ISSN 2355-0082

Developing HIV/AIDS Patient Profile Model Using K-Means Clustering Method

Rena Nainggolan[1], Fenina Adline Twince Tobing[2]
[1] Komputer Akuntansi, Universitas Methodist Indonesia, Medan , Indonesia
[2] Teknik Informatika, Universitas Multimedia Nusantara, Tangerang, Indonesia
[1]renanainggolan@methodist.ac.id

Accepted 08 June 2023
Approved 12 July 2023

Abstract*— RSUD Dr. Pirngadi is one of the Regional General Hospitals in North Sumatra that handles services for HIV/AIDS patients. Patients who actively take ARV therapy every month are low, namely around 20%. People infected with HIV become carriers and transmitters of the HIV throughout HIVes, HIV though they don't feel sick and look healthy, the sufferer still carries HIV. To be able to carry out more effective and efficient handling of the prevention and control from HIV/AIDS transmission, it is very important for the government and related parties, such as the Health Office, Social Services, and Hospital management, especially in VCT/CST (Voluntary Counseling and Testing/Care Support Treatment) to find out about understanding patient profiles, and prevent the development of HIV/AIDS disease transmission is very important. This knowledge can be used by the government to carry out programs that can prevent and break the chain of transmission of the HIV/AIDS virus as early as possible and help those involved in health services to become more familiar with the situation of their patients. This research takes the service area at RSUD Dr. Pirngadi Medan as one of the research domains in the field of data mining with data sources from RSUD Dr. Pirngadi Medan. This is done as information is known to the VCT/CST services at RSDU Dr. Pirngadi Medan. With data groupings like this, it is hoped that the Government or related agencies can create programs and implement them so that they can prevent and overcome the spread of HIV/AIDS in Indonesia. Obtained the number of patients in each cluster where for cluster 1 there were 7 patients and cluster 2 obtained as many as 3 clusters and cluster 3 obtained as many as 10 clusters.9 It is on this basis that the authors are interested in taking the title of the study regarding the formation of a patient profile model using the K-Mean clustering method.

***Keywords*— Clustering; K-Means Clustering; HIV/AIDS, Patient Profile**

I. Introduction

In Various fields of life today, a lot of data is generated but the data is hidden, to be able to find out the hidden information from the data, needed for to do data processing of the data. Especially in the current, it is very demanding for business people to use computer technology so they can complete. In the current era of globalization, which is called e-commerce, transactions in the Internet world are defined as e-commerce.

Data mining techniques are focused on building methods for disclosing knowledge stored in data and are used to uncover hidden information in data that is not visible on the surface but has the potential to be used. Data mining is a semi-automated process that applies statistical, learning, artificial intelligence, math, machine learning techniques to identify and extract .Until now, many algorithms have been developed to answer this problem, each of these algorithms is used based on the technique/function approach to processing input into the desired output. Broadly speaking, the techniques often used in data mining include association rule mining, clustering, classification, neural networks, and nearest country[1].

Cluster analysis plays an important role in classifying objects. Depending on the application, objects can be signals, customers, patients, news, plants, and others [2]. The algorithm that is often applied for clustering techniques because it makes estimates that are efficient and does not require many parameters is the K-Means AlgorithmK-Means uses predefined k groups (first k groups as centroids) [3].

The K-Means Clustering method is a method for grouping or classifying objects (data) into K-groups (clusters) based on certain criteria. Classification of data is done by calculating the shortest distance between cluster center data. The main concept of this method is to compile K centroid values or the average (mean) of a group of data with N dimensions, where this method requires that the K value be determined first.. The K-means algorithm starts with the formation of a prototype cluster at the beginning and then iteratively repairs the prototype cluster so that a convergent condition is achieved, namely a condition where no significant changes were found in the prototype cluster [4]. This turnover is measured by applying the objective function D which is defined as the sum or average distance of each data item from its centroid group.

