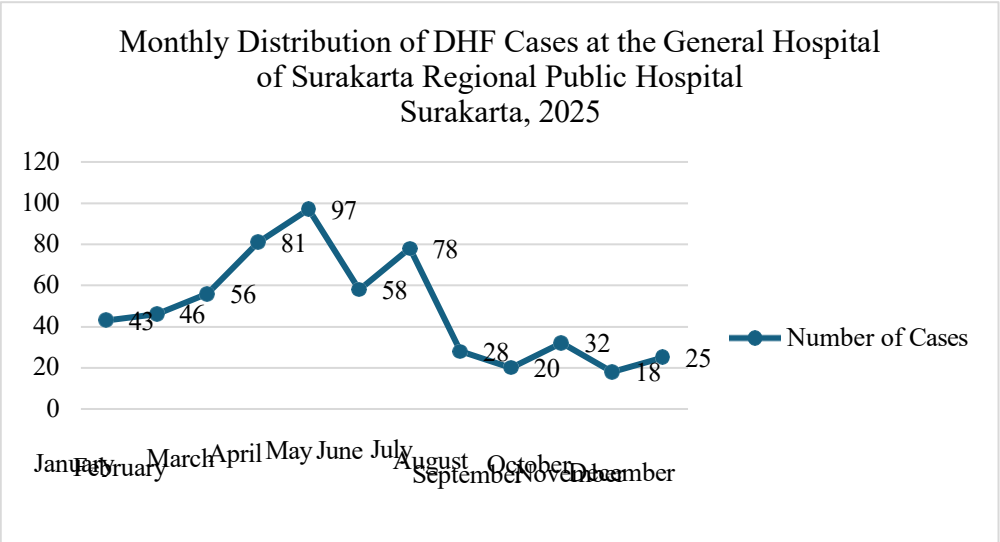

Figure 2. Monthly Distribution of DHF Cases at the General Hospital of Surakarta Regional Public Hospital Surakarta, 2025

2. Data Preparation

Of the 610 initial records, 18 were excluded due to missing values, yielding 592 valid records (97.05%) for analysis. Descriptive statistics of the five variables are presented in Tables 2 and 3.

Table 1. Descriptive Statistics of Clinical Variables (n = 592)

Variable	Min	Max	Mean	Std. Deviation
Age (years)	0	80	19.08	14.68
Platelet (x10³/µL)	3	331	80.04	48.96
Hematocrit (%)	25	59	39.45	5.21
Hemoglobin (g/dL)	8.7	20.4	13.43	1.88
Leukocyte (x10³/µL)	0.97	16.46	4.83	2.59

Based on table 1, the mean patient age was 19.08 ± 14.68 years, indicating a predominantly young patient population. Mean platelet count of 80.04 ± 48.96 ×10³/µL confirms thrombocytopenia characteristic of DHF. Mean leukocyte count of 4.83 ± 2.59 ×10³/µL reflects leukopenia associated with dengue infection.

3. K-Means Clustering Results

Table 2. Cluster Membership Distribution

Cluster	Number of Members	Percentage (%)
Cluster 0	251	42.40
Cluster 1	72	12.16
Cluster 2	269	45.44
Total	592	100.00

K-Means Clustering with K=3 was applied to the 592 valid records using RapidMiner Studio. Table 4 presents the cluster membership distribution and Table 5 the centroid values for each cluster.

The cluster membership distribution in Table 4 shows that Cluster 2 contained the largest proportion of patients (269 cases, 45.44%), followed closely by Cluster 0 (251 cases, 42.40%), while Cluster 1 was the smallest group (72 cases, 12.16%). To characterize the clinical profile underlying each of these groups, Table 5 presents the centroid values of the five clinical variables for each cluster, which form the basis for interpreting the severity level represented by each group.

Table 3. Cluster Centroids for Clinical Variables

Variable	Cluster 0	Cluster 1	Cluster 2
Age (years)	22.1	41.5	11.8
Platelet (x10³/µL)	47.3	80.6	108.2
Hematocrit (%)	44.5	35.1	36.4
Hemoglobin (g/dL)	15.1	12.1	12.5
Leukocyte (x10³/µL)	4.6	8.3	4.0

Based on Table 5, the centroid values reveal three distinct clinical profiles. Overall, the three centroid profiles, differentiated primarily by age, platelet count, and hemoglobin/hematocrit levels, provide a clinically interpretable basis for assigning each cluster to one of the three DHF severity categories defined in national guidelines.

Table 4. Mapping of Clusters to DHF Severity Levels and Clinical Management

Cluster	Severity Level	n (%)	Centroid Characteristics	Clinical Management
Cluster 0	Dengue Without Warning Signs	251 (42.40%)	Low platelet (47.3 x10³/µL), elevated hematocrit (44.5%), mean age 22.1 years	Outpatient, close monitoring, danger sign education
Cluster 1	Dengue With Warning Signs	72 (12.16%)	Moderate platelet (80.6 x10³/µL), lower hematocrit (35.1%), elevated leukocyte (8.3 x10³/µL), older age (41.5 years)	Hospitalization, IV isotonic crystalloid 5–7 mL/kg/h
Cluster 2	Severe Dengue	269 (45.44%)	Higher platelet (108.2 x10³/µL), young patients (11.8 years), leukopenia (4.0 x10³/µL)	Emergency fluid resuscitation 10–20 mL/kg/h, ICU care

4. Cluster Quality Evaluation

The Davies-Bouldin Index (DBI) for K=3 was 0.745. As DBI values closer to zero indicate better clustering quality, with greater intra-cluster cohesion and inter-cluster separation, a value of 0.745 demonstrates that the three-cluster model produces a well-defined, clinically meaningful partition.

5. Severity Level Classification

Each cluster was mapped to a DHF severity level by comparing centroid characteristics against the clinical criteria of the national guideline [5]. The mapping, with corresponding clinical management, is summarized in Table 6.

DISCUSSION

This study successfully applied K-Means Clustering with the CRISP-DM framework to classify 592 DHF patients at a general hospital in Surakarta into three clinically meaningful severity levels. The following discussion addresses the key findings from each of the five specific objectives of this research.

The seasonal distribution of cases, peaking in April–June (38.69%), aligns with DHF epidemiological patterns in tropical Indonesia [4], where the rainy season promotes *Aedes aegypti* breeding [2]. This finding supports targeted public health interventions during these months, including enhanced vector control and heightened hospital preparedness.

Cluster 0 (n=251, 42.40%) was classified as *Dengue Without Warning Signs*. The centroid platelet value of 47.3 ×10³/µL is low, but the presence of elevated hemoglobin (15.1 g/dL) and elevated hematocrit (44.5%) suggests hemoconcentration consistent with early dengue without

progression to plasma leakage. These patients may be managed as outpatients with monitoring and caregiver education regarding warning signs, per national guidelines [5,11].

Cluster 1 (n=72, 12.16%) was classified as *Dengue With Warning Signs*. The combination of moderate platelet count (80.6 $\times 10^3/\mu L$), lower hematocrit (35.1%), elevated leukocyte count (8.3 $\times 10^3/\mu L$), and older mean age (41.5 years) distinguishes this cluster. The leukocytosis pattern may reflect a secondary dengue infection or co-morbidity. Although this cluster is the smallest (12.16%), these patients require immediate hospitalization and intravenous fluid therapy to prevent progression to severe dengue [5,17].

Cluster 2 (n=269, 45.44%) was classified as *Severe Dengue*. This cluster was predominantly composed of young patients (mean age 11.8 years), consistent with evidence indicating higher DHF severity risk in the pediatric age group [5]. The leukopenia (4.0 $\times 10^3/\mu L$) and clinical context support severe dengue classification [5,29]. These patients require emergency fluid resuscitation and intensive care monitoring [5].

First, regarding DHF case distribution (Objective 1): the monthly case data revealed a pronounced peak in May 2025 (97 cases, 15.90%), with the second quarter (April–June) accounting for 38.69% of annual cases. This seasonal pattern aligns with the rainy season in Surakarta and confirms the strong relationship between tropical climate cycles and DHF transmission dynamics, consistent with epidemiological evidence for Indonesia [2,4]. This finding has direct implications for hospital resource planning and preventive public health campaigns, as similarly observed in prior DHF distribution studies [7,8].

Second, regarding data preparation (Objective 2): of the 610 initial records, 592 (97.05%) were retained after excluding 18 records with complete missing values. The high data completeness rate reflects the reliability of the electronic medical records system and supports the validity of subsequent analysis [12]. Descriptive statistics confirmed clinically significant findings: the mean platelet count of 80.04 $\pm$48.96 $\times 10^3/\mu L$ confirms population-level thrombocytopenia, the mean leukocyte count of 4.83 $\pm$2.59 $\times 10^3/\mu L$ indicates leukopenia consistent with dengue [5], and the mean patient age of 19.08 years reflects the vulnerability of the young population to DHF infection [11].

