Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

Implementation of K-Means Clustering for Grouping Post-Flood Disease Patterns in Affected Residential Settlements

Hotler Manurung[1)], Marto Sihombing[2)], Ratih Puspadini[3)]
[1,2,3)] Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Kaputama, Binjai, Indonesia
[1)*] manurunghotler0@gmail.com

ABSTRACT

Flooding in Binjai City increases post-flood disease incidence in slum settlements due to inadequate sanitation and high environmental vulnerability, while disease records are often scattered, poorly organized, and late, slowing health interventions. This study groups post-flood disease data in affected residential areas using K-Means clustering based on disease type, village, and slum level. A total of 1,100 records were collected from five districts in Binjai City; after excluding incomplete records, 850 valid records were used in the 2-cluster scenario and 1,086 in the 3-cluster scenario, implemented in a Matlab-based application. In the 2-cluster scenario, Cluster 1 contained 536 records with centroid (3.64, 12.48, 2.41) and Cluster 2 contained 314 records with centroid (4.18, 16.25, 2.09). In the 3-cluster scenario, Cluster 1 contained 498 records with centroid (3.45, 11.62, 2.31), Cluster 2 contained 312 records with centroid (4.02, 16.47, 2.05), and Cluster 3 contained 276 records with centroid (2.91, 8.35, 2.76), all dominated by diarrhea in medium-to-high slum-level areas. Internal validity indices (Silhouette, Davies-Bouldin, Calinski-Harabasz), computed on a reference sample, support retaining the 3-cluster scheme for its finer, more actionable risk stratification. The results show that K-Means clustering groups affected areas by disease and slum characteristics, and a centroid-derived priority ranking of the clusters is proposed to support health priority setting, medical resource distribution, and data-driven post-flood disease mitigation.

Keywords: Binjai City; Data Mining; K-Means Clustering; Post-Flood Diseases; Slum Settlements

INTRODUCTION

Flooding is one of the natural disasters that frequently occurs in various regions of Indonesia, including Binjai City, which has experienced repeated flood events in recent years. This phenomenon is triggered by high rainfall, inadequate drainage systems, and climate change that causes extreme weather (Kiparisov et al., 2023). Floods not only damage houses and infrastructure but also affect public health by increasing the risk of infectious diseases, especially those related to sanitation, water quality, garbage accumulation, and highly degraded slum environments that become breeding grounds for various disease vectors (H. Li et al., 2025).

Post-flood conditions worsen public health problems, with increasing cases of diseases such as diarrhea, acute respiratory infection (ARI), skin diseases, dengue hemorrhagic fever, and leptospirosis. Based on existing records, the collection and monitoring of post-flood disease cases remain difficult. Available data are often scattered, poorly organized, and arrive late, which leads to inefficient interventions and sometimes biased allocation of health resources (Pratama et al., 2024).

Considering these challenges, a technology-based approach for grouping flood-affected areas according to disease distribution patterns is highly needed. One method that can be used to map disease distribution more efficiently and systematically is K-Means clustering (Liu et al., 2025). This method can group affected villages in slum environments based on the similarity of their disease patterns, making it easier to determine priority zones for health interventions. This research direction is also aligned with the fourth national priority agenda of Indonesia, namely strengthening human resource development, science and technology, education, and health (Prabowo & Gibran, 2024).

Using K-Means clustering, this study groups residential areas in Binjai City based on post-flood disease cases (Joo et al., 2025; Song et al., 2025). This approach is expected to produce cluster zoning that represents health risk levels, which can then be used to prioritize the distribution of medicines, medical

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

personnel, and other health interventions (J. Li et al., 2025).

Prior studies have applied K-Means clustering to grouping tasks such as employment-program participants based on income and contribution (Zema et al., 2022), new-student admission profiles (Agneresa et al., 2022), thesis-topic records (Sembiring et al., 2022), single-disease patient records without location information (Ordila et al., 2020; Adiputra, 2021), and heart-disease patients without an environmental risk indicator (Wala et al., 2024). None of these studies combined disease type with village-level location and an official slum-severity indicator, and none produced a health-risk zoning output at the village level; this combination constitutes the specific research gap addressed in this work. Building on this gap, the present study proposes a K-Means-based, slum-level-aware zoning of post-flood disease risk at the village level for the affected residential settlements of Binjai City. Its contributions are threefold: (1) a clustering model that jointly uses disease type, village, and official slum level as clustering variables; (2) a quantitative, validity-index-supported comparison of a 2-cluster and a 3-cluster zoning scheme on 1,100 disease records, together with a centroid-derived intervention-priority ranking of the resulting clusters; and (3) a Matlab-based application that supports fast, data-driven decision-making in post-flood public health management.

LITERATURE REVIEW

Several previous studies have proven the effectiveness of the K-Means algorithm in grouping data. The application of K-Means with 3 clusters on BPJS Employment participant data successfully grouped 1,000 participants based on income, contribution paid, and the program taken (Zema et al., 2022). Another study on new student data of Universitas Buana Perjuangan Karawang in 2020/2021, with a total of 2,479 records, produced 2 clusters with a Davies-Bouldin Index value close to 0, namely Cluster 1 with 1,954 records (78.82%) as the high-level cluster and Cluster 2 with 525 records (21.18%) as the low-level cluster (Agneresa et al., 2022). A comparable experiment obtained a clustering distribution of 29% in C1, 21% in C2, 22% in C3, 13% in C4, and 15% in C5 (Sembiring et al., 2022). These studies demonstrate the general effectiveness of K-Means for grouping structured records, but each relied on a single domain-specific variable set (e.g., income and contribution, or admission scores) without an environmental risk indicator such as settlement slum level, and none targeted post-disaster health-risk zoning at the village level, which motivates the variable combination used in the present study.

Data mining is the process of extracting useful information and patterns from very large data, covering data collection, extraction, analysis, and statistics; it is also known as knowledge discovery, knowledge extraction, and data or pattern analysis (Angin & Sihombing, 2024). Clustering is a data mining technique that divides data into groups of objects with similar characteristics and plays an important role in applications such as scientific data exploration, information retrieval, text mining, spatial databases, and web analysis (Manurung et al., 2024).

K-Means is a popular clustering algorithm that is widely used in industry. The algorithm is built on a simple idea: a number of clusters is determined at the beginning, initial objects are selected as cluster centroids, and the algorithm then repeats the assignment and update steps until stability is reached, that is, until no object can be moved between clusters (Fahry et al., 2025). Similarity between objects is fundamental in cluster analysis and is commonly measured by distance, such as the Euclidean and Manhattan distances (Van Doorsselaere et al., 2019). In this study, only the Euclidean distance is implemented in the clustering procedure, consistent with the default sqEuclidean configuration of the Matlab kmeans function; the Manhattan distance is presented here as a theoretical alternative and is not used in the experiments, and a comparative evaluation of both distance measures is left for future work.

METHOD

This study was carried out in five chronological stages. First, interviews were conducted with the prospective users of the system to ensure that the clustering of disease data in slum settlements matches user needs. Second, a literature study was performed on references related to clustering, disease types,

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

residential settlements, and slum areas. Third, field research was conducted through direct observation of disease problems in residential settlements located in slum areas, after which the collected data were prepared for the clustering process. Fourth, the method was implemented as Matlab code to obtain the cluster calculation results and to identify the closest relationship among disease type, village, and slum level. Fifth, an evaluation was performed to draw conclusions and formulate suggestions.

The data used in this study are 1,100 post-flood disease records of residential areas in Binjai City for 2025-2026, collected from five districts, namely Binjai Kota, Binjai Barat, Binjai Timur, Binjai Utara, and Binjai Selatan. Each record consists of three variables: disease type, village, and slum level, where the slum level refers to the official designation of slum housing and settlement locations issued by the local Housing and Settlement Agency (PERKIM). Disease type and village are nominal categorical variables, while slum level is an ordinal category (Low, Medium, High); all three were converted into integer codes assigned in the order the categories were first encountered in the dataset, so that they could be processed by the Euclidean-distance-based K-Means procedure. This label-encoding is a practical simplification rather than a measurement of true semantic distance between categories, since it implicitly imposes an ordinal relationship between nominal categories that may not correspond to their real-world similarity; this simplification is acknowledged as a methodological limitation in the Discussion section. Because data collection was carried out on a rolling basis while post-flood disease reports were still being compiled from the primary healthcare centers (Puskesmas) in the five districts, records with incomplete disease-type, village, or slum-level information at the time of each testing session were excluded prior to clustering. As a result, the 2-cluster experiment was conducted on 850 valid records extracted at an earlier stage of data compilation, while the 3-cluster experiment was conducted on 1,086 valid records extracted after additional reports had been compiled and cleaned; both extractions were drawn from the same overall pool of 1,100 collected post-flood disease records. For the manual calculation of the algorithm, a reference sample of 20 records was used, as shown in Table 1. This reference sample includes 6 skin disease cases (30%), 4 diarrhea cases (20%), 3 gastroenteritis cases (15%), 2 asthma cases (10%), 2 dengue hemorrhagic fever (DHF) cases (10%), and 1 case each of tuberculosis, malaria, and hepatitis (5% each), spread across 19 different villages, with 10 records (50%) at a high slum level, 9 records (45%) at a medium slum level, and 1 record (5%) at a low slum level.

