



Comparative Analysis of K-Means Clustering and K-Medoids Clustering Methods in Clustering Neonatal Infant Mortality Rates in West Java Province

Intan Putri S^{1*}, Nana Suarna², Agus Bahtiar³

¹²³ Teknik Informatika, STMIK IKMI CIREBON

nan.imut21@gmail.com^{1*}, st_nana@yahoo.com², agusbahtiaar038@gmail.com³

Abstract

Neonatal mortality rate is an important indicator in assessing public health conditions. This study aims to cluster neonatal mortality data in West Java Province using the *K-Means Clustering* and *K-Medoids Clustering* methods, as well as compare the performance of both methods in producing the best clusters. The study used secondary data obtained from Open Data West Java. The research stages included data selection, preprocessing, clustering, and evaluation using the *Davies-Bouldin Index* (DBI). The experiments were conducted using cluster variations (k) from 2 to 8. The results showed that the *K-Means Clustering* method produced the best performance with a DBI value of 0.430 at $k = 3$. The clustering results generated three categories: low-risk cluster with 408 data points, medium-risk cluster with 65 data points, and high-risk cluster with 13 data points. The differences in cluster characteristics indicate variations in neonatal mortality risk levels among regions in West Java Province. The findings of this study are expected to support decision-making and more targeted health policy planning.

Keywords: *K-Means Clustering*, *K-Medoids Clustering*, *Davies-Bouldin Index*, Neonatal Mortality.

1. Introduction

Neonatal mortality rate is an important indicator in assessing the quality of public health because it reflects maternal and infant health conditions as well as the effectiveness of healthcare services[1][2]. Neonatal mortality remains a global health problem influenced by various maternal, clinical, socioeconomic, and healthcare system factors. Maternal factors such as very young or advanced maternal age, anemia, hypertension, obesity, and a history of pregnancy complications significantly increase the risk of neonatal death. In addition, clinical conditions including preterm birth, low birth weight, congenital abnormalities, and neonatal infections are major causes of neonatal mortality[3][4]. Socioeconomic factors such as low maternal education, poverty, and limited access to healthcare services also contribute to poor neonatal health outcomes [5].

Efforts to reduce neonatal mortality require data-driven decision-making capable of identifying high-risk regions. Regional health data analysis using data mining techniques, particularly clustering methods, can be applied to group regions based on similarities in health indicators, mortality rates, demographic characteristics, and healthcare access. Clustering methods such as K-Means and K-Medoids can assist policymakers in identifying patterns of neonatal mortality distribution, enabling more targeted healthcare interventions[6].

Several studies have shown that clustering methods are effective in segmenting health regions into homogeneous groups to support public health policy planning [7]. K-Means is widely used because of its simplicity and computational efficiency in grouping data based on centroids. However, this method is sensitive to noise and outliers. In contrast, K-Medoids uses actual data objects as cluster centers, making it more stable and robust against extreme data values [8][9]. In the context of healthcare data, which often contains high variability, comparing these two methods is important to determine the most suitable clustering algorithm.

Furthermore, clustering-based healthcare data analysis can help identify hidden patterns in regional health data and classify regions with similar risk characteristics. This approach supports evidence-based decision-making, optimization of healthcare resource allocation, and improvement of public health intervention effectiveness[10][11][12]. Therefore, a comparative analysis of K-Means Clustering and K-Medoids Clustering in grouping neonatal mortality rates in West Java Province is important to identify the most effective clustering method for supporting regional health evaluation and policy planning.

2. Literature Review

2.1. Theoretical Basis

The theoretical foundation serves as the conceptual basis of a study to explain the variables, methods, and approaches used in the research. It functions as a reference in developing the research framework and assists researchers in understanding the research problem systematically and scientifically. In this study, the theoretical foundation includes concepts related to health data analysis, particularly the processing of neonatal mortality data using a data mining approach. The existence of a theoretical foundation provides a strong basis for explaining the investigated phenomena as well as the methods applied in the analysis process.

2.2. Data Mining

Data mining is a major component of data analysis that aims to extract meaningful patterns, relationships, and knowledge from large and complex datasets using statistical, machine learning, and computational techniques [10]. In the healthcare sector, data mining plays an important role in transforming raw data into valuable information to support evidence-based decision-making, disease prediction, patient stratification, and healthcare resource optimization[11][12]. One of the commonly used techniques in data mining is clustering, which belongs to unsupervised learning methods. Clustering is used to group data based on similarity characteristics without predefined labels, making it useful for identifying groups with similar risk profiles in healthcare data [13]. Methods such as K-Means and K-Medoids are widely applied in health data analysis because of their ability to reveal hidden structures within complex datasets. In this study, clustering techniques are applied to group regions based on neonatal mortality rates in order to identify high-risk areas and support more targeted healthcare policies. Therefore, data mining provides an effective approach for analyzing complex health data and generating valuable insights to support accurate and data-driven decision-making.

