CLUSTERING HIV/AIDS DISEASE USING K-MEANS CLUSTERING ALGORITHM

Indah Purnama Sari[1*]
Pipit Putri Hariani MD[2]
Al-Khowarizmi[3]
Fanny Ramadhani[4]
Oris Krianto Sulaiman[5]
Andy Satria[6]
Asrar Aspia Manurung[7]

[*1, 2, 3, 7]Universitas Muhammadiyah Sumatera Utara
[4]Universitas Negeri Medan
[5]Universitas Islam Sumatera Utara
[6]Universitas Dharmawangsa
[*1]*email:* indahpurnama@umsu.ac.id

Abstract: The HIV (Human Immunodeficiency Virus) is an infection-causing virus that targets the immune system, making it more vulnerable to illness and infection. East Java (33,043), Papua (25,586), West Java (24,650), and Central Java (18,038 persons) all had more HIV/AIDS infections in 2017 than DKI Jakarta (46,378), which was followed by East Java (33,043), West Java (24,650), and a smaller number—25,586—in Papua. Nationally, West Java is still among the four provinces with the highest HIV/AIDS cases. This shows that HIV/AIDS has become a threat to the wider community because in addition to threatening the lives of sufferers, this disease is at risk of transmission that will increase. The increase in HIV/AIDS cases can be a problem in the psychology of the sufferer, because this disease can have a negative impact in the form of physical, psychological, social and spiritual problems that cause PLWHA (People with HIV AIDS) to live a stressful life. In this study, calculations were carried out using the K-Means algorithm with the Optimize Parameters Grid on data on the spread of HIV / AIDS cases in 2019-2021 sourced from the West Java Provincial Health Office. K-Means is one of the algorithms in data mining that can be used for grouping / clustering of data. The data used in this study were 971 records. The purpose of this study was conducted to determine the cluster of t h e spread of HIV/ AIDS as an effort to assist the government in reducing the number of HIV / AIDS cases in West Java province. The results o f this study are comparing DBI with the K-Means method from k-2 to k-20 contained in the table above, it can be seen that the cluster that is close to 0 is k-2, with a DBI value of 0.414. Because the value of k-2 is the smallest value compared to other k, it can be concluded that k-2 with a value of 0.414 which is closest to 0 is the best cluster result.

Keywords: Clustering, HIV/AIDS, K-Means algorithm

Introduction

HIV/AIDS cases continue to increase every year, in 2022 it was reported that the cumulative number of HIV/AIDS sufferers in Indonesia amounted to 242,699 people. The prevalence of HIV/AIDS cases is still a health problem in Indonesia. Nationally, West Java is still among the five provinces with the highest HIV/AIDS cases [2]. HIV (Human Immunodeficiency Virus) is a virus that attacks the immune system which can weaken the body's ability to fight infection and

disease. While AIDS is a condition where HIV is already in the final stage of infection. At this stage the body's ability to fight incoming infections is very low making it very vulnerable to various infections (Krisdayanti, et al, 2019).
The K-Means algorithm is the technique used to cluster HIV/AIDS cases in West Java. Due to its ease of use and effectiveness in splitting data into various groups, the K-Means algorithm is one of the clustering algorithms that is frequently employed. An approach that can reduce the distance between data and its cluster is the K-Means algorithm. Because of its ease of use and effectiveness in combining data with enormous volumes, clustering K-Means is a popular method and is most frequently employed in data processing. One non-hierarchical clustering technique is K-Means, which aims to divide existing data into one or more clusters. (Salami, et al, 2021).
In an effort to aid the government in lowering the number of HIV/AIDS cases in the province of West Java, the study's goal was to identify the cluster of the spread of HIV/AIDS. In order to assist the government in reducing the number of HIV/AIDS cases in West Java, the goal of this research is to identify the cluster of HIV/AIDS cases in West Java based on age characteristics.