II. Method

A. Data Definition

In Webster's New World's Dictionary, it is written that a datum: is something known or assumed*. That is, a datum (a single form of data) is something that is known/assumed. Thus, data can provide an overview of a situation or problem. Meanwhile, data according to the Oxford Dictionary is The Facts. So, it can be concluded that data is something that is known or assumed to be used for analysis, discussion, scientific presentation, or statistical tests. The data can be divided into 4 parts, namely:

1. Types of Data Based on Their Nature.

Data type can be like divided according to their nature, according to their sources, according to how they are obtained, and according to when they are collected. The nature of the data can be divided into two types, namely qualitative data (non-metric) and quantitative data (metric). Then the type of qualitative data is further divided into two types, namely nominal data and ordinal data. Likewise, the type of quantitative data is divided into two types, namely interval data and ratio data.

2. Types of Data by Source

The division of data types according to their sources is based on the sources of data acquisition, namely internal data and external data. Data grouped by an organization to define organizational conditions or activities that are related and useful for daily needs and internal control. For example production data, sales data at the company, employment data, financial data, and so on are internal data

External data is data collected to describe a situation or activity outside the organization. Examples of external data such as population data and national income data were obtained from the local statistical center office. A company needs external data such as population to predict demand potential, while national income data to determine the level of people's purchasing which is useful for the basis of price level policy.

3. Types of Data According to How to Obtain It.

Based on how to get it, data is divided into two types, namely secondary data and primary data. Secondary data is data obtained in finished form and has been processed by other parties. Secondary data is usually in the form of publications while primary data is data that is processed and collected by individuals or organizations, which are taken directly from the object. For example, a company wants to obtain the average use of a product by residents in a place by conducting direct interviews with the community.

4. Types of Data According to Time of Collection.

Based on the time of collection, data can be divided into two types, namely cross-sectional data and periodic data (time series). Cross-section data is data collected in a certain period, usually describing the conditions or activities in that period. For example, the results of the 2014 population census describe the condition of Indonesia in 2014 according to age, gender, religion, level of education, and so on.

Periodic data (time series) is data collected from time to time. Its purpose is to describe the development of activity over time. For example, the development of production in a company over the last five years, the development of product sales over the last five years, and so on. This type of data is also often referred to as historical data.

B. Data Mining

Data mining is a data processing method to find hidden patterns in data. The results of data processing with this data mining method can be applied to decision making in the future.

Data mining is a method of processing large amounts of data, therefore data mining has an important role in the fields of finance, industry, science, technology and weather. Broadly speaking, data mining studies discuss methods such as classification, clustering, regression, market basket analysis and variable selection, and [4]. Data mining is divided into several groups based on the tasks that can be done, namely

1. Description

Sometimes analytical research just wants to find a way to explain the patterns and trends present in the data. For example, voting committees may not be able to get hold of evidence or facts if they are not professional enough to gain little support in the presidential election. Pattern and trend descriptions often provide possible explanations for a pattern or trend.

2. Estimation

Estimation is almost the same as classification, the difference is that the target variable is estimated more numerically than categorically. The model is built by implementing a complete record that gives the value of the target variable as a predictive value. After that, in the next review, an estimate of the value of the target variable is made based on the value of the predicted variable. For example, in hospital patients it will be based on the patient's age, weight index, gender and blood sodium level to predict systolic blood pressure. The estimation model is generated by the value of predictive variables in the learning process. For other new cases, it can be obtained from other estimation models.

3. Predictions

Prediction is almost the same as estimation and classification, except that it estimates the value of future results. Examples of predictions in business and research are:

A. Estimated percentage increase in traffic accidents in the coming year if the lower speed limit is increased. Several techniques and methods applied in classification and estimation can be used (under suitable conditions) for predictions
B. Estimated price of rice in the next three months

4. Classification (Classification)

In classification, there are target categorical variables. For example, income classification can be grouped into three types, namely low income, medium income and high income

5. Clustering

Clustering is the process of observing, by collecting records or observing and creating classes of similar objects. Clusters are groups of records that correspond to each other and differ from records in other clusters. Clustering has no target variable in it so it is different from classification. Clustering does not attempt to estimate, classify, or estimate the value of the target variable. However, the clustering algorithm attempts to include all data into equal groups, where the similarity of records in one group will be high, while the similarity with records in other groups will be small.