2.3. Clustering

Clustering is one of the techniques in data mining that belongs to the *unsupervised learning* method, which is used to group data without predefined labels or categories. The main objective of clustering is to classify data objects into several clusters based on similarity characteristics, so that objects within the same cluster have a higher level of similarity compared to those in different clusters [8]. Clustering plays an important role in revealing hidden patterns and structures within complex data, particularly high-dimensional healthcare data. This technique is used for pattern discovery, data simplification, and knowledge extraction from large and complex datasets[14]. In the healthcare field, clustering is widely applied to identify homogeneous subgroups within heterogeneous populations, such as patient stratification, disease pattern recognition, and regional segmentation based on specific health conditions[13][15]. In addition to pattern identification, clustering also functions as a data reduction technique by simplifying large datasets into several representative groups, thereby improving data interpretation and decision-making efficiency. Clustering algorithms such as K-Means and K-Medoids are widely used in healthcare data analysis because they can group data based on similarity without requiring predefined labels, making them suitable for exploratory analysis[8]. In this study, clustering techniques are used to group regions based on neonatal mortality rates in West Java Province. Through this grouping process, regions with similar characteristics can be identified, which may assist in determining high-risk areas and support more targeted healthcare policy planning.

2.4. K-Means Clustering & K-Medoids Clustering

K-Means and K-Medoids are partition-based clustering methods widely used in data analysis to group data based on similarity characteristics. K-Means uses the mean value (*centroid*) as the center of a cluster, while K-Medoids uses actual data objects (*medoids*) as cluster centers[8]. Both methods work iteratively by assigning data to the nearest cluster and updating the cluster center until stable clusters are formed. K-Means is known for its simplicity, fast computation, and efficiency in handling large datasets. However, it is sensitive to noise, outliers, and the selection of initial centroids. In contrast, K-Medoids is more robust to noise and extreme values because it uses real data points as cluster centers, making it more stable for complex and irregular datasets[14] [16]. Both algorithms are widely applied in healthcare data analysis because they can reveal hidden patterns and simplify complex data into meaningful groups. In this study, K-Means and K-Medoids are used to group regions based on neonatal mortality rates in West Java Province in order to identify high-risk areas and determine the clustering method that provides better performance and interpretation results.

2.5. Comparison of K-Means and K-Medoids

K-Means and K-Medoids are partition-based clustering methods used to group data based on similarity characteristics. The main difference between these methods lies in the cluster center used, where K-Means uses the mean value (*centroid*), while K-Medoids uses actual data objects (*medoids*) as the cluster center [8]. K-Means has advantages in computational speed and efficiency, making it suitable for large datasets. However, this method is sensitive to noise and outliers because it relies on average values as cluster centers. In contrast, K-Medoids is more stable and robust to noise and extreme values since it uses real data objects as cluster centers [9]. In addition, K-Means generally uses Euclidean distance and is more appropriate for data with simple distributions, whereas K-Medoids is more flexible in the use of distance measures, allowing it to handle more complex and irregular data distributions [16]. From an interpretability perspective, K-Medoids clustering results are easier to understand because the cluster centers represent actual data points within the dataset [13].

In this study, K-Means and K-Medoids are compared to determine the clustering method that produces the best cluster quality in grouping regions based on neonatal mortality rates in West Java Province.

2.6. Cluster Evaluation Using Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) is an internal evaluation method used to measure clustering quality based on the compactness of data within a cluster (*intra-cluster*) and the separation between clusters (*inter-cluster*) [8]. A smaller DBI value indicates better clustering quality because the clusters are more compact and clearly separated from one another [9]. DBI is widely used to compare the performance of clustering algorithms such as K-Means and K-Medoids and to determine the optimal number of clusters due to its computational efficiency and ease of interpretation [13][15]. In this study, DBI is applied to evaluate the clustering results of neonatal mortality data in West Java Province in order to identify the method that produces the best cluster quality. In addition, the application of data mining and machine learning in healthcare has been widely used for neonatal mortality analysis and prediction. Previous studies have shown that data-driven approaches can assist in identifying risk patterns, grouping high-risk regions, and supporting evidence-based decision-making in healthcare policy planning [17][18][19]. In this study, clustering methods are used to classify regions into high-, medium-, and low-risk categories based on neonatal mortality rates in West Java Province.

3. Research Methods

3.1. Research Design

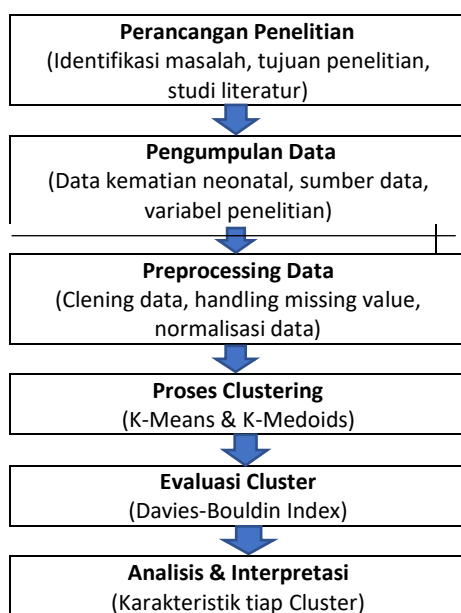


Fig.1: Types of Research Design

This research was conducted using a quantitative approach based on data mining, which focuses on processing and analyzing numerical data to find hidden patterns in the data.