Table 1. Sample of Post-Flood Disease Data

No.	Name	Age	Village	Disease Type	Slum Level
1	Budi Harahap	34	Sejahtera	Skin Disease	Medium
2	Indah Saputra	25	Jati Karya 3	Diarrhea	Low
3	Indah Lestari	55	Lorong Sungai	Tuberculosis	Medium
4	Nur Saputra	53	Jambi	Diarrhea	Medium
5	Ahmad Lubis	79	Juanda 1	Malaria	High
6	Rahmat Harahap	74	Mungkur	Asthma	Medium
7	Fajar Syahputra	31	Sinembah 2	DHF	Medium
8	Muhammad Pratama	41	Kopi	Skin Disease	Medium
9	Ahmad Lubis	5	Satria 1	Skin Disease	High
10	Tono Lestari	50	Satria 1	Skin Disease	High
11	Fitri Syahputra	63	satria 2	DHF	High
12	Aisyah Hasibuan	80	Satria 3	Diarrhea	High
13	Agus Lestari	36	Satria 4	Asthma	High
14	Rizki Simanjuntak	29	Syafiah	Gastroenteritis	High
15	Agus Santoso	47	Amal	Hepatitis	Medium
16	Rahmat Santoso	39	Seksama	Gastroenteritis	Medium
17	Ahmad Saputra	30	Mawar	Gastroenteritis	Medium
18	Muhammad Lubis	47	Kartini 1	Diarrhea	High
19	Putri Wijaya	25	Kartini 2	Skin Disease	High
20	Dewi Siregar	6	Berngam Atas	Skin Disease	High

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

The K-Means clustering method was selected because it has been proven effective and simple to apply in grouping health-related and social data (Angin & Sihombing, 2024; Sembiring et al., 2022). In K-Means, the similarity between objects X = [X1, X2, ..., Xp] and Y = [Y1, Y2, ..., Yp] of p dimensions is measured by a distance function, namely the Euclidean distance in equation (1) and the Manhattan distance in equation (2); only the Euclidean distance in equation (1) is applied in the clustering experiments reported in this study, while the Manhattan distance is retained as a theoretical alternative for comparison in future work.

$$dEuclidean : X, Y = \sqrt{\Sigma_i (Xi - Yi)^2} \quad (1)$$

$$dManhattan : X, Y = \sqrt{\Sigma_i |Xi - Yi|} \quad (2)$$

The procedure begins by determining the number of clusters, after which the initial centroids are chosen randomly from the data. The distance of every object to each centroid is then computed using the Euclidean distance, and each object is assigned to the cluster with the nearest centroid. New centroids are subsequently computed as the average of the members of each cluster, and the process is repeated until no object moves between clusters. The workflow of the method is shown in Figure 1.

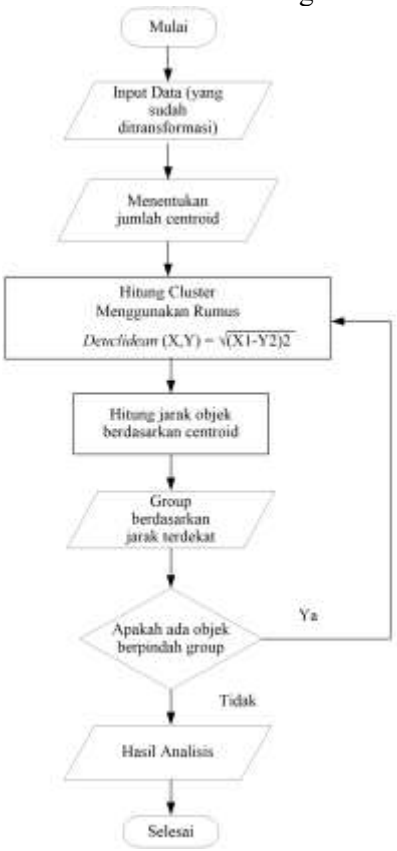

Figure 1. Flowchart of the K-Means Clustering Method

Cluster Validity and Selection of the Number of Clusters.

The number of clusters was tested for k = 2 and k = 3, following the two-tier and three-tier health-risk zoning scheme commonly used in comparable clustering studies for exploratory, decision-support purposes (Agneresa et al., 2022; Zema et al., 2022). To support this choice with an internal validity assessment rather than an unverified assumption, the Silhouette Coefficient, the Davies-Bouldin Index (DBI), and the Calinski-Harabasz Index were computed for both configurations on the reference sample of 20 records shown in Table 2, using the disease type, village, and slum-level variables as clustering features. For k = 2, the resulting values were a Silhouette Coefficient of 0.52, a DBI of 0.63, and a Calinski-Harabasz Index of

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

36.83, while for k = 3 the values were a Silhouette Coefficient of 0.38, a DBI of 0.80, and a Calinski-Harabasz Index of 32.61. These indices indicate that, on the reference sample, the 2-cluster configuration exhibits slightly stronger internal cohesion and separation than the 3-cluster configuration. The 3-cluster scheme is nevertheless retained as the primary scenario reported in the Result and Discussion sections because it provides a finer three-level risk stratification (high, medium, and low priority, as detailed in Table 5) that is more directly actionable for targeted health interventions, while the 2-cluster scheme is reported alongside it as a simpler, more internally cohesive alternative. Both configurations are reported so that readers and future studies can weigh internal cluster quality against the operational resolution required for zoning decisions; a full re-computation of these indices on the complete 1,100-record dataset is recommended as part of future validation.

The results were measured and evaluated in two complementary ways that serve different purposes. The manual calculation on the 20-record reference sample (Table 1 and Table 2) was used to demonstrate and verify the correctness of the K-Means procedure step by step, by checking the convergence of cluster memberships between iterations; this manual result is illustrative and is not intended to be numerically identical to the large-scale result, since it uses a much smaller sample. The full dataset of post-flood disease records was separately tested in the Matlab application using two scenarios, namely 2 clusters (850 valid records) and 3 clusters (1,086 valid records), which were then compared based on cluster sizes, the interpretation of the resulting centroids, and the internal validity indices reported above.

RESULT

Before clustering, the categorical data were initialized into numerical form so that each record can be expressed by the independent variables Disease Type (X), Village (Y), and Slum Level (Z). The transformed data of the 20-record sample are shown in Table 2.

Table 2. Data Transformation

No.	Disease Type (X)	Village (Y)	Slum Level (Z)
1	2	17	2
2	4	4	1
3	7	10	2
4	4	3	2
5	5	5	3
6	1	12	2
7	3	19	2
8	2	8	2
9	2	13	3
10	3	14	3
11	4	15	3
12	1	16	3
13	6	20	3
14	8	1	2
15	6	18	2
16	6	11	2
17	4	6	3
18	2	7	3
19	2	2	3
20	5	9	3

In the first iteration, the initial centroids were chosen randomly from the transformed data, namely Centroid 1 = (2, 17, 2) taken from record 1, Centroid 2 = (7,10,2) taken from record 3, and Centroid 3 = (5, 5, 3) taken from record 5. The distance of each record to every centroid was computed using equation (1), and each record was assigned to the nearest centroid. The results of the first iteration are shown in Table 3.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

Table 3. Results of Iteration I

No	Disease Type (X)	Village (Y)	Slum Level (Z)	Distance to C1	Distance to C2	Distance to C3	Group
1	2	17	2	0	8.60	12.41	1
2	4	4	1	13.19	6.78	2.44	3
3	7	10	2	8.60	0	5.47	2
4	4	3	2	14.14	7.61	2.44	3
5	5	5	3	12.41	5.47	0	3
6	1	12	2	5.09	6.32	8.12	1
7	3	19	2	2.23	9.84	14.17	1
8	2	8	2	9	5.38	4.35	3
9	2	13	3	4.12	5.91	8.54	1
10	3	14	3	3.31	5.74	9.22	1
11	4	15	3	3	5.91	10.04	1
12	1	16	3	1.73	8.54	11.70	1
13	6	20	3	5.09	10.10	15.03	1
14	8	1	2	17.08	9.05	5.09	3
15	6	18	2	4.12	8.06	13.07	1
16	6	11	2	7.21	1.41	6.16	2
17	4	6	3	11.22	5.09	1.41	3
18	2	7	3	10.05	5.91	3.60	3
19	2	2	3	15.03	9.48	4.24	3
20	5	9	3	8.60	2.44	4	2

The first iteration changed the memberships from the initial group {0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0} to the new group {1,3,2,3,3,1,1,3,1,1,1,1,1,3,1,2,3,3,3,2}. Because the memberships changed, the calculation continued to the next iteration with new centroids computed from the average of the members of each group, namely Centroid 1 = (3, 16, 3), Centroid 2 = (6, 10, 2), and Centroid 3 = (4, 5, 2). The results of the second iteration are shown in Table 4.

Table 4. Results of Iteration II

No	Disease Type (X)	Village (Y)	Slum Level (Z)	Distance to C1	Distance to C2	Distance to C3	Group
1	2	17	2	1.73	8.06	12.16	1
2	4	4	1	12.20	6.40	1.41	3
3	7	10	2	7.28	1	5.83	2
4	4	3	2	13.07	7.28	2	3
5	5	5	3	11.18	5.19	1.41	3
6	1	12	2	4.58	5.38	7.61	1
7	3	19	2	3.16	9.48	14.03	1
8	2	8	2	8.12	4.47	3.60	3
9	2	13	3	3.16	5.09	8.30	1
10	3	14	3	2	5.09	9.11	1
11	4	15	3	1.41	5.47	10.05	1
12	1	16	3	2	7.87	11.44	1
13	6	20	3	5	10.05	15.16	1
14	8	1	2	15.84	9.22	5.65	3
15	6	18	2	3.74	8	13.15	1
16	6	11	2	5.91	1	6.32	2
17	4	6	3	10.05	4.58	1.41	3
18	2	7	3	9.05	5.09	3	3
19	2	2	3	14.03	9	3.74	3
20	5	9	3	7.28	1.73	4.24	2

The second iteration produced exactly the same memberships as the first iteration, namely {1,3,2,3,3,1,1,3,1,1,1,1,1,3,1,2,3,3,3,2}, so the iteration was stopped. From the 20 sample records, three groups were formed: group 1 with 9 records, group 2 with 3 records, and group 3 with 8 records. Cluster 1,

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

with centroid (3, 16, 3), is characterized by dengue hemorrhagic fever (DHF) cases in the Satria 4 area with a high slum level. Cluster 2, with centroid (6, 10, 2), is characterized by gastroenteritis cases in the Lorong Sungai area with a medium slum level. Cluster 3, with centroid (4, 5, 2), is characterized by diarrhea cases in the Satria 2 area with a medium slum level.

The complete dataset of 1,100 post-flood disease records was then processed in the Matlab-based application. In the 2-cluster scenario, whose result is shown in Figure 2, using the 850 valid records extracted for this experiment, Cluster 1 has 536 members with centroid (3.64, 12.48, 2.41), dominated by diarrhea cases in Kartini 1 Village with a high slum level, which indicates a relatively high health risk that requires special attention in post-flood disease prevention and treatment. Cluster 2 has 314 members with centroid (4.18, 16.25, 2.09), dominated by diarrhea cases in Sumberkarya Village with a medium slum level, indicating a lower health risk than Cluster 1 that nevertheless still requires continuous monitoring.

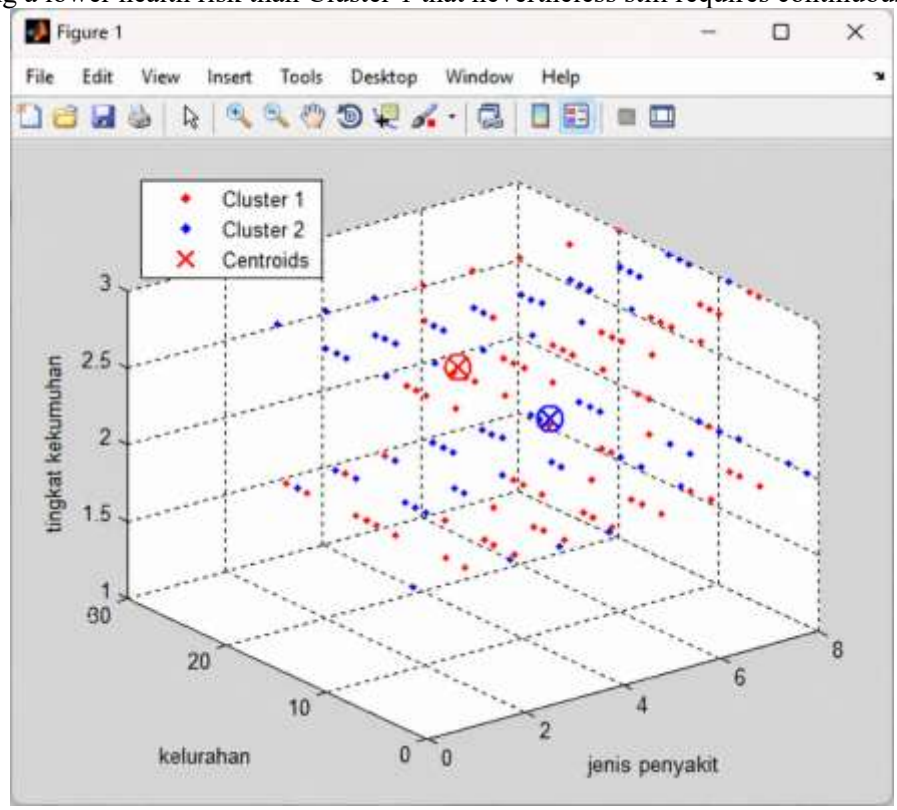

Figure 2. Clustering Result of the 2-Cluster Scenario

In the 3-cluster scenario, whose result is shown in Figure 3, using the 1,086 valid records extracted for this experiment, Cluster 1 has 498 members with centroid (3.45, 11.62, 2.31), dominated by diarrhea cases in Kartini 2 Village with a high slum level. Cluster 2 has 312 members with centroid (4.02, 16.47, 2.05), dominated by diarrhea cases in Sumberkarya Village with a medium slum level. Cluster 3 has 276 members with centroid (2.91, 8.35, 2.76), dominated by diarrhea cases in Tanjung Jati Village with a high slum level. These centroid values show that each cluster has distinct characteristics, so the affected areas can be distinguished according to their disease patterns and environmental conditions.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

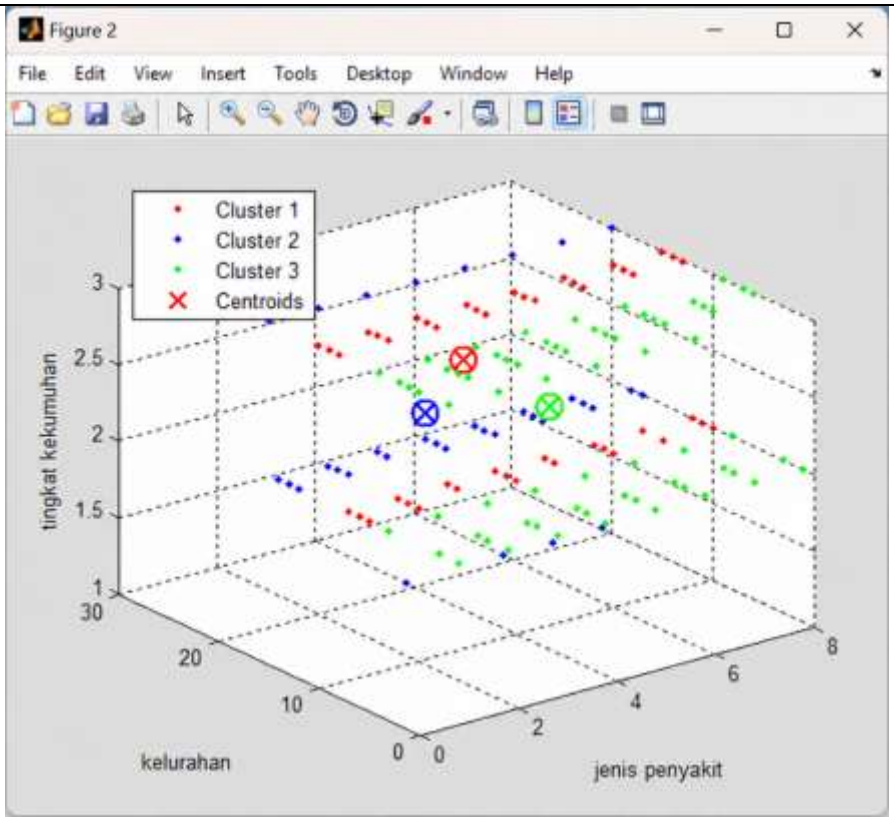

Figure 3. Clustering Result of the 3-Cluster Scenario

The dominant disease category differs between the 20-record reference sample, where Cluster 1, 2, and 3 are respectively characterized by DHF, gastroenteritis, and diarrhea (Table 3 and Table 4), and the full-scale clustering result, where diarrhea is the dominant disease in every cluster of both the 2-cluster and 3-cluster scenarios. This difference arises because the reference sample is too small (20 records) to reflect the true disease distribution of the full dataset, whereas diarrhea is overwhelmingly the most frequent post-flood disease at full scale, consistent with the well-established mechanism by which flooding contaminates water sources and disrupts sanitation, thereby increasing waterborne disease transmission (H. Li et al., 2025). Consequently, at full scale the clusters are differentiated mainly by their village location and slum-level dimensions rather than by disease type, so disease type should be interpreted as a secondary descriptive attribute of each cluster rather than its primary distinguishing feature.

Table 5 summarizes the profile of each cluster in both scenarios, together with an intervention-priority rank derived from the average slum-level (Z) value of each cluster, where a higher Z indicates a more severe, higher-priority zone.

Table 5. Cluster Profile Summary and Intervention-Priority Ranking

Scenario	Cluster	N	Centroid (X, Y, Z)	Dominant Disease	Dominant Village	Priority Rank
2-cluster	Cluster 1	536	(3.64, 12.48, 2.41)	Diarrhea	Kartini 1	1 (highest)
2-cluster	Cluster 2	314	(4.18, 16.25, 2.09)	Diarrhea	Sumberkarya	2 (lowest)
3-cluster	Cluster 3	276	(2.91, 8.35, 2.76)	Diarrhea	Tanjung Jati	1 (highest)
3-cluster	Cluster 1	498	(3.45, 11.62, 2.31)	Diarrhea	Kartini 2	2
3-cluster	Cluster 2	312	(4.02, 16.47, 2.05)	Diarrhea	Sumberkarya	3 (lowest)