Examples of clustering in business and research are:

1. Clustering the expression of genes, to obtain similar behavior from a large number of genes.
2. For accounting audits, namely to separate financial behavior in good or suspicious conditions.

6. Association (Association)

The way associations work in data mining is to look for attributes that appear at one time. In the business world it is more often referred to as shopping cart analysis.

Examples of associations in business and research are:

A. Finding products in the supermarket that are never bought together and those that are bought together.
B. Menarahakan customer groups by targeting the marketing of a product for companies that do not have large marketing funds.
C. Examine the number of cellular telecommunications company subscribers.
D. which is expected to provide a positive response to the service upgrade offers provided.

One technique known in data mining is clustering. The definition of scientific clustering in data mining is grouping several objects or data into clusters (groups) so that each group has data that is very suitable and different from data in other groups. Until now, scientists are still making various records to improve the cluster model and calculate how many clusters are optimal so that the best clusters can be produced. There are two clustering methods that we are familiar with, namely partition and hierarchical clustering. hierarchical clustering method itself consists of complete link clustering whereas the partitioning method itself consists of k-means and fuzzy k-means, average link grouping, single link grouping, and central link grouping.

C. *K-Means Clustering*

K-Means is an algorithm for grouping n objects based on attributes into k partitions, where k < n. The K-means method is the most common and simplest clustering method. This is because K-means is able to classify large amounts of data in a relatively fast and efficient time. The following figure shows the k-means clustering algorithm in action, for the two-dimensional case. Randomly generated initial centers to show more detailed stages. The partition space background is only for illustration and is not generated by the k-means algorithm[5].

The K-Means algorithm begins with a random determination of K, where K is the number of clusters you want to form. After that, assign K values randomly, for a while this value is the center of the cluster or commonly called the mean, centroid or "means". Find the shortest distance from each data to the centroid. By using the Euclidian formula to Classify each data based on its proximity to the centroid. Do this process so that the centroid value is fixed or does not change (stable). K-Means is an iterative clustering algorithm.

Stages of the K-Means Clustering Method [6]

1. The first stage determines the number of clusters.
2. The second stage is the determination of the cluster center. In this study, the determination of the cluster center was carried out randomly.
3. To determine the object/data to be placed in the cluster, calculations are performed based on the distance between the two objects, as well as the distance between the object and the center of the cluster. To calculate the distance of all data to the cluster center point.

Use Euclidean theory:[7]

$$D(i, j) = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + .. + (Xki - Xkj)^2}$$

Where:

D(i,j)= Distance of data i to the center of Cluster j
xki = Data to i on attribute data to k
Xkj = jth center point on the kth attribute

4. Recalculate the cluster center with the current cluster membership. The cluster center is the average of all data/objects in a particular cluster.
5. If the cluster center has not changed anymore, then the clustering process stops, or returns to step 3 until the cluster center remains.

D. HIV/AIDS

The Human Immunodeficiency Virus (HIV) is a virus belonging to the Ribonucleic Acid (RNA) class that specifically undermines the human body's immune system and causes Acquired Immunodeficiency Syndrome (AIDS) [6]. They are a potential source of infection for other people. People who have been infected with HIV and their bodies have formed antibodies (anti-bodies) against the virus are HIV positive.

AIDS (Acute Immunodeficiency Syndrome/SIDA) is a combination of clinical symptoms due to a weakening of the body's immune system that arises as a result of HIV infection. AIDS often manifests with the emergence of various opportunistic infectious diseases, malignancies, metabolic disorders, and others [8]

E. Patient Profile

The patient profile is influenced by the Adherence factor (Adherence/ Therapy or Medication Taking). Adherence or adherence to therapy is a condition where the patient adheres to his medication based on his awareness, not just because he obeys the doctor's orders. This is important because it is hoped that it will further increase the level of medication adherence. Compliance or compliance must always be considered and evaluated regularly at every consultation. Failure of ARV therapy or taking medication is often caused by patient disobedience in taking drugs or ARVs. [9] Patient characteristics include sociodemographic factors (gender, age, race/ethnicity, education, income, literacy/illiteracy/, health insurance, and group origin in society, for example (commercial sex workers or transgender) and psychosocial factors (drug use, mental health). Narcotics, Alcohol, Psychotropics, and other Addictive Substances) environment and social support, knowledge and behavior towards HIV and its therapy).

III. Result and Discussion

A. Data Transformation

For data to be managed by applying the k-mean clustering method, data of nominal data type such as the city of origin (address/city of origin/city of origin), risk factors, gender, occupation, and first the data must be initialized in the form of numbers. Process initialize the address/city of origin/city of origin/city of origin, the steps are as follows:

1. In the city of origin data, the region is first divided into several regions.
2. Then these areas are sorted Based on the number of patients coming from the area seen from the highest frequency of origin from the area.
3. The number of origins from the highest patient area will be given a number 1. After that the area with the second highest frequency will be given a number 2, and so on until the area with the lowest frequency.Apart from the city of origin, major, risk factors, ethnicity/ethnicity, education, and occupation are also included in the nominal data type, so it needs to be initialized as a number. As in the city of origin, department, risk factors, ethnicity/ethnicity, education occuand passion, they are also given an initialization based on the frequency of the number of HIV/AIDS patients present.

B. Development Techniques

This research will be carried out with the following steps.

1. Literature study and guidance consultation

At this stage, research materials are collected through various sources of literature, either in the form of books, journals, proceedings, magazines, and so on as supporting materials, and also carry out consultations with the thesis supervisor.

2. Field data collectionAt this stage field observations are carried out to collect the required data and problems that are often encountered.
3. Data initialization
   At this stage, data identification is carried out to determine the validity of the data and the variables that will be used. If the data is not valid then field observations are carried out again.
4. Preparation of test datasets After obtaining valid data, testing methods are now being developed so that the research objectives are fulfilled.
5. Data mining application design
   At this stage, the application is designed using Matlab.
6. Implementation of tests using applications and evaluation of results.
7. This stage is to test the data using the application program and analyze the test results and evaluate errors.
8. Compile the final assignment book This final stage is the documentation of supporting theories, application system design, test results, and analysis, as well as suggestions and conclusions.

Below is data on 20 patients for testing the Modified K-Mean Clustering Algorithm by applying the calculation of the Sum of Squared Error (SSE) value to determine the center of a cluster

Table 1. Preliminary Data

No	Initial	Daerah	Pekerjaan	Umur
0	A	Perjuangan Medan	Tenaga Kerja Indonesia	33
1	B	Selayang Medan	Pegawai Swasta	25
2	C	Belawan Medan	Pedagang	44
3	D	Perjuangan Medan	Tenaga Kerja Indonesia	31
4	E	Perjuangan Medan	Tenaga Kerja Indonesia	37
5	F	Selayang Medan	Pegawai Swasta	40
6	G	Kota Medan	Pegawai Swasta	24
7	H	Belawan Medan	Pegawai Negeri Sipil	42
8	I	Kota Medan	Pegawai Negeri Sipil	38
9	J	Kota Medan	Therapy	28
10	K	Selayang Medan	Tenaga Kerja Indonesia	27
11	L	Selayang Medan	Tenaga Kerja Indonesia	22
12	M	Perjuangan Medan	Tenaga Kerja Indonesia	27
13	N	Perjuangan Medan	Theraphys	30
14	O	Kota Medan	Pegawai Swasta	44
15	P	Belawan Medan	Pedagang	41
16	Q	Belawan Medan	Pedagang	24
17	R	Belawan Medan	Tenaga Kerja Indonesia	22
18	S	Belawan Medan	Pedagang	27
19	T	Selayang Medan	Pegawai Swasta	26

A. *Data Transformation*

In this study, to process the above data using the K-Means Clustering method, nominal data such as occupation and region must first be initialized in the form of numbers.

To initialize the region, sort from the highest must be based on the frequency of patients coming from that region. After that, an initial with the number 1 will be given to the area with the highest frequency, and an initial with the number 2 will be given to the area with the second largest frequency. The following can be seen in the initialization results table for regional categories.Apart from the region, occupation is also included in the nominal data type, so it needs to be initialized as a number. As with the region, the job is also given an initialization based on the frequency of the patient's work. The results of the initialization of the work can be seen in Table 2

Table 2. Job Data Initialization

No	Pekerjaan	Frekuensi	Inisialisasi
1	Tenaga Kerja Indonesia	7	1

2	Pegawai Swasta	5	2
3	Pedagang	4	3
4	PNS	2	4
5	Therapy	2	5

The data can be grouped using the K-Mean Clustering method. After processing all patient data is transformed to numbers. The next process needs to be grouped data into several clusters, namely the following stages:

1. Determine in advance the number of clusters. In this study, the existing data will be grouped into three clusters.
2. Then in each cluster determine the starting point. In this study, the initial center point was generated randomly. The cluster center on the initial solution can be seen in Table..

B. *Test Results*

The distance of each patient's data to the new cluster center in the 1st iteration

a. *Patient data 1*

The first process will calculate the distance from the patient data to the cluster center

$$D(1,1) = \sqrt{(2-2)^2 + (1-1)^2 + (33-37)^2}$$
$$= \sqrt{(0)^2 + (0)^2 + (-4)^2}$$
$$= \sqrt{0+0+16}$$
$$= \sqrt{16}$$
$$= 4$$

The result is that the distance between patient 1's data and the center of the first cluster is 4. From the calculation results above. Then the distance from the first patient data to the second cluster center will be calculated as below..

$$D(1,2) = \sqrt{(2-2)^2 + (1-1)^2 + (33-33)^2}$$
$$= \sqrt{(0)^2 + (0)^2 + (0)^2}$$
$$= \sqrt{0+0+0}$$
$$= \sqrt{0}$$
$$= 0$$

The results of the above calculation results show that the distance from the center of the second cluster to patient 1 is 0. Then the distance from the center of the third cluster to patient 1's data will be calculated as below

$$D(1,3) = \sqrt{(2-3)^2 + (1-1)^2 + (33-27)^2}$$
$$= \sqrt{(-1)^2 + (0)^2 + (6)^2}$$
$$= \sqrt{1+0+36}$$
$$= \sqrt{37}$$
$$= 6.083$$

The distance between the first patient data and the third cluster center was obtained at 6.083. So it can be concluded that patient 1's data is grouped into cluster 3 because the minimum distance is 6,043, this means that patient 1's data will become a member of cluster 3 (C3).

Table 3. The Distance of Each Patient Data to the Centroid Point in the 1st Iteration

No	**Initial**	**W**	**P**	**U**	**Jarak Ke**			**Min *Cluster***			
					C1	C2	C3	C1	C2	C3	
0	A	2	1	33	4.0	0.0	6.083		*		C2
1	B	3	2	25	12.083	8.124	2.236			*	C3
2	C	1	3	44	7.348	11.225	17.234	*			C1
3	D	2	1	31	6.0	2.0	4.123		*		C2
4	E	2	1	37	4.0	10.05	10.05	*			C1
5	F	3	2	40	3.317	7.141	13.038	*			C1
6	G	4	2	24	13.191	9.274	3.317			*	C3
7	H	1	4	42	5.916	9.539	15.427	*			C2
8	I	4	4	38	3.742	6.164	11.446	*			C2
9	J	4	5	28	10.05	6.708	4.234			*	C3
10	K	3	1	27	10.05	6.083	0.0			*	C3
11	L	3	1	22	15.033	11.045	5.0			*	C3
12	M	2	1	27	10.0	6.0	1.0			*	C3
13	N	2	5	30	8.062	5.0	5.099		*		C2
14	O	4	2	44	7.348	11.225	17.059	*			C1
15	P	1	3	41	4.583	8.307	14.283	*			C1
16	Q	1	3	24	13.191	9.274	4.123			*	C3
17	R	1	1	22	15.033	11.045	5.385			*	C3
18	S	1	3	27	10.247	6.403	2.828			*	C3

19	T	3	2	26	11.091	7.141	1.414			*	C3

Do the same process at the job of every patient.