3.2. Data Sources and Types

This study uses secondary quantitative data obtained from the official regional government portal, Open Data Jabar, through the dataset “*Jumlah Kematian Berdasarkan Jenis dan Penyebab Kematian di Jawa Barat*”.

link:

<https://opendata.jabarprov.go.id/id/dataset/jumlah-kematian-berdasarkan-jenis-dan-penyebab-kematian-di-jawa-barat>

The dataset contains information on mortality based on gender, cause of death, region, and time period in West Java Province. The data are used for clustering analysis using K-Means and K-Medoids methods to identify patterns of neonatal mortality across regions. In clustering analysis, data quality is important to produce accurate and representative results. Therefore, data preprocessing and normalization are conducted to improve clustering performance and support the identification of regional neonatal mortality risk patterns in West Java Province.

3.3. Data Analysis Technique

The data analysis technique in this study uses a clustering-based data mining approach to group regions based on neonatal mortality rates. The clustering methods applied are K-Means and K-Medoids, where K-Means is used for its computational efficiency, while K-Medoids is more robust to noise and outliers [8][9]. The analysis begins with descriptive analysis to understand the characteristics of the dataset,

followed by the clustering process using both methods. The clustering results are then evaluated using the Davies-Bouldin Index (DBI), where a smaller DBI value indicates better cluster quality [16][13]. Furthermore, the clustering results from K-Means and K-Medoids are compared to determine the best-performing method. The final stage involves interpreting the clusters to identify regional characteristics based on neonatal mortality rates, which can support healthcare decision-making and policy planning [15].

4. Results and Discussion

4.1. Implementation of Research Stages

This study applies a data mining approach using clustering techniques with K-Means and K-Medoids methods on neonatal mortality data in West Java Province. The research process includes data collection and preprocessing, clustering analysis, evaluation using the Davies-Bouldin Index (DBI), and interpretation of the clustering results. The implementation stages are carried out systematically to produce optimal regional clustering and provide an overview of neonatal mortality patterns in West Java Province.

4.1.1. Data Collection

The dataset used in this study is a dataset of neonatal mortality rates in West Java Province, consisting of 487 records and 10 attributes for the years 2017-2019. The dataset is sourced from Open Data Jabar (opendata.jabarprov.go.id). The search results are in the form of a soft file document in XLS format, as shown in Figure 4.1.

id	kode_provinsi	nama_provinsi	kode_kabup...	nama_kabu...	jenis_kemat...	penyebab_k...	jumlah_kem...	satuan	tahun
21	32	JAWA BARAT	3201	KABUPATEN ...	LAHIR MATI	ASFIKSIA	20	JIWA	2017
22	32	JAWA BARAT	3201	KABUPATEN ...	LAHIR MATI	BBLR	44	JIWA	2017
23	32	JAWA BARAT	3201	KABUPATEN ...	LAHIR MATI	KELAINAN	10	JIWA	2017
24	32	JAWA BARAT	3201	KABUPATEN ...	LAHIR MATI	LAIN-LAIN	16	JIWA	2017
25	32	JAWA BARAT	3201	KABUPATEN ...	LAHIR MATI	SEPSIS	5	JIWA	2017
26	32	JAWA BARAT	3201	KABUPATEN ...	LAHIR MATI	TETANUS	2	JIWA	2017
47	32	JAWA BARAT	3202	KABUPATEN ...	LAHIR MATI	ASFIKSIA	87	JIWA	2017
48	32	JAWA BARAT	3202	KABUPATEN ...	LAHIR MATI	BBLR	54	JIWA	2017
49	32	JAWA BARAT	3202	KABUPATEN ...	LAHIR MATI	KELAINAN	27	JIWA	2017
50	32	JAWA BARAT	3202	KABUPATEN ...	LAHIR MATI	LAIN-LAIN	27	JIWA	2017
51	32	JAWA BARAT	3202	KABUPATEN ...	LAHIR MATI	SEPSIS	2	JIWA	2017
52	32	JAWA BARAT	3202	KABUPATEN ...	LAHIR MATI	TETANUS	1	JIWA	2017
73	32	JAWA BARAT	3203	KABUPATEN ...	LAHIR MATI	ASFIKSIA	69	JIWA	2017
74	32	JAWA BARAT	3203	KABUPATEN ...	LAHIR MATI	BBLR	66	JIWA	2017

ExampleSet (486 examples, 0 special attributes, 10 regular attributes)

Fig.2: Infant Mortality Rate Dataset

In RapidMiner, the initial step in importing neonatal mortality data in West Java Province is an Excel file (xlsx). To enter the data into RapidMiner, use the "read excel" operator. The "read excel" function reads data in Excel format. Other operators, such as "Read CSV," "XML," and "URL," depend on the type of data format being imported into RapidMiner.