Literature Review

Data Mining

Data Mining is the process of finding meaningful new correlations, patterns and trends by sorting through large amounts of data stored in repositories, using pattern recognition technology as well as statistical and mathematical techniques [6]. One method that occurs in data mining is clustering. Clustering is a data analysis method that can be used to solve problems in a data grouping. Data mining is often also called knowledge discovery in databases (KDD). KDD is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets. Data Mining is the core of the knowledge Discovery in Database (KDD) process which involves algorithms to explore data, develop models and find previously unknown patterns, also known as pattern recognition which is used to find hidden patterns from processed data. (Ramadhani, et al, 2023).

Knowladge in Database (KDD)

Knowledge in Database (KDD) is a method used to obtain knowledge originating from existing databases. The results of the knowledge obtained can be used as a knowledge base that is used for decision making purposes [4]. Knowledge in Database (KDD) is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets. Data Mining is the core of the Knowledge Discovery in Database (KDD) process which involves algorithms for exploring data, developing models and finding previously unknown patterns is also known as pattern recognition which is used to find hidden patterns from processed data. (Indra, et al, 2021).

Clustering

Clustering is the process of dividing data in a set into several groups where the similarity of the data in one group is greater than the similarity of the data with data in other groups. Basically, clustering is a method for finding and grouping data that has similar characteristics (similarity) between one data and another. other data. (Sari, et al, 2019).

K-Means Clustering Algorithm

The K-Means algorithm is an iterative clustering algorithm that partitions a data set into a number of k clusters that have been determined at the beginning [4]. The K-Means algorithm has quite

high accuracy regarding object size, so this algorithm is relatively more scalable and efficient for processing large numbers of objects. Apart from that, the K-Means algorithm is not affected by the order of objects. One of the important stages in implementing K-Means Cluster is determining the centroid, number of clusters and centroid distance. (Sari, et al, 2019).

Method

A. K-means Clustering Algorithm

Knowledge extracted from existing databases is obtained using the Knowladge in Database (KDD) approach. The outcomes of the knowledge acquired may be used to create a knowledge foundation for use in making decisions [4]. Knowledge in Databases (KDD) is the process of gathering and using historical data to look for patterns, regularities, or correlations in huge data sets. Data mining, also known as pattern recognition, is the central component of the knowledge discovery in databases (KDD) process. It uses algorithms to examine data, create models, and identify previously undiscovered patterns.

The K-Means approach divides the dataset into a preset number of k clusters using iterative clustering [4]. The K-Means technique is comparatively more scalable and effective for processing a large number of objects because it has a reasonably high precision for the object's size. The order of the objects has no impact on the K-Means method either. Finding the centroid, the number of clusters, and the centroid distance is one of the crucial steps in implementing K-Means Cluster [8]. Sustainable tourism as defined by The World Tourism Organization (UNWTO) is tourism that takes full account of current and future economic, social and environmental impacts.

B. General Architecture

A methodology and research limits are required for the paper's direction. so that a universal architecture can be constructed to prevent writing errors. In Fig. 1, the overall architecture is displayed.

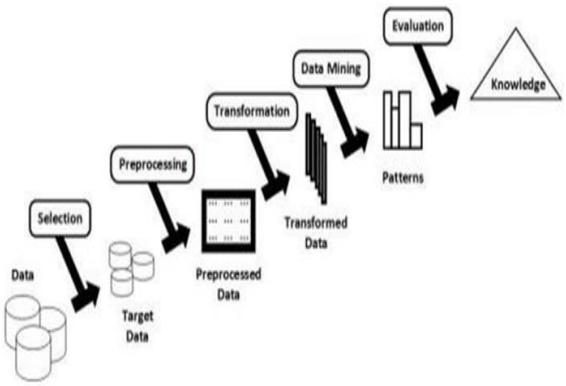

Figure 1 : Stages of the KDD Process

In this design stage using the Knowledge Discovery in Database (KDD) stage, the KDD stages are as follows:

1. Data Selection
   At this stage, the process of selecting which attributes will be used in the data mining process is carried out. Of the 10 attributes, namely Id, City District Name, Age Group, Gender,

Number of Cases, Year, Unit, Province Code, Province Name, City District Code. These attributes will be selected into 6 attributes, namely Id, City District Name, Age Group, Gender, Number of Cases, Year.