The following will show a comparison of the number of patients in each cluster, where in cluster 1 the number of patients is 7 people, in cluster 2 there are 3 people and in cluster 3 there are 10 people.

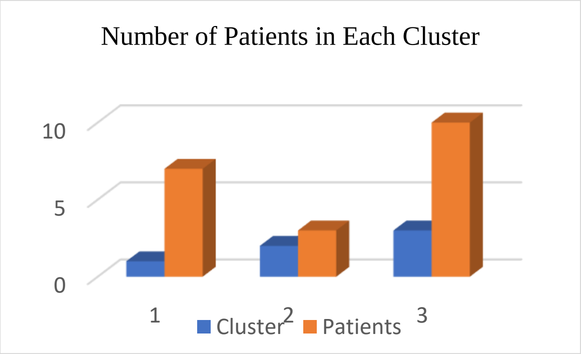

Fig 1. Number of Patients in Each Cluster

.

IV. Conclusion

Based on the test results and analysis of the results obtained, a conclusion can be drawn, namely:

1. From the results of testing and analysis we can see that the first cluster center point determines the number of iterations produced to obtain the final clustering.
2. In the K-Means Clustering Algorithm, data processing is very fast. This algorithm often experiences premature convergence so that its accuracy is not guaranteed. In this algorithm, the results look unsatisfactory, because it is not guaranteed that the distance between each centroid does not span so that if there are two or more groups with adjacent centroid points.
3. It cannot be concluded that the most optimum cluster center with the K-Mean Clustering method will produce the minimum iterations.
4. Obtained the number of patients in each cluster where for cluster 1 there were 7 patients and cluster 2 obtained as many as 3 clusters and cluster 3 obtained as many as 10 clusters

References

[1] Yuli Mardi, “Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika,” *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2019.

[2] M. W. Talakua, Z. A. Leleury, and A. W. Talluta, “Analisis Cluster Dengan Menggunakan Metode Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014,” *J. Ilmu Mat. dan Terap.*, vol. 11, no. 2, pp. 119–128, 2017.

[3] B. Harahap, “Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung),” *Reg. Dev. Ind. Heal. Sci. Technol. Art Life*, pp. 394–403, 2019, [Online]. Available: https://ptki.ac.id/jurnal/index.php/readystar/article/view/82

[4] W. M. P. Dhuhita, “Clustering Metode K-Means Untuk Menentukan Status Gizi Balita,” *J. Inform.*, vol. 15, no. 2, pp. 160–174, 2015.

[5] N. Wakhidah, “Clustering Menggunakan K-Means Algorithm (K-Means Algorithm Clustering),” *Fak. Teknol. Inf.*, vol. 21, no. 1, pp. 70–80, 2014.

[6] A. Rauf, Sheeba, S. Mahfooz, S. Khusro, and H. Javed, “Enhanced K-mean clustering algorithm to reduce number of iterations and time complexity,” *Middle East J. Sci. Res.*, vol. 12, no. 7, pp. 959–963, 2012, doi: 10.5829/idosi.mejsr.2012.12.7.1845.

[7] E. Z. Khulaidah and N. Irsalinda, “FCM using squared euclidean distance for e-commerce classification in Indonesia,” *J. Phys. Conf. Ser.*, vol. 1613, no. 1, 2020, doi: 10.1088/1742-6596/1613/1/012071.

[8] H. Sur, “Hubungan Pengetahuan HIV / AIDS dengan Stigma terhadap Orang dengan HIV / AIDS di Kalangan Remaja 15-19 Tahun di Indonesia (Analisis Data SDKI Tahun 2012) Relationship HIV / AIDS Knowledge related Stigma towards People Living with HIV / AIDS among Adole,” vol. 1, no. 2, pp. 35–43, 2017.

[9] Y. Marlinda and M. Azinar, “Jurnal of Health Education,” vol. 2, no. 2, pp. 192–200, 2017.