4.1.2. Understanding And Data Selection

In the data understanding and selection section, the collected dataset was reviewed to determine the structure and content of the data used in the study. To facilitate the analysis process, the data was first entered into the RapidMiner tool using the Read Excel operator. This process aims to display the data so that an understanding of the available attributes can be obtained, as shown in Figure 4.2.

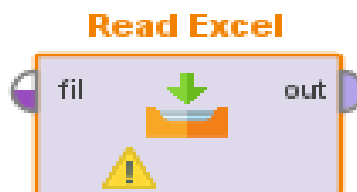


Fig.3: Excel Read Operator

After selecting the data to be entered using the Read Excel data import tool, the selected data will be displayed. The displayed data contains ten attributes: ID, district code, province name, district code, district name, type of death, cause of death, number of deaths, unit, and year, as shown in Figure 4.3.

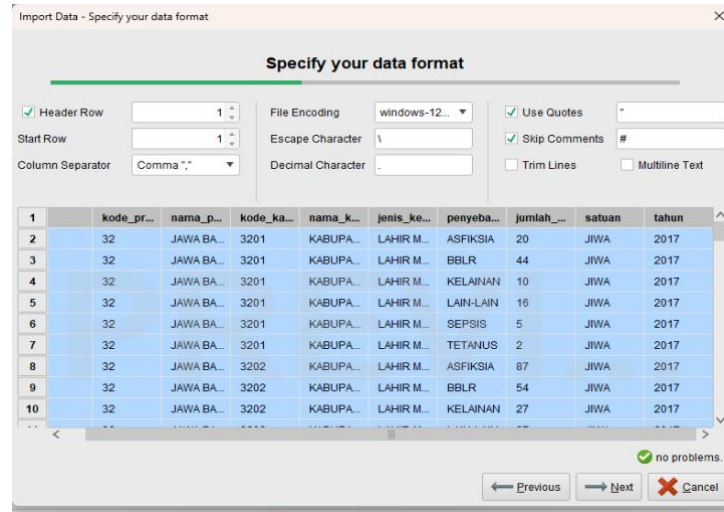


Fig.4: Data Import Results

After ensuring all attributes are correct, we simply click next and enter the next step. The Select Attributes operator is used to select only relevant attributes or columns to focus the analysis on variables that influence the grouping process, as shown in Figure 4.4.

Select Attributes



Fig.5: Select Attribute

The parameters used in the Select Attribute operator can be explained in Figure 4.5.

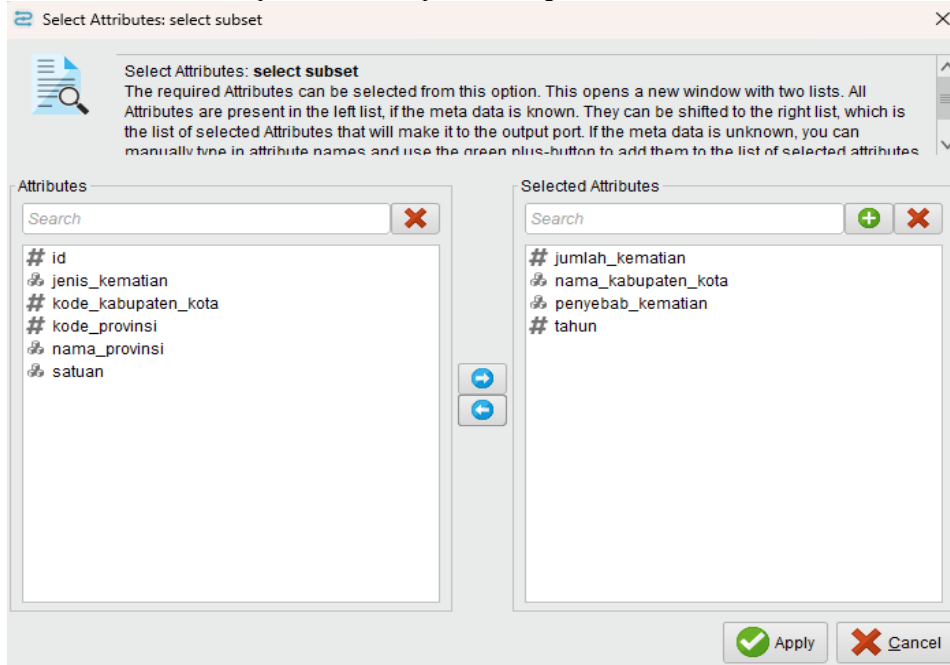


Fig.6: Selected Attributes Results

After ensuring all the selected attributes are correct, we simply click apply and enter the next stage, data preprocessing.

4.1.3. Preprocessing Data

In this section, there is a data cleaning process so that it can be processed by the K-Means and K-Medoids algorithms properly. The operators used in the data process are to change the data type to suit and what is changed are the sub-district and gender attributes using Nominal to Numerical, as in Figure 4.6.

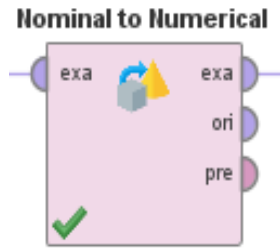


Fig.7: Attribute becomes numeric

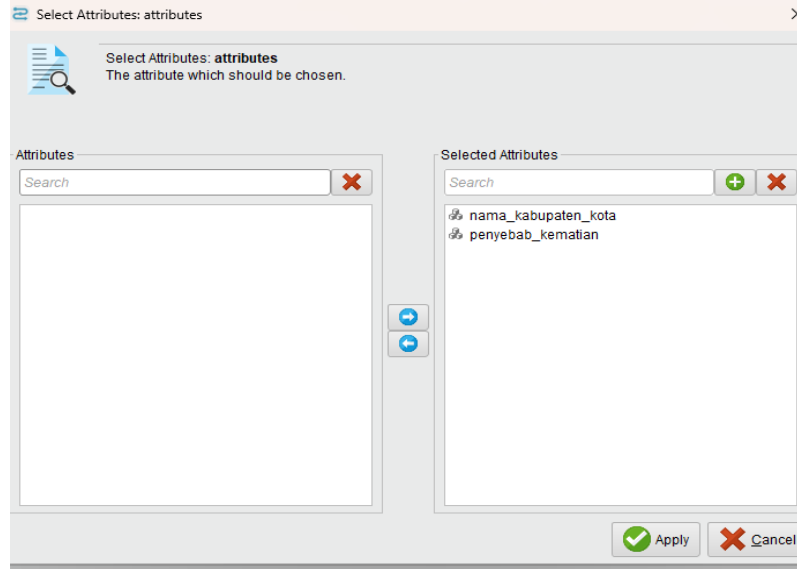


Fig.8: Select Attributes

For the Nominal to Numerical attribute, in the parameter in the attribute filter type column, select subset. For the attributes column, select select attributes, and in this column, apply district_city_name and cause_of_death. For the coding type column, select dummy coding. As shown in Table 4.1.

Tabel 1: Parameter

Parameters	
Attribute filter type	Subset
Attributes	Select attribut 1. Nama_kabupaten_kota 2. Penyebab_kematian
Coding type	Dummy coding

The results of applying nominal attributes to numeric attributes using the “Nominal to Numerical” operator can be seen in Figure 4.8.

Name	Type	Missing	Statistics	Filter (35 / 35 attributes):
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	Search for Attributes
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	
nama_kabupaten_kota = KABU...	Integer	0	Min: 0, Max: 1, Average: 0.037	

Fig.9: Nominal to Numeric operator results

4.1.4. Clustering Process

In this section, the clustering process is performed to form data groups based on the similarity of characteristics shared by each data item. The clustering process uses two methods, K-Means and K-Medoids, with the aim of comparing the obtained clustering results. This section also uses the Cluster Distance Performance operator to directly measure the quality of the clustering results. This operator produces evaluation values such as the Davies-Bouldin Index (DBI) or Average Within Centroid Distance, which indicate the degree of closeness between data within a cluster and the degree of separation between clusters. As shown in Figure 10

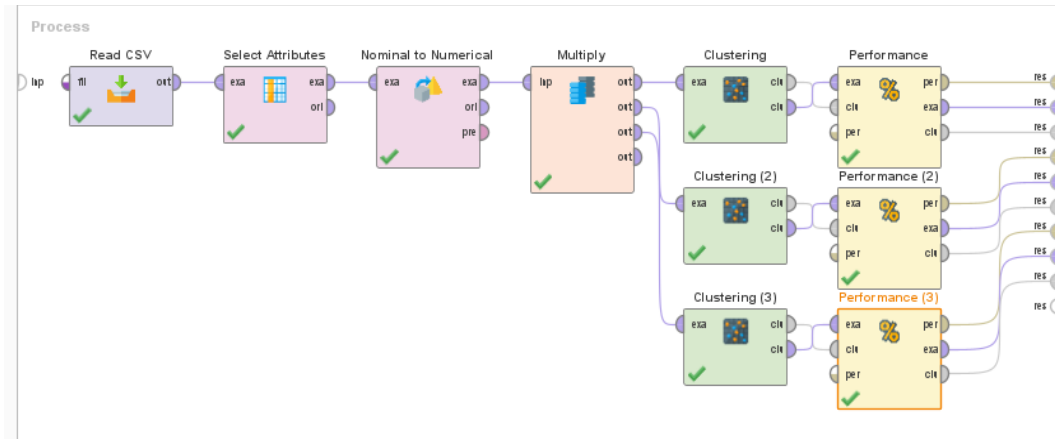


Fig.10: Display of the application of the Clustering methods

Using the Cluster Distance Performance operator, a test is performed to find the best results from 2, 3, 4, 5, 6, 7, and 8 clusters tested. The Cluster Distance Performance operator is used to assess the DBI results for each tested cluster. The following displays the cluster test results in Figure 4.10.