2. Preprocessing
   Preprocessing is the process of cleaning data from missing values or noise, namely irrelevant or inconsistent data. In this preprocessing stage, the data obtained will be cleaned of errors, missing, or incomplete data criteria.
3. Data Transformation
   Data Transformation is the process of changing the data type into numeric data type, so that the data can be processed using the K-Means algorithm.
4. Data Mining
   This data mining stage is done by applying an algorithm or knowledge search. In this research, the algorithm used is the k-means clustering algorithm.
5. Interpretation/Evaluation
   At the interpretation or evaluation stage, it is done by analyzing the results of the experiments that have been carried out.

Result and Discussion

Using the Rapidminer tool, the k-means clustering algorithm application is created. The phases of Knowledge Discovery in Databases (KDD) were used in this study. The following are some examples of how KDD was used in this study :

A. Data Selection

Read Excel, in the RapidMiner application the Read Excel feature functions as an excel file reader. This operator is used to import or enter excel data on the user's computer into the rapidminer process. Read Excel is the most basic operator used before starting a process. This operator can be used to load data from the Microsoft Excel spreadsheet. The dataset used is a dataset of HIV/AIDS cases in West Java.

Figure 2 : Excel Read Operator

The parameters in the Read Excel operator use the default parameters.
Set Role, this feature serves to distinguish the naming line of coordinate attributes and position predictions that will be entered into the 'label' category. So that when categorizing the 'label' data does not participate in the calculation and change the results.

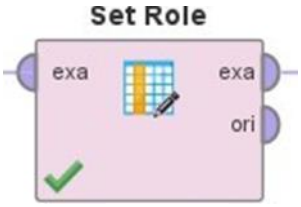

Figure 3 : Operator Set Role

The parameters of the Set Role operator used are shown in the table below.

Table 1. Second File Data Set Statistic

No	Parameters	Contents
1	Attribute Filter Type	Subset
2	Selected Attributes	id, gender, number of cases, age group, district name

Select Attributes, this operator is used to filter what data will be used in data processing. The attributes in the HIV/AIDS case dataset are filtered from 10 attributes to 6 attributes used. The attributes used are id, kelmin type, number of cases, age group, city district name, year.

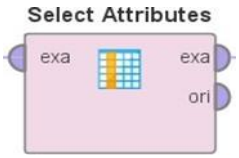

Figure 4 : Select Attributes Feature

B. Preprocessing

Since there are no missing values or data that do not have a value in the HIV/AIDS case dataset, there are no missing values. Preprocessing is not performed.

Figure 5 : Statistical Results

C. Data Transformation

Nominal to Polynominal, used to convert non-numeric attribute types into polynominal types. In the HIV/AIDS case dataset, the year attribute becomes polynominal.

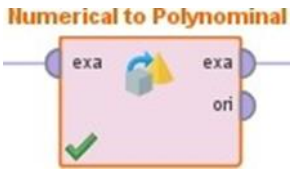

Figure 6 : Nominal to Polynominal

Nominal to Numerical, used to convert non-numeric attribute types to numeric types. In the HIV/AIDS Case dataset the attributes gender, age group, city district name, year are converted to numeric.

Figure 7 : Nominal o Numerical

D. Data Mining

In this stage the algorithm used is k-means clustering.

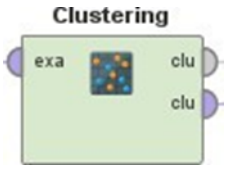

Figure 8 : K-Means Clustering

The parameters of the K-Means Clustering operator used are shown in the table below. Parameters of the operator This K-Means Clustering uses default parameters.