Third, regarding K-Means Clustering results (Objective 3): the algorithm produced three distinct clusters with markedly different centroid profiles. Cluster 0 exhibited the lowest platelet values (47.3 $\times 10^3/\mu L$) combined with elevated hematocrit (44.5%) and hemoglobin (15.1 g/dL), signaling hemoconcentration. Cluster 1 showed elevated leukocyte count (8.3 $\times 10^3/\mu L$) and the oldest mean age (41.5 years), suggesting either a secondary dengue infection or concurrent comorbidity. Cluster 2 was characterized by the youngest patients (mean age 11.8 years) and leukopenia (4.0 $\times 10^3/\mu L$), consistent with the pediatric high-risk profile for severe dengue. The K-Means algorithm proved well-suited to this classification task, in line with findings from prior DHF data mining studies [7,8,9]. The distinct age profile of each cluster is consistent with national guidelines on DHF clinical parameters [5], and the application of RapidMiner Studio supported efficient and reproducible clustering [6,28].

Fourth, regarding cluster quality evaluation (Objective 4): the Davies-Bouldin Index of 0.745 confirms good clustering quality, as values approaching zero indicate superior intra-cluster cohesion and inter-cluster separation. This result compares favorably with related validation studies [10] and demonstrates that the three-cluster solution produces a well-defined, clinically meaningful partition of DHF patients. The DBI value supports the a priori choice of K=3 as an appropriate and valid number of clusters for this dataset.

Fifth, regarding severity level determination and clinical recommendations (Objective 5): the mapping of clusters to WHO-defined severity levels yielded clinically coherent and actionable outcomes. Cluster 0 (42.40%, Dengue Without Warning Signs) can be managed in an outpatient setting with monitoring and danger sign education. Cluster 1 (12.16%, Dengue With Warning Signs) requires prompt hospitalization and IV fluid therapy. Cluster 2 (45.44%, Severe Dengue) demands

emergency fluid resuscitation and ICU-level care. The high proportion of Severe Dengue cases (45.44%) in this dataset underscores the critical importance of early severity identification to guide resource allocation and prevent mortality [5,11]. Vikhe et al. (2024) demonstrated that platelet count is among the most accurate predictors of dengue severity and can be used for patient stratification, supporting the classification observed in Cluster 0 and Cluster 2 [30]. Furthermore, Marandi et al. (2023) confirmed that hematocrit elevation is a key indicator of plasma leakage and severe dengue, consistent with the elevated hematocrit and hemoglobin values found in Cluster 2 [29]. Laksono et al. (2024) and Herdiaman et al. (2024) similarly demonstrated that K-Means Clustering applied to inpatient medical record data in Indonesian hospital settings produces clinically meaningful groupings that support decision-making [27,28]. Operationally, integrating this clustering model into the hospital information system could enable automatic patient stratification at admission, improving triage efficiency particularly during the peak transmission season.

CONCLUSION

The application of K-Means Clustering with the CRISP-DM framework successfully classified DHF patients at General Hospital of Surakarta into three severity levels: Cluster 0 (*Dengue Without Warning Signs*), which formed the largest group, Cluster 1 (*Dengue With Warning Signs*), which formed the smallest group, and Cluster 2 (*Severe Dengue*), which formed the second-largest group. The Davies-Bouldin Index confirmed good clustering quality. The study was limited by reliance on secondary EMR data, which precluded verification of clinical outcomes for each severity group. Future research is recommended to prospectively validate the model, incorporate additional variables (e.g., NS1 antigen results), and integrate the algorithm into the hospital information system to support real-time clinical decision-making in DHF management.

Acknowledgments: The authors express their gratitude to the Director and staff of the General Hospital of Surakarta for granting data access, and to the Faculty of Health Sciences, Universitas Duta Bangsa Surakarta, for institutional support.
Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Frida N. Mengenal Demam Berdarah Dengue. Semarang: Alprin; 2019.
2. Wijayanti SPM. Karakteristik dan Pola Penyebaran Penyakit Demam Berdarah Dengue di Wilayah Endemis. Purwokerto: UNSOED Press; 2019.
3. Kementerian Kesehatan RI. Profil Kesehatan Indonesia Tahun 2023. Jakarta: Kemenkes RI; 2024.
4. Dinas Kesehatan Jawa Tengah. Profil Kesehatan Jawa Tengah Tahun 2023. Semarang: Dinkes Jateng; 2024.
5. Kementerian Kesehatan RI. Keputusan Menteri Kesehatan RI Nomor HK.01.07/MENKES/4636/2021: Pedoman Nasional Pelayanan Kedokteran Tata Laksana Infeksi Dengue Anak dan Remaja. Jakarta: Kemenkes RI; 2021.
6. Sulianta F. Basic Data Mining From A to Z. Bandung: Universitas Widyatama; 2023.
7. Khusaeri A. Analisis Algoritma K-Means Clustering dalam Pengelompokan Daerah Penyebaran Penyakit Demam Berdarah Dengue. J Inf Syst Inform Comput. 2024;8(2):434–449.
8. Mohsa AA, Rosiana PS, Umaidah Y. Implementasi K-Means dalam Pengelompokan Penyebaran Penyakit DBD di Jawa Barat. JITET. 2023;11(3):782–788.
9. Adiputra NM. Clustering Penyakit DHF pada Rumah Sakit Dharma Kerti Menggunakan Algoritma K-Means. INSERT J. 2021;3(1):99–105.
10. Kapita S, Mubarak A, Abdullah SD, Fhadli M. Penerapan Algoritma Clustering Kohonen-SOM dengan Validasi Davies Bouldin Index. Indonesian J Inf Syst. 2022;7(2):134–143.