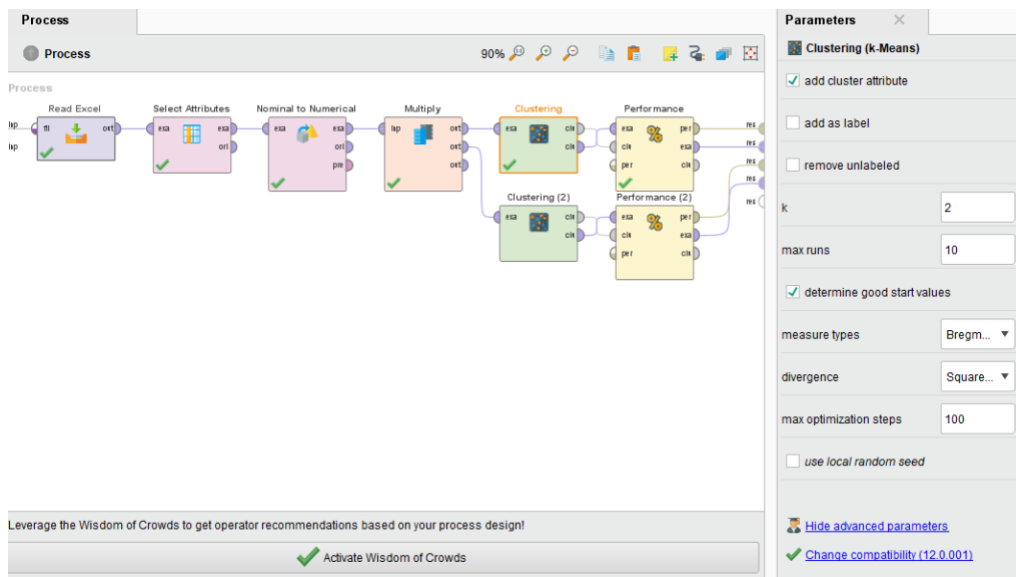


Fig.11: K-Means and K-Medoids Algorithms with K2

Figure 4.10 above shows the process of calculating the Davies Bouldin Index value. The calculated K2 values for the K-Means and K-Medoids clusters will yield the Davies Bouldin Index value. The next step is shown in Figure 4.11.

```

PerformanceVector:
Avg. within centroid distance: 139.966
Avg. within centroid distance_cluster_0: 82.588
Avg. within centroid distance_cluster_1: 505.100
Davies Bouldin: 0.482
    
```

Fig.12: DBI K-Means K2 results

In Figure 4.11 is the calculation of the value results using DBI K-Means with K2, namely the result is 0.482, then the results of the DBI K-Medoids K2 calculation can be seen in Figure 4.12.

PerformanceVector

```

PerformanceVector:
Avg. within centroid distance: 320.333
Avg. within centroid distance_cluster_0: 26.683
Avg. within centroid distance_cluster_1: 712.808
Davies Bouldin: 1.019
  
```

Fig.13: DBI value results of K-Medoids K2

Figure 4.12 shows the calculation of the value results using DBI K-Medoids with K2, namely the result is 1.019. Next, there is Figure 4.13.

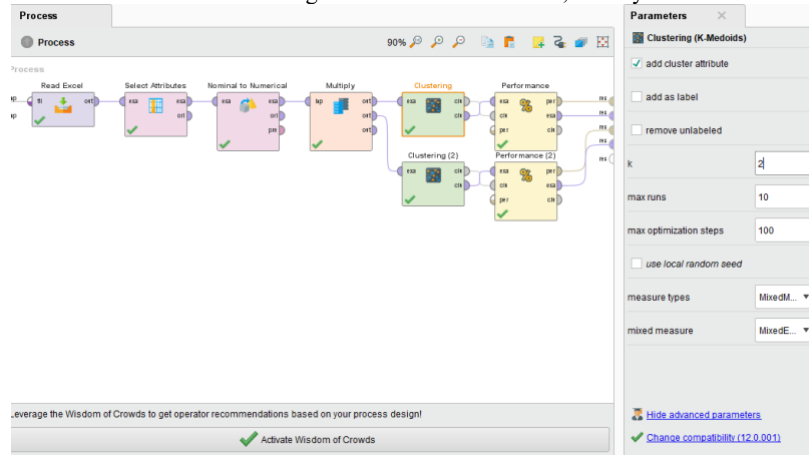


Fig.14: K-Means and K-Medoids Algorithm with K8

Figure 4.13 above shows the process of calculating the Davies Bouldin Index value. The K8 value in the K-Means and K-Medoids clusters is calculated to determine the resulting Davies Bouldin Index value. The next step is shown in Figure 4.14.

PerformanceVector

```

PerformanceVector:
Avg. within centroid distance: 10.733
Avg. within centroid distance_cluster_0: 9.654
Avg. within centroid distance_cluster_1: 29.422
Avg. within centroid distance_cluster_2: 15.294
Avg. within centroid distance_cluster_3: 6.391
Avg. within centroid distance_cluster_4: 24.587
Avg. within centroid distance_cluster_5: 17.095
Avg. within centroid distance_cluster_6: 16.313
Avg. within centroid distance_cluster_7: 16.444
Davies Bouldin: 0.501
  
```

Fig.15: Dbi K-Means K8 Results

Figure 4.14 shows the calculation of the value results using DBI K-Means with K8, namely the result is 0.501, then the results of the DBI K-Medoids K8 calculation can be seen in Figure 4.15.

PerformanceVector

```

PerformanceVector:
Avg. within centroid distance: 33.689
Avg. within centroid distance_cluster_0: 5.659
Avg. within centroid distance_cluster_1: 4.603
Avg. within centroid distance_cluster_2: 605.786
Avg. within centroid distance_cluster_3: 25.263
Avg. within centroid distance_cluster_4: 17.113
Avg. within centroid distance_cluster_5: 44.905
Avg. within centroid distance_cluster_6: 5.608
Avg. within centroid distance_cluster_7: 43.487
Davies Bouldin: 1.586

```

Fig.16: DBI value results of K-Medoids with K8

Figure 4.15 shows the calculation of the value results using DBI K-Medoids with K8, namely the result is 1,586.

4.1.5. Clustering Evaluation Result

This stage displays the comparison results of K-Means and K-Medoids, including information on the number of clusters and the DBI values generated by the K-Means and K-Medoids algorithms. To measure the effectiveness of the clustering obtained by each algorithm, the DBI value is used for evaluation. The following details the comparison results as shown in Table 4.2.

Tabel 2: DBI K-Means dan K-Medoids

Jumlah Cluster	DBI Algoritma	
	K-Means	K-Medoids
2	0.482	1.019
3	0.430	0.574
4	0.472	0.610
5	0.517	0.707
6	0.502	1.293
7	0.472	1.521
8	0.501	1.586

Based on the results of Table 4.2, a DBI value of 0.430 was found using the K-Means algorithm, with a total of 3 clusters indicating that this value is the best for grouping, because the DBI value is smaller than the results of other K-Means DBIs. Thus, it can be concluded that the best number of clusters from the comparison of K-Means and K-Medoids for grouping neonatal mortality rates in West Java province is 3 clusters as seen in Figure 4.16.

Cluster Model

```

Cluster 0: 408 items
Cluster 1: 13 items
Cluster 2: 65 items
Total number of items: 486

```

Fig.17: Smallest Cluster Model K-Means

Based on Figure 4.16, the results of the K3 cluster model from K-Means show that Cluster 0 has a total of 408 items, Cluster 1 contains 13 items, and Cluster 2 contains 65 items. So the total number of items reaches 486.

4.1.6. Analysis and Interpretation of Results

Based on the results of the clustering and evaluation process, the best method was K-Means, with a Davies-Bouldin Index (DBI) of 0.430 for the number of clusters (k) = 3. The clustering results show that the data is divided into several clusters with different characteristics. Each cluster represents a different neonatal mortality rate in each region. The first cluster indicates a group of regions with low neonatal

mortality rates, the second cluster shows moderate rates, and the third cluster shows high rates. These differences indicate variations in conditions between regions in West Java Province. These results indicate that the clustering method is able to effectively group data based on similar characteristics. This clustering allows us to identify regions that require more attention in efforts to manage and prevent neonatal mortality. Therefore, the results of this study can be used as a basis for more targeted decision-making and health policy planning.

4.2. Discussion

This study compares the K-Means and K-Medoids methods for clustering neonatal mortality rates in West Java Province through data mining stages, from preprocessing to evaluation. Both methods were tested with varying numbers of clusters ($k = 2-8$) and evaluated using the Davies-Bouldin Index (DBI), where a lower value indicates better clustering quality. The evaluation results showed that the K-Means method produced the lowest DBI value with an optimal number of three clusters ($k = 3$). The clustering results divided regions into three categories: low-risk (408 regions), moderate-risk (65 regions), and high-risk (13 regions). This distribution indicates disparities in neonatal mortality rates between regions. High-risk clusters indicate the need to improve access to and quality of maternal and neonatal health services, while moderate-risk clusters require strengthening preventive interventions. Low-risk clusters still require ongoing monitoring to maintain existing achievements.

5. Conclusion and Recommendations

This study demonstrates the success of the K-Means and K-Medoids methods in clustering neonatal mortality rates in West Java Province through systematic data mining. Based on an evaluation using the Davies-Bouldin Index (DBI), the K-Means method with an optimal number of three clusters ($k = 3$) yielded the best performance, with the lowest DBI value (0.430). The clustering results divided regions into three categories: high, medium, and low risk, indicating differences in neonatal mortality risk levels between regions. These findings confirm that the clustering method is effective in identifying data distribution patterns and can be used as a basis for prioritizing region-based health policies and interventions. The results of this study can inform regional-based health policymaking. The West Java Provincial Government and Health Office are advised to prioritize areas in high-risk clusters by strengthening maternal and neonatal health services, including improving facilities, health workers, and coverage of prenatal and newborn care. Low-risk areas require regular monitoring to maintain existing achievements, while moderate-risk areas must maintain and enhance existing health programs to prevent future increases in risk.

References

- [1] D. Tembo, F. D. N. Abobo, P. Kaonga, and et al., "Risk factors associated with neonatal mortality among neonates admitted to neonatal intensive care unit of the University Teaching Hospital in Lusaka," *Sci. Rep.*, vol. 14, p. 5231, 2024, doi: 10.1038/s41598-024-56020-6.
- [2] M. Nabila, A. Baidani, Y. Mourajid, M. Chebabe, and H. Abderraouf, "Analysis of risk determinants of neonatal mortality in the last decade: A systematic literature review (2013–2023)," *Pediatr. Rep.*, vol. 16, no. 3, pp. 696–716, 2024, doi: 10.3390/pediatric16030059.
- [3] A. Abu Sharkh et al., "Factors contributing to neonatal mortality in a tertiary center in south West Bank: A single-center retrospective cohort study," *PLOS Glob. Public Heal.*, vol. 5, no. 7, p. e0004885, 2025, doi: 10.1371/journal.pgph.0004885.
- [4] K. Taye, Y. Kebede, D. Tsegaw, and W. Ketema, "Predictors of neonatal mortality among neonates admitted to the neonatal intensive care unit at Hawassa University Comprehensive Specialized Hospital," *BMC Pediatr.*, vol. 24, p. 237, 2024, doi: 10.1186/s12887-024-04689-z.
- [5] OECD and WHO, *Neonatal mortality*. OECD Publishing, 2024. doi: 10.1787/51fed7e9-en.
- [6] E. Sabna and M. Mitra, "Analysis of K-mean clustering in grouping child mortality rate in Indonesia," *J. Intell. Syst. Softw. Eng.*, vol. 1, no. 1, pp. 27–31, 2024, doi: 10.25311/joisse/Vol1.Iss1.1813.
- [7] B. Nafandra, D. Sulistiowati, and N. Amalita, "Comparison of K-means and DBSCAN algorithms in regional segmentation based on mortality and fertility indicators," *J. Mat. dan Stat. serta Apl.*, vol. 14, no. 1, 2025, doi: 10.24252/msa.v14i1.61961.
- [8] D. Jing, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny)*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.
- [9] X. Zhou, "A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective," *Symmetry (Basel)*, vol. 15, no. 9, p. 1679, 2022, doi: 10.3390/sym15091679.
- [10] M. Sharma and W. Verbeke, "Improving diagnosis and prognosis of healthcare using data mining techniques: A review," *Comput. Biol. Med.*, vol. 125, p. 103960, 2020, doi: 10.1016/j.compbiomed.2020.103960.
- [11] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, 2020, doi: 10.1056/NEJMr1814259.
- [12] F. Jiang et al., "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 6, no. 2, pp. 230–243, 2021, doi: 10.1136/svn-2020-000454.
- [13] A. Aljohani, "Optimizing patient stratification in healthcare: A comparative analysis of clustering algorithms for EHR data," *Int. J. Comput. Intell. Syst.*, vol. 17, p. 173, 2024, doi: 10.1007/s44196-024-00568-8.
- [14] Z. Li, H. Zhou, Z. Xu, and Q. Ma, "Machine learning and public health policy evaluation: Research dynamics and prospects for challenges," *Front. Public Heal.*, vol. 13, p. 1502599, 2025, doi: 10.3389/fpubh.2025.1502599.
- [15] Y. Hu, H. Yan, and M. Liu, "Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records," *BMC Med. Res. Methodol.*, vol. 24, p. 309, 2024, doi: 10.1186/s12874-024-02422-z.
- [16] A. A. Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Comput. Sci.*, vol. 10, p. e2286, 2024.
- [17] H. J. Do, K. M. Moon, H.-S. Jin, J. Y. Kwon, S.-Y. Moon, and E.-K. Kim, "Machine learning models for predicting mortality in 7,472 very low birth weight infants using data from a nationwide neonatal network," *Diagnostics*, vol. 12, no. 3, p. 625, 2022, doi: 10.3390/diagnostics12030625.
- [18] A. Sheikhtaheri, M. R. Zarkesh, R. Moradi, and F. Kermani, "Prediction of neonatal deaths in NICUs: Development and validation of machine learning models," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 131, 2021, doi: 10.1186/s12911-021-01497-8.
- [19] I. B. Mboya, M. J. Mahande, M. Mohammed, J. Obure, and H. G. Mwambi, "Prediction of perinatal death using machine learning models: A birth registry-based cohort study in northern Tanzania," *BMJ Open*, vol. 10, no. 10, p. e040132, 2020, doi: 10.1136/bmjopen-2020-040132.