Table 2 : Parameters of K-Means Clustering Preator

No	Parameters	Contents
1	K	2-20

From the reading results of the K-Means Clustering operator with default parameters, the following information is obtained.

Table 3 : K-Means Clustering Results

Itertion	Clusters	Measure Type	Main Criterion	Davies Bouldin
1	2	Euclidean Distance	Davies Bouldin	0.414
2	3	Euclidean Distance	Davies Bouldin	0.570
3	4	Euclidean Distance	Davies Bouldin	0.511
4	5	Euclidean Distance	Davies Bouldin	0.615
5	6	Euclidean Distance	Davies Bouldin	0.614
6	7	Euclidean Distance	Davies Bouldin	0.703

7	8	Euclidean Distance	Davies Bouldin	0.742
8	9	Euclidean Distance	Davies Bouldin	0.804
9	10	Euclidean Distance	Davies Bouldin	0.813
10	11	Euclidean Distance	Davies Bouldin	0.798
11	12	Euclidean Distance	Davies Bouldin	0.806
12	13	Euclidean Distance	Davies Bouldin	0.842
13	14	Euclidean Distance	Davies Bouldin	0.831
14	15	Euclidean Distance	Davies Bouldin	0.844
15	16	Euclidean Distance	Davies Bouldin	0.844
16	17	Euclidean Distance	Davies Bouldin	0.912
17	18	Euclidean Distance	Davies Bouldin	0.927
18	19	Euclidean Distance	Davies Bouldin	0.830
19	20	Euclidean Distance	Davies Bouldin	0.854

Furthermore, the Performance operator is used with the Davies-Bouldin Index (DBI) method. To find out the DBI value her.

Figure 9 : Cluster Distance Performance

From the reading results of the Performance operator, the following information is obtained.

No	K	DBI
1	2	0.414
2	3	0.570
3	4	0.511
4	5	0.615
5	6	0.614
6	7	0.703
7	8	0.742
8	9	0.804

9	10	0.813
10	11	0.798
11	12	0.806
12	13	0.842
13	14	0.831
14	15	0.844
15	16	0.844
16	17	0.912
17	18	0.927
18	19	0.830
19	20	0.854

E. Interpretation / Evaluation

After comparing DBI with the K-Means method from k-2 to k-20 in the table above, it can be seen that the cluster closest to 0 is k-2, with a DBI value of 0.414. Because the value of k-2 is the smallest value compared to other k, it can be concluded that k-2 with a value of 0.414 which is closest to 0 is the best cluster result.

Conclusion

Utilizing the K- Means Clustering technique, data testing utilizing RapidMiner tools on HIV/AIDS illness clustering in West Java was performed. The cluster that is closest to 0 is k-2, with a DBI value of 0.414, as determined using the K-Means method, which produced the DBI value from k-2 to k-20 in the table above. Because the value of k-2 is the smallest value compared to other k, it can be concluded that k-2 with a value of 0.414 which is closest to 0 is the best cluster result.

References

E. Krisdayanti and J. I. Hutasoit, "THE EFFECT OF COPING STRATEGIES ON MENTAL HEALTH AND QUALITY OF LIFE OF POSITIVE HIV/AIDS PATIENTS," 2019.

S. Salami et al., "Qualitative Study of Coping Strategies of HIV AIDS Patients in Bandung City," *Faletehan Health Journal*, vol. 8, no. 1, pp. 22-30, 2021, [Online]. Available at: www.journal.lppm- stikesfa.ac.id/ojs/index.php/FHJ

R. Indra, L. Sinaga, W. Saputra, H. Qurniawan, and S. Tunas Bangsa, "Grouping the Number of Aids Disease Cases by Province Using the K-Means Method," 2021. [Online]. Available at: https://pusdatin.kemkes.go.id

Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H (2021). "Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process." *Journal of Computer Science, Information Technology and Telecommunication Engineering*, 2(1), 139-144.

