Infant Routine Monitoring System During 0-12 Months Immunization Using Agglomerative Hierarchical Clustering Algorithm

Eva Darnila[1], Maryana[2], Martunis[3]
[1,2,3]Faculty of Informatic Technology, University of Malikussaleh, Lhokseumawe, Indonesia

ABSTRACT

Data mining is the process of finding patterns from large data sets using description, estimation, prediction, classification, clustering and association techniques. One of the mininng data techniques used to group is the Agglomerative Hierarchical Clustering algorithm. The process of grouping data using Agglomerative Hierarchical Clustering aims to group objects based on the distance between the two clusters. There are several methods on Agglomerative Hierarchical Clustering algorithm namely Single Linkage, Average Linkage and Complete Linkage. In this study, the authors used the Average Linkage method. By using average linkage the author is interested to conduct research on infant immunization data in Kotajuang Health Center and Gandapura Health Center in Bireuen Regency. Immunization is the process to make a person immune or immune to a disease. This process is carried out by administering a vaccine that stimulates the immune system to be immune to the disease. The purpose of this study was to look at the grouping of infant immunizations based on 3 data variables namely gender, address and type of immunization. From the results of the study, the results were obtained for one of the groups consisting of 3 villages, namely Pante Sikumbang Village, Palohme Village and Ie Rhop Village. It has a male infant immunization rate of 54% and a female infant immunization rate of 77%. The information shown is that awareness of immunizations for baby boys and baby girls in this group "has started well".

Keyword : Data Mining, Agglomerative Hierarchical Clustering, Average Linkage, Immunizations, Bireuen.

Corresponding Author:
Name : Eva Darnila
Department : Faculty of Informatic Technology
Affiliation : University of Malikussaleh
Address : Lhokseumawe, Indonesia
Email : eva.darnila@unimal.ac.id

Article history:
Received Jan, 2021
Revised Jan, 2021
Accepted Jan, 2021

1. INTRODUCTION

Indonesia is a country with a high population growth rate. According to the National Population and Family Planning Agency (BKKBN), Indonesia's population growth rate still reaches 1.19 percent or around 2.3 million per year. Which means, there are about 2.3 million babies who will later become the nation's successors born each year. Maintaining baby's health is very important for its growth [1].

One of the efforts to maintain the health of infants is to provide immunizations. Immunization is the process of making a person immune or immune to a disease. This process is done by giving a vaccine that stimulates the immune system to be immune to the disease. In Indonesia itself, the whole process of giving immunizations has been going very well. By the end of 2019, the percentage had reached 93%. However, there are several provinces that have not yet reached this percentage. In Aceh Province itself, the percentage of immunization only reached 50% in the same year [2].

Puskesmas as one of the managers who organize immunization services plays an important role in straightening this understanding so that people have more confidence in immunizing. Therefore, in order to provide a more specific understanding to community groups, the puskesmas can monitor every time the community carries out immunizations. The data obtained from the monitoring results can later be grouped to find out which groups of people have regularly immunized babies and people who are not aware of infant immunization. This grouping can be done using data mining.

Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The selection of the right method or algorithm is very dependent on the objectives and the overall KDD process. One of the data mining techniques used to perform clustering is the Agglomerative Hierarchical Clustering algorithm [1] [3].

The process of grouping data using Agglomerative Hierarchical Clustering aims to group objects based on the distance between two clusters. There are several methods in the Agglomerative Hierarchical Clustering algorithm, namely Single Linkage, Average Linkage and Complete Linkage. In this study, the author uses the Average Linkage method. The results of this method will be analyzed to see the grouping of infants who are immunized.

2. RESEARCH METHOD

A. Average Linkage

Agglomerative Hierarchical Clustering is a hierarchical grouping method with a bottom-up approach. The grouping process starts from each data as a group, then recursively searches for potential groups based on distance as a pair to join as one larger group. The process is repeated continuously so that it appears to move up (agglomerative) to form a hierarchy (hierarchy). The key to the operation of the Agglomerative Hierarchical Clustering method is the use of a measure of proximity between two groups, or the cluster proximity parameter. Proximity can be defined as a measure that distinguishes groups [4].

One of the methods in Agglomerative Hierarchical Clustering is the average linkage method. Average linkage calculates the distance between two clusters which is called the average distance which is calculated in each cluster. The average linkage formula is as follows.

$$\boldsymbol{d_{(uv)w} = \frac{\sum i \sum k\, d_{ik}}{V_{(uv)} N_w}} \qquad (1)$$

dik = the distance of the object cl in the cluster (uv) and the object k in the cluster w.
V(uv) = number of objects in the cluster uv
Nw = number of objects in the cluster w

However, before doing the average linkage, first calculate the distance of each data so that later it can be clustered. The most frequently used distance calculation formula is the Euclidean distance.

Euclidean distance is the metric that is most often used to calculate the similarity of two vectors because it has a high level of accuracy and productivity. Euclidean distance is the distance between objects in a straight line. Euclidean distance formula is as follows.

$$\boldsymbol{d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}} \qquad (2)$$

d = distance between x and y
x = cluster center data
y = data on attribute
i = each data
n = number of data
xi = data at the center of the cluster to i
yi = data on each data to i

B. Scheme

The system schema for data processing using the algorithm is C4.5 as follows:

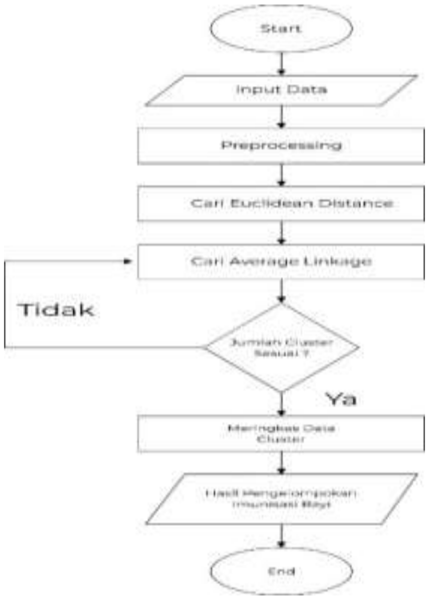

Fig 1. Scheme system

B. Average Linkage

The immunization data that has been taken previously from the research site is then transformed into a format that is in accordance with the system required, namely by removing unnecessary fields and replacing the immunization schedule value with binaryization, so that the transformation results are obtained with the format as shown in the table below.

Table 1. Transformation result data

No.	Gender	Address	Hb 0	Bcg	…	Polio 4	Ipv	Measles	Full Imunization
1	male	Ie Rhop	1	1	…	1	0	1	0
2	male	Ie Rhop	0	1	…	0	0	0	0
		…					…		
137	female	Cot Puuk	1	0	…	1	0	0	0

3. RESULTS AND DISCUSSION

A. Infant Immunization Preprocessing Data

From the calculation step, the search results are obtained as shown in table (2) data on preprocessing infant immunization in Gandapura District. From the search results, the average value was obtained based on the sex of the baby at the time of the immunization process. The gender of the baby boy is indicated by the letter x and the sex of the baby girl is indicated by the letