Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

K-Means Algorithm for Grouping Models of Dengue Fever Prone Areas in Cirebon City

Aida Safitri[1]*, Ade Irma Purnamasari[2], Agus Bahtiar[3], Edi Tohidi[4]

[1,2,3,4] STMIK IKMI Cirebon

aidasafitri382@gmail.com[1], irma2974@yahoo.com[2], Agusbahtiar038@gmail.com[3], editohidi00@gmail.com[4]*

Abstract

Dengue hemorrhagic fever (DHF) is an infectious disease transmitted through the Aedes aegypti mosquito. DHF cases in Cirebon City show a significant increase every year. This study aims to classify dengue prone areas based on case data per health center in 2020-2024 obtained from the Cirebon City Health Office. The method used is the K-Means algorithm with the Knowledge Discovery in Database (KDD) approach, which includes data selection, preprocessing, data transformation, data mining, evaluation, and knowledge. Evaluation using Davies-Bouldin Index (DBI) showed optimal results at k = 6 with a DBI value of -0.445. The clustering results produced six clusters: cluster 5 (437 dengue cases in 34 health centers) showed high risk; cluster 0 (244 cases), cluster 2 (129 cases), and cluster 3 (279 cases) showed medium risk; while cluster 1 (69 cases) and cluster 4 (86 cases) showed low risk. This study shows that the K-Means algorithm is effective in identifying DHF risk distribution patterns and provides a strategic basis for the Cirebon City Health Office to prioritize interventions and develop more effective prevention strategies.

Keywords: *K-Means; Dengue fever; Clustering; DBI; Vulnerable areas*

1. Introduction

Dengue fever is an acute illness caused by the Dengue virus, which belongs to the arthropod-borne virus, genus Flavivirus, and Flaviviridae, the virus is transmitted through the bite of Aedes aegypti and Aedes albopictus mosquitoes.[1] In Indonesia, dengue fever is still a major health problem. This mosquito has a lifestyle in the tropics and subtropics.[2] At this time dengue fever is spreading in all corners of Indonesia including in Cirebon City.

Based on data obtained from the Cirebon City Health Office, dengue fever cases show an increase every year, so it is necessary to classify areas prone to dengue fever to provide a more effective disease control strategy. Previous research, such as that conducted by [3] in Samarinda City, with the title "k-means for clustering dengue fever prone areas" shows that the K-Means algorithm is effective for grouping dengue fever prone areas into three clusters based on the level of risk. Another study by [4] in Tasikmalaya City showed that the k-means algorithm was effective in clustering areas by producing two groups of data, namely sub-district data groups and puskesmas data groups.

Research by [5] in Bontang City also showed the success of K-Means in clustering areas with different levels of dengue cases. Meanwhile, research [6] entitled Implementation of K-Means in Clustering the Spread of DHF Disease in West Java, research [7] [8] integrated the K-Means method with other approaches such as Elbow Method and Knowledge Discovery in Database (KDD) which resulted in optimal clustering and accurate case prediction.

This research aims to apply the K-Means algorithm in clustering dengue fever prone areas in Cirebon City. The contribution of this research is to assist the Cirebon City Health Office in designing effective dengue fever control strategies and also provide input for the development of further research related to the application of the K-Means algorithm in the Health context.

2. Literature Review

2.1. Data Mining

Data mining is a process that uses data analysis techniques to find patterns, relationships, or hidden information in very large and complex data sets. The goal of this process is to discover and understand patterns, which results in useful knowledge. One method that is often used in data mining is clustering, which serves to group data based on certain similarities.

2.2. Clustering

Data clustering is a process or phase of grouping data into specific parts to obtain results that correspond to a type or kind of data. Grouping data into several specific parts to obtain results that are of the same type or kind.[9] In grouping dengue fever (DHF), clustering is used to

group cases based on certain similarities. This method aims to identify patterns of disease distribution, such as areas with high, medium and low case rates.

2.3 K-Means

The K-Means algorithm is a clustering method that divides data into several groups or clusters based on similarities to obtain similar or similar results. This algorithm works by determining a predetermined number of clusters (k), the k-means algorithm grouping uses the K formula, where K=2 indicates group 2, and K=3 indicates group 3, and then 2.3.[10].

3. Research Methods

All paragraphs must be justified alignment. With justified alignment, both sides of the paragraph are straight.

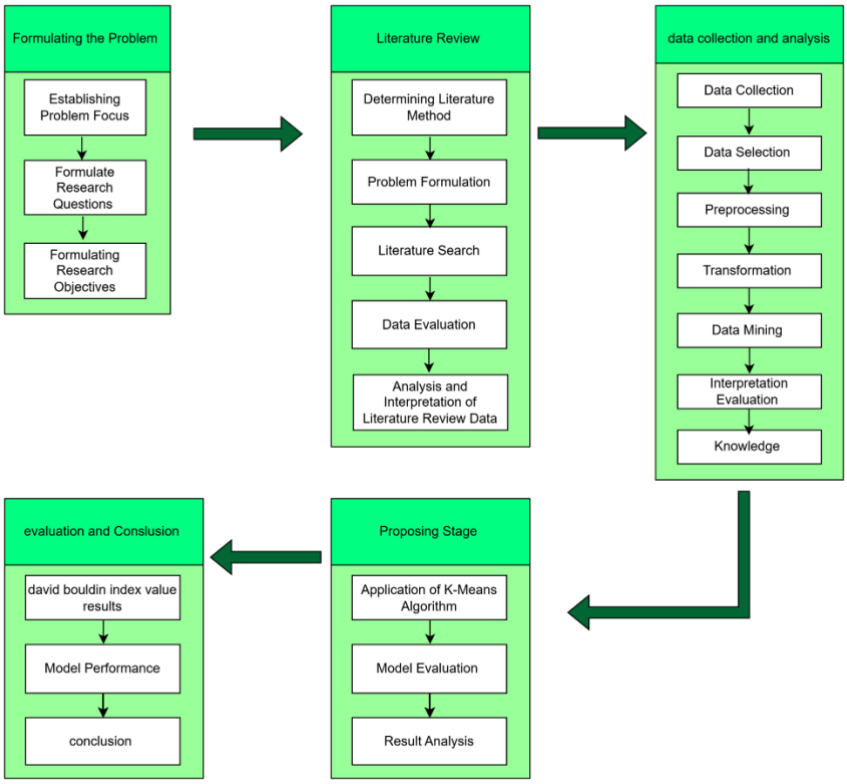

Fig. 1: Description of Research Method Activities

Table 1: Description of Research Method Activies

Stages	activity	activity description
Formulating the Problem	Establishing Problem Focus	Identify the problems encountered in clustering dengue fever prone distribution areas with the K-Means method.
	Formulate Research Question	Formulate clear and specific research questions.
	Formulate Research Objectives	Formulate specific, measurable and achievable research objectives based on research questions with available resources and time. Therefore, the objectives set are as follows: 1. Apply the k-means algorithm in clustering dengue prone areas. 2. analyzing the results of clustering dengue prone areas using the k-means algorithm. 3. evaluate the results of clustering dengue prone areas with the k-means algorithm.
Literature Review	Determining Literature Method	Choosing the right method for literature searches such as keyword searches that are suitable for research such as k-means, clustering, and dengue fever.
	Problem Formulation	Develop research questions by reviewing the literature reviewed to determine the next research process.
	Literature Search	Collecting relevant literature according to the research topic taken by searching keywords or others, from various sources such as publish or perish, google scholar, Scopus.
	Data Evaluation	To assess the quality and relevance of data found in the literature to ensure validity and reliability.
	Analysis and Interpretation of Literature Review Data	Analyze the data from the literature collected and interpret the findings to identify patterns of gaps and research implications.

Data Collection and analysis	Data Collection	Data was obtained by observation at the Cirebon City Health Office. Dengue Fever Case Data with vulnerable years 2020-2024 has several information such as City name, Sub-district name, UPT Puskesmas, Number of Cases, Unit and Year.
	Data Selection	Formulate the problems encountered in grouping dengue fever prone distribution areas with the K-Means method. And formulate objectives,
	Preprocessing	The dataset will undergo a cleaning process or called Data Cleaning.
	Transformation	Convert category data to numeric
	Data Mining	Apply data mining modeling techniques using the K-Means Clustering algorithm to classify Dengue Fever Case Data in Cirebon City.
	Interpretation Evaluation	Evaluate the results of the data mining process, such as the modeling that has been formed, the most optimal David Bouldin Index (DBI) value that has been obtained during cluster testing.
	Knowledge	Provide visualization of dengue fever spread patterns in various health center and DBI value results.
Proposing Stage	Application of K-Means Algorithm	Applying K-Means Algorithm to generate the optimal k value
	Model Evaluation	Using David bouldin index test data to measure model performance
	Result Analysis	Analyze the clustering results to understand the characteristics of each cluster.
Evaluation and Conslusion	david bouldin index value results	Presenting the results of the DBI value test, the lowest DBI value shows the most optimal cluster.
	Model Performance	Produce a visualization of the clustering model of dengue fever cases.
	conclusion	Summarize the results, provide recommendations for future research, and discuss limitations.

4. Result and Discussion

4.1. Result

The results of this study present data on the number of dengue fever cases recorded in Cirebon City from 2020 to 2024, which were taken from various sub-districts and health centers. The next step will group the data using the k-means algorithm by applying the Knowledge Discovery in Database (KDD) method.

4.1.1 Data Selection

The first stage is to select data and attributes that are relevant and in accordance with the research to be carried out. The data to be processed is data on dengue fever cases in Cirebon City from 2020-2024 with attributes id, Bps_name_Kecamatan, Upt_Puskesmas, Jumlah_kasus, and Year. The dataset table can be seen below.

Id	Bps_nama_Kecamatan	Upt_Puskesmas	Jumlah_kasus	Tahun
1.	Kejaksan	kesenden	4	2020
2.	Kejaksan	Kejaksaan	0	2020
3.	Kejaksan	Sukapura	5	2020
4.	Kejaksan	Kebonbaru	2	2020
5.	kesambi	Kesambi	5	2020
…	…	…	…	…
106	Harjamukti	Kalitanjung	16	2024
107	Harjamukti	Larangan	37	2024
108	Harjamukti	Perumnas Utara	23	2024
109	Harjamukti	Sitopeng	45	2004
110	Harjamukti	Kalijaga permai	86	2004

Fig. 2: Data Selection

4.1.2 Preprocessing Data

The second stage after data selection is data preprocessing. This process aims to eliminate missing values, inconsistent data, and data noise. At this stage, researchers use 2 operators, namely the Set Role and Select attribute operators.

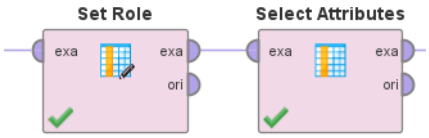

Fig. 3 Operator Preprocessing Data

a. Set Role: Used to change the role of attributes in this operator, researchers only change the Upt_Puskesmas attribute as id. So that in the display of the results of the operator set role the name UPT Puskesmas is marked blue.

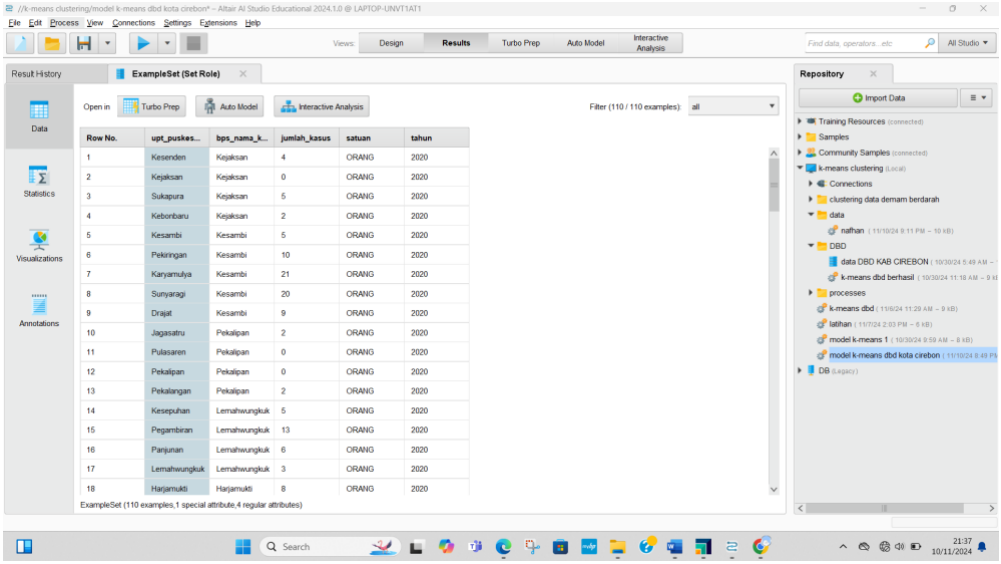

Fig. 4: Set Role result

b. Select Attributes: The second step of the select attributes operator is used to select several attributes based on the list of attribute names. In the parameters of the select attributes operator, there are several that must be changed, starting from the attribute filter type to a subset, this is used to remove attributes that are not used in this study. After doing the attribute selection process, researchers only selected 4 attributes from the dataset which originally amounted to 5 attributes. The attributes used are bps_name_kecamatan, upt_puskesmas, number_cases, year.

ExampleSet (Select Attributes)

Open in Turbo Prep Auto Model Interactive Analysis

Row No.	upt_puskes...	bps_nama_k...	jumlah_kasus	tahun
1	Kesenden	Kejaksan	4	2020
2	Kejaksan	Kejaksan	0	2020
3	Sukapura	Kejaksan	5	2020
4	Kebonbaru	Kejaksan	2	2020
5	Kesambi	Kesambi	5	2020
6	Pekiringan	Kesambi	10	2020
7	Karyamulya	Kesambi	21	2020
8	Sunyaragi	Kesambi	20	2020
9	Drajat	Kesambi	9	2020
10	Jagasatru	Pekalipan	2	2020
11	Pulasaren	Pekalipan	0	2020
12	Pekalipan	Pekalipan	0	2020
13	Pekalangan	Pekalipan	2	2020
14	Kesepuhan	Lemahwungkuk	5	2020
15	Pegambiran	Lemahwungkuk	13	2020
16	Panjunan	Lemahwungkuk	6	2020
17	Lemahwungkuk	Lemahwungkuk	3	2020
18	Harjamukti	Harjamukti	8	2020

ExampleSet (110 examples,1 special attribute,3 regular attributes)

Fig. 5: Select Attributes result

4.1.3. Transformation Data

Transformation is the process of changing the format or value of data to facilitate analysis or processing into another form suitable for the modeling process. This process aims to ensure that the K-Means algorithm can process data in an appropriate format, because this algorithm requires input in the form of numerical data.

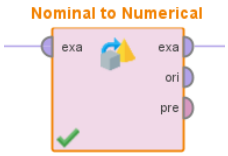

Fig. 6: Operator Transformation

a. Nominal to Numerical: The Nominal to Numerical operator is used to change categories of data (nominal), such as text or labels, into numeric data so that it can be processed by analysis algorithms that require a number format.

4.1.4 Data Mining

In the data mining stage, researchers use the k-means algorithm involving several operators such as Read Excel, Set Role, Select Attributes, Nominal to Numerical, Clustering, and performance.

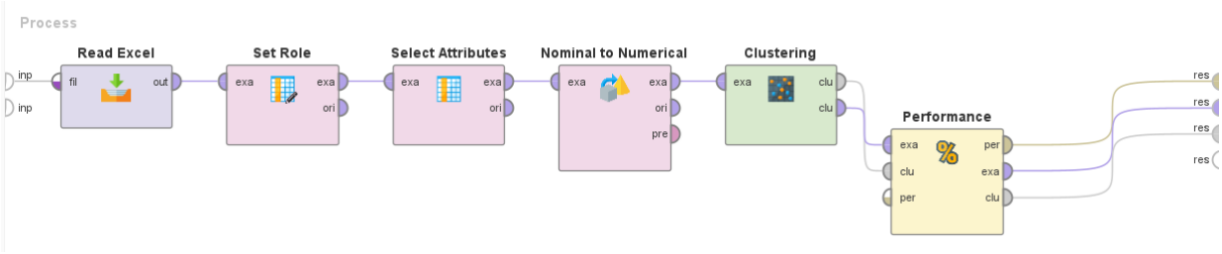

Fig. 8: K-Means algorithm model

4.1.5 Evaluation

At the evaluation stage, the cluster results will be tested using the DBI value for the Cluster Distance Performance parameter. The cluster with the smallest DBI value or close to 0 is used as the best cluster. Researchers conducted experiments from cluster 2 to cluster 10 to determine the best cluster.

Table.2: test results for David Bouldin index values

Cluster	David Bouldin Index (DBI)
2	-0.477
3	-0.524
4	-0.501
5	-0.557
6	-0.445
7	-0.546
8	-0.615
9	-0.668
10	-0.585

From the David Bouldin Index (DBI) value table above, it can be concluded that the optimal cluster value is cluster 6 (k=6) with a DBI result of -0.445. The lower the DBI value, the better the cluster formed.

4.1.6 Knowledge

Based on evaluation with DBI, the best number of clusters is k = 6 with a DBI value of -0.445. with cluster members 0: 59 items, cluster 2: 3 items, cluster 3: 12 items, cluster 4: 1 items, cluster 5: 34 items. The following are the cluster model results:

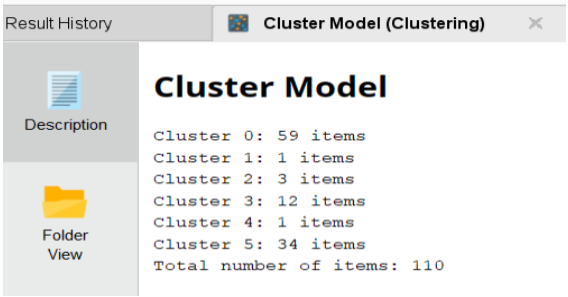

Fig. 9: Description Cluster Model

After completing the cluster model, then displaying a plot of the analysis results using the k-means algorithm, you can get the results of grouping dengue fever case data based on the UPT Puskesmas attributes, number of cases and year. Resulting in 6 clusters, the plot can be seen below:

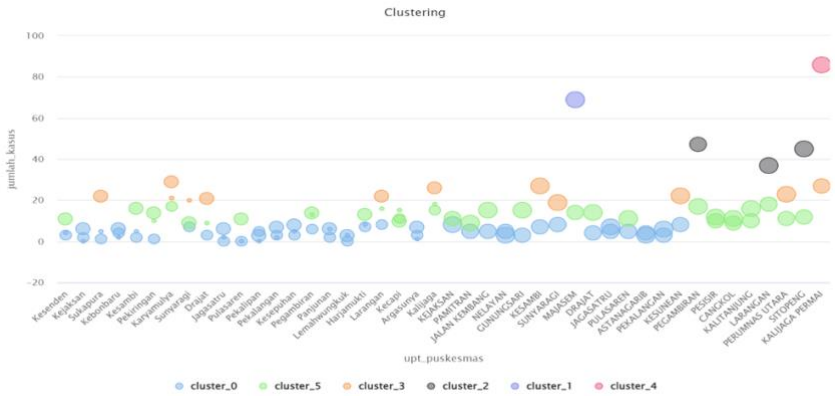

Fig. 10: Plot results of dengue cases 2020-2024

In the image above, it can be seen that the Scatter Plot image shows the distribution of the number of cases based on the location of health centers in Cirebon City.
With cluster 0 in blue having a case susceptibility of 0-8 cases of dengue fever in 2020-2024, cluster 1 in purple having the highest caseload of 69 cases of dengue fever which only occurred in 2024 alone, cluster 2 in gray having a case susceptibility of 37- 47 cases of dengue fever occurred in 2023-2024, cluster 3 in orange has a case susceptibility of 20-29 cases of dengue fever in 2023. 2020-2024, cluster 4 pink has 86 cases of dengue fever which only occurred in 2024 alone. Cluster 5 in green is susceptible to 9-18 cases of dengue fever each year.

5. Conclusion and Suggestions

5.1 conclusions

This study successfully applied the K-Means algorithm to classify dengue prone areas in Cirebon City, using puskesmas case data from 2020-2024. The analysis determined six optimal clusters with different risk levels: high (cluster 5), medium (clusters 0, 2, and 3), and low (clusters 1 and 4). This study provides insights for the Cirebon City Health Office to focus resources on high-risk areas and develop targeted prevention strategies.