11. Tuwa MD, Talakua AC. Mendiagnosa Penilaian Keparahan Gigitan Demam Berdarah Dengue Melalui Gejala Menggunakan Sistem Pakar. 2025;5(1):21–28.
12. Notoatmodjo S. Metodologi Penelitian Kesehatan. Jakarta: PT Rineka Cipta; 2018.
13. Sugiyono. Metode Penelitian Kuantitatif Kualitatif dan R&D. Bandung: Alfabeta; 2025.
14. Dewi DAIC, Pramita DAK. Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids. Jurnal Matrix. 2019;9(3):102–109.
15. Iskandar MS, Fatah Z. Implementasi Metode Algoritma K-Means Clustering untuk Menentukan Penerima PIP. Gudang J Multidisiplin Ilmu. 2024;2(10):1–8.
16. Hosizah, Maryati W. Statistik Rumah Sakit. Jakarta: Kemenkes RI; 2018.
17. Safra CE, Hariyanti I. Penerapan Algoritma K-Means Clustering dalam Identifikasi Pola Penggunaan Barang. J Ilmiah Teknik. 2025;4(2):1–18.
18. Suharmanto. Metodologi Penelitian Kesehatan. Yogyakarta: Pustaka Panesa; 2021.
19. Sudra RI. Rekam Medis. Tangerang Selatan: Universitas Terbuka; 2020.
20. Adiputra NM. Penerapan Data Mining pada Rekam Medis untuk Analisis Penyakit DHF. INSERT J. 2023;5(1):44–52.
21. Peraturan Menteri Kesehatan No. 24 Tahun 2022 tentang Rekam Medis. Jakarta: Kemenkes RI; 2022.
22. Iskandar MS, Fatah Z. Implementasi Data Mining pada Analisis Kluster Kesehatan. J Inf Syst Comput Appl. 2024;12(1):55–65.
23. Dewi DAIC, Pramita DAK. Evaluasi Algoritma Clustering K-Means pada Data Pasien Rawat Inap. J Biostatistik. 2020;17(2):88–96.
24. Mohsa AA, Rosiana PS. Validasi K-Means Clustering pada Data Epidemiologi Berbasis Kewilayahan. J Epidemiol Indones. 2023;8(4):210–219.
25. Adiputra NM, Suryasa IW. Machine Learning Applications in Hospital Medical Records Analysis: A Systematic Review. Bali Med J. 2023;12(2):900–907.
26. Fathurrahman, Harini S, Kusumawati R. Evaluasi Clustering K-Means dan K-Medoid pada Persebaran Covid-19 di Indonesia dengan Metode Davies-Bouldin Index (DBI). Jurnal MNEMONIC. 2023;6(2):117–128.
27. Herdiaman EA, Sudiarjo A, Hikmatyar M. Klasterisasi Pasien Pada RSUD Ciamis Menggunakan Metode K-Means. Jurnal Informatika dan Teknik Elektro Terapan (JITET). 2024;12(3S1):3538–3547.
28. Laksono WB, Syahidin Y, Yunengsih Y. Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering. Jurnal Teknologi Sistem Informasi dan Aplikasi. 2024;7(2):621–627.
29. Marandi RZ, Leung P, Sigera C, Murray DD, Weeratunga P, Fernando D, Rodrigo C, Rajapakse S, MacPherson CR. Development of a Machine Learning Model for Early Prediction of Plasma Leakage in Suspected Dengue Patients. PLOS Negl Trop Dis. 2023;17(3):e0010758.
30. Vikhe VB, Khandol D, Faruqi AA, Reddy A. A Study on the Clinical Profile of Patients Presenting With Dengue Fever and the Precision Prediction of Dengue Severity Using Platelet Count at Presentation. Cureus. 2024;16(6):e63255.