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

DISCUSSIONS

The clustering results provide a K-Means-based, slum-level-aware zoning of post-flood disease risk at the village level, summarized as an intervention-priority ranking in Table 5. In the 3-cluster scenario, Cluster 3 (Tanjung Jati Village, Z = 2.76) is ranked as the top-priority zone despite having fewer members than Cluster 1, because its members are the most homogeneously concentrated in the high slum-level category; Cluster 1 (Kartini 2 Village, Z = 2.31) is ranked second, and Cluster 2 (Sumberkarya Village, Z = 2.05) is ranked third and mainly requires continuous monitoring rather than urgent intervention. In practice, this ranking can be used directly: Cluster 3 and Cluster 1 should be prioritized for oral rehydration salts and basic medicine distribution, temporary health post placement, and early inclusion in a post-flood disease early-warning system, while Cluster 2 can be scheduled for periodic monitoring visits; all three clusters' locations and priority ranks can further be integrated into a local government or health-office dashboard to support real-time, area-based decision-making. Compared with the 2-cluster scenario, the 3-cluster scenario separates the affected areas into more operationally specific groups, which is why it is retained as the primary scenario in this study even though, as reported above, its internal validity indices are marginally lower than those of the 2-cluster scenario on the reference sample.

These findings are consistent with previous studies that applied K-Means to health data, such as the clustering of dengue hemorrhagic fever patients at Dharma Kerti Hospital (Adiputra, 2021), the grouping of patient medical records by disease type (Ordila et al., 2020), and the clustering of heart disease patients (Wala et al., 2024), all of which showed that K-Means produces meaningful groups for health decision support. The results also agree with the finding that environmental disasters increase the burden of disease in affected populations (Mpakosi et al., 2024). In contrast to those studies, this work combines disease type with the official slum level of settlements, so the resulting clusters directly represent post-flood risk zones at the village level. Unlike Adiputra (2021) and Ordila et al. (2020), whose clusters were formed from a single hospital's or clinic's records and therefore reflect clinical rather than geographic risk patterns, the clusters obtained in this study are shaped jointly by disease type and settlement slum level, which explains why diarrhea, rather than a single dominant clinical category as in those single-source studies, emerges as the shared descriptive feature across clusters while village-level slum severity becomes the primary differentiator; this reflects the environmental, sanitation-driven transmission pathway typical of post-flood outbreaks (Mpakosi et al., 2024), rather than a purely clinical clustering pattern.

This study has several limitations. The analysis uses only three variables, and the slum level is taken from an official designation that may change over time. Disease type and village were converted into integer codes for computational purposes, which implicitly imposes an ordinal relationship between nominal categories under the Euclidean distance measure; this is a simplification rather than a semantically accurate distance measure, and alternative approaches such as K-Prototypes clustering, Gower distance, or one-hot encoding should be explored in future work. The Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index reported in the Method section were computed only on the 20-record reference sample rather than on the complete dataset, so re-computing these indices on the full 850-record and 1,086-record extractions is recommended to confirm the optimal number of clusters at full scale. The performance of the Matlab-based application itself, including its computation time, usability, and scalability to larger datasets, was not benchmarked in this study and remains an open evaluation task. Finally, the scope of the data is limited to Binjai City and to the extractions available at the time of each testing session, so generalization to other regions or to the complete, fully reconciled dataset must be done carefully.

CONCLUSION

This study implemented a K-Means-based, slum-level-aware zoning of post-flood disease risk at the village level for the affected residential settlements of Binjai City, using disease type, village, and slum level as clustering variables. In the 2-cluster scenario, tested on 850 valid records, the method produced Cluster 1 with 536 records and centroid (3.64, 12.48, 2.41) and Cluster 2 with 314 records and centroid (4.18, 16.25, 2.09); in the 3-cluster scenario, tested on 1,086 valid records, it produced Cluster 1 with 498

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

records and centroid (3.45, 11.62, 2.31), Cluster 2 with 312 records and centroid (4.02, 16.47, 2.05), and Cluster 3 with 276 records and centroid (2.91, 8.35, 2.76); both extractions were drawn from the overall pool of 1,100 collected records. The 3-cluster scheme was retained as the primary scenario because it provides a finer three-level risk stratification that supports the centroid-derived intervention-priority ranking presented in Table 5, even though internal validity indices computed on the reference sample indicate marginally stronger cohesion for the 2-cluster scheme. The contribution of this study is therefore threefold: a joint disease-type, village, and slum-level clustering model; a validity-index-supported, ranked zoning of post-flood health risk; and a Matlab-based application that can be used by local governments and health agencies to determine treatment priorities, distribute medicines and medical personnel, and design mitigation strategies. Future work should enrich the analysis with additional variables such as population size, vulnerable age groups, rainfall, and inundation depth; recompute internal cluster validity measures on the complete, fully reconciled dataset; benchmark the computational performance and usability of the Matlab application; evaluate alternative encodings or distance measures such as K-Prototypes or the Manhattan distance for the nominal disease-type and village variables; and develop a web-based or geographic information system to present the clustering and priority-ranking results interactively.

ACKNOWLEDGMENT

The authors thank the Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia, which supported and funded this research in 2026, and STMIK Kaputama for the institutional support during the research.

REFERENCES

Adiputra, I. N. M. (2021). Clustering penyakit DBD pada Rumah Sakit Dharma Kerti menggunakan algoritma K-Means. INSERT: Information System and Emerging Technology Journal, 2(2), 99-105. https://ejournal.undiksha.ac.id/index.php/insert/article/view/41673

Agneresa, Hananto, A. L., Hilabi, S. S., Hananto, A., & Tukino. (2022). Strategi promosi penerapan data mining mahasiswa baru dengan metode K-Means clustering. Dirgamaya: Jurnal Manajemen dan Sistem Informasi, 2(2), 25-34.

Angin, S. P., & Sihombing, M. (2024). Clustering data on underage marriage. 3(November), 195-200.

Fahry, F., Miswaty, T. C., & Harun, H. (2025). Analyzing marketplace reviews using Word2Vec, CNN, and deep K-Means with sociolinguistic approaches. Jurnal Teknik Informatika, 6(6), 5489-5502. https://jutif.if.unsoed.ac.id/index.php/jurnal/article/view/5340

Joo, D., Na, R., Kim, H., Yoo, S., & Lee, S. (2025). Analysis of application of design standards for future climate change adaptive agricultural reservoirs using cluster analysis. 1-19.

Kiparisov, P., Lagutov, V., & Pflug, G. (2023). Quantification of loss of access to critical services during floods in Greater Jakarta: Integrating social, geospatial, and network perspectives. Remote Sensing, 15(21), 1-23.

Li, H., Huang, J., Zhang, X., Meng, Z., Fan, Y., & Wu, X. (2025). Flash flood risk classification using GIS-based fractional order k-means clustering method. Fractal and Fractional, 9(9), 1-18.

Li, J., Meng, Z., Zhang, J., Chen, Y., Yao, J., & Li, X. (2025). Prediction of seawater intrusion run-up distance based on K-Means clustering and ANN model. Journal of Marine Science and Engineering, 13(2), 1-18.

Liu, C., Feng, Q., Zhou, W., Zhang, C., & Zhang, X. (2025). Flow field evaluation method of high water-cut reservoirs based on K-Means clustering algorithm. Symmetry, 17(6), 1-18.

Manurung, H. (2024). Pengelompokan UMKM Kota Binjai menggunakan metode clustering K-Means untuk mengidentifikasi pola. 8(2), 93-99.

Mpakosi, A., Cholevas, V., Tzouvelekis, I., Passos, I., Kaliouli-Antonopoulou, C., & Mironidou-Tzouveleki, M. (2024). Autoimmune diseases following environmental disasters: A narrative review of the literature. Healthcare, 12(17), 1767. https://www.mdpi.com/2227-

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.9210

Submitted : 2026-07-14
Accepted : 2026-07-21
Published : 2026-07-22

9032/12/17/1767/htm

Ordila, R., Wahyuni, R., Irawan, Y., & Sari, M. Y. (2020). Penerapan data mining untuk pengelompokan data rekam medis pasien berdasarkan jenis penyakit dengan algoritma clustering. Jurnal Ilmu Komputer, 9(2), 148-153. https://jik.htp.ac.id/index.php/jik/article/view/181

Prabowo, & Gibran. (2024). Asta Cita: Visi misi Prabowo-Gibran.

Pratama, R. Z., Sihombing, M., & Ambarita, I. (2024). Application of data mining to measure the level of satisfaction with public facilities and services at STMIK Kaputama Binjai using linear regression method. Journal of Artificial Intelligence and Engineering Applications, 4(1), 411-418.

Sembiring, C. S. D. B., Hanum, L., & Tamba, S. P. (2022). Penerapan data mining menggunakan algoritma K-Means untuk menentukan judul skripsi dan jurnal penelitian (studi kasus FTIK UNPRI). Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA), 5(2), 80-85.

Song, H., Li, S., Yu, B., Sun, Y., Tao, Q., & Peng, L. (2025). Automatic text summary method based on optimized K-Means clustering algorithm with symmetry and maximal-marginal-relevance algorithm. 1-24.

Van Doorsselaere, T., Shariati, H., & Debosscher, J. (2019). Time series optimization on data mining. Journal of Physics: Conference Series, 1235(1), 012014. https://iopscience.iop.org/article/10.1088/1742-6596/1235/1/012014

Wala, J., Herman, H., & Umar, R. (2024). Implementasi K-Means clustering pada pengelompokan pasien penyakit jantung. JISKA (Jurnal Informatika Sunan Kalijaga), 9(3), 205-216. https://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/4458

Zema, Maulita, Y., & Arliana, L. (2022). Penerapan data mining pengelompokan peserta BPJS Ketenagakerjaan berdasarkan program yang diambil menggunakan metode clustering. Jurnal Sistem Informasi Kaputama, 6(2), 152-164.

* Corresponding author