5.2 Sugestions

Based on the results of research regarding the grouping of areas prone to dengue fever in Cirebon City using the K-Means algorithm, there are several suggestions that researchers will provide for further research, namely:

a. In further research, it is recommended to add data attributes such as population density, rainfall, environmental conditions, temperature, so that it can provide better results in identifying risk factors for cases of dengue fever.
b. To improve model quality, you can use grouping algorithms such as DBSCAN, Silhouette Score.
c. Visualization and Collaboration: Use geospatial maps or interactive dashboards to view results and work with the Health Service to create more effective disease prevention policies based on the analysis results.

References

[1] M. A. Sembiring, "PENERAPAN METODE ALGORITMA K-MEANS CLUSTERING UNTUK PEMETAAN PENYEBARAN PENYAKIT DEMAM BERDARAH DENGUE (DBD)," *J. Sci. Soc. Res.*, vol. 4, no. 3, p. 336, Oct. 2021, doi: 10.54314/jssr.v4i3.712.
[2] K. Gustipartsani, N. Rahaningsih, R. Danar Dana, and I. Yulia Mustafa, "Data Mining Clustering Menggunakan Algoritma K-Means Pada Data Kunjungan Wisatawan Di Kabupaten Karawang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3595–3601, 2024, doi: 10.36040/jati.v7i6.8282.
[3] N. Puspitasari, A. Ardin Maulana, Rosmasari, and F. Alameka, "K-Means untuk Klasterisasi Daerah Rawan Penyakit Demam Berdarah K-Means for Clustering of Dengue Fever Prone Areas," *J. Sisfotenika*, vol. 13, no. 1, pp. 40–52, 2023, doi: 10.30700/jst.v13i1.1337.
[4] M. Rivalda, E. M. Hidayat, M. A. Gunawan, and D. Defriyanto, "Penerapan Metode Clustering Dalam Upaya Pencegahan Penyakit Demam Berdarah Menggunakan Algoritma K-Means (Studi Kasus: Kota Tasikmalaya)," *J. Larik Ldng. Artik. Ilmu Komput.*, vol. 3, no. 1, pp. 1–10, Jul. 2023, doi: 10.31294/larik.v3i1.1774.
[5] S. H. Widiastuti and R. Jumardi, "Pengelompokan Daerah Rawan Demam Berdarah dengan Metode K-Means Clustering," *J. Inf. dan Teknol.*, vol. 4, no. 4, pp. 185–190, Aug. 2022, doi: 10.37034/jidt.v4i4.213.
[6] A. Apriliansyah Mohsa, P. Silva Rosiana, and Y. Umaidah, "Implementasi K-Means Dalam Pengelompokan Penyebaran Penyakit Dbd Di Jawa Barat," *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3, pp. 782–788, 2023, doi: 10.23960/jitet.v11i3.3344.
[7] N. Tri Hartanti, "Mengukur Tingkat Pemahaman Mahasiswa Pada Mata Kuliah Pemrograman dengan Algoritma K-Means Clustering Measuring Students' Level of Understanding in Programming Courses with the K-Means Clustering Algorithm," *J. Sisfotenika*, vol. 12, no. 1, pp. 62–73, 2022, doi: 10.30700/jst.v12i1.1210.
[8] J. Laurenso, D. Jiustian, F. Fernando, V. Suhandi, and T. Herlina Rochadiani, "Implementation of K-Means, Hierarchical, and BIRCH Clustering Algorithms to Determine Marketing Targets for Vape Sales in Indonesia," *J. Appl. Informatics Comput.*, vol. 8, no. 1, pp. 62–70, 2024, doi: 10.30871/jaic.v8i1.4871.
[9] A. Nugroho Sihananto, A. Puspita Sari, H. Khariono, R. Akhmad Fernanda, and D. Cakra Mudra Wijaya, "Implementasi Metode K-Means Untuk Pengelompokan Kasus Covid-19 Tingkat Provinsi Di Indonesia," *J. Inform. dan Sist. Inf.*, vol. 3, no. 1, pp. 76–85, Apr. 2022, doi: 10.33005/jifosi.v3i1.472.
[10] I. Tri Gustiane, M. Martanto, and T. Suprapti, "Clustering Hasil Cek Darah Diabetes Lansia Menggunakan Metode K-Means Di Posbindu Kp. Lebakjero Desa Ciherang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 2125–2129, 2024, doi: 10.36040/jati.v8i2.9281.