G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Application of K-Means Algorithm for Drug Data Clustering," *National Journal of Information Technology and Systems*, vol. 5, no. 1, pp. 17-24, Apr 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H (2021). "Optimization of the FP-Growth Algorithm in Data Mining Techniques to Get the Electric Power Theft Pattern for the Development of Smart City." *4th International Conference of Computer and Informatics Engineering (IC2IE)*, 293-298.

Ramadhani, F., Satria, A., & Sari, I.P (2023). "Implementasi Metode Fuzzy K-Nearest Neighbor dalam Klasifikasi Penyakit Demam Berdarah." *Hello World Jurnal Ilmu Komputer* 2(2), 58-62.

Rohani M.M., & Yusoff, A. S. (2015). Tahap Kesediaan Pelajar Dalam Penggunaan Teknologi, Pedagogi, Dan Kandungan (TPACK) Dalam Pembelajaran Kurikulum di IPT. *Proceeding of the 3rd International Conference on Artificial Intelligence and Computer Science*, Pulau Pinang.

Sari, I.P., Batubara, I.H., & Al-Khowarizmi, A (2021). "Sensitivity Of Obtaining Errors In The Combination Of Fuzzy And Neural Networks For Conducting Student Assessment On E-Learning." *International Journal of Economic, Technology and Social Sciences (Injects)*, 2(1), 331-338.

Sari, I.P., Fahroza, M.F., Mufit, M.I., & Qathrunad, I.F (2021). "Implementation of Dijkstra's Algorithm to Determine the Shortest Route in a City." *Journal of Computer Science, Information Technology and Telecommunication Engineering*, 2(1), 134-138.

Ramadhani, F., Al-Khowarizmi, A., & Sari, I.P (2021). "Improving the Performance of Naïve Bayes Algorithm by Reducing the Attributes of Dataset Using Gain Ratio and Adaboost." *2021 International Conference on Computer Science and Engineering (IC2SE).* Vol. 1, pp.1-5.

Abbas A K. 2006-2007,"Congenital and Acquired Immuno deficiencies." In: Basic Immunology: Function and Disorders of the Immune System.

Rosella, M. 2013. "Faktor-faktor yang berpengaruh terhadap harapan hidup 5 tahun pasien Human Immundeficiency Virus (HIV) / Acquired Immune Deficiency Syndrome (AIDS) di RSUP DR. Kariadi Semarang," *Karya Tulis Ilmiah, Universitas Diponegoro, Semarang.*

Triharti, S. A. 2016. "Identifikasi Penyakit Tuberkulosis (TB) pada manusia menggunakan metode Naive Bayes," Skripsi, Universitas Sanata Dharma, Yogyakarta.

Hanum, S. Y. M., 2009. "Hubungan Kadar CD4 dengan Infeksi Jamur Superfisialis pada penderita HIV di RSUP H. Adam Malik Medan."

Marwati, L. 2016. "Aplikasi diagnosa penyakit TBC menggunakan metode Naive Bayes," Skripsi, Universitas Muhammadiyah Surakarta, Surakarta.

H. Priyatman, F. Sajid, and D. Haldivany, "JEPIN (Journal of Informatics Education and Research) Clustering Using K-Means Clustering Algorithm to Predict Student Graduation Time," 2019.

Z. Nabila, A. Rahman Isnain, and Z. Abidin, "DATA MINING ANALYSIS FOR CLUSTERING COVID-19 CASES IN LAMPUNG PROVINCE WITH K-MEANS ALGORITHMA," Journal of Technology and Systems Information (JTSI), vol. 2, no. 2, pp. 100, 2021, [Online]. Available at: http://jim.teknokrat.ac.id/index.php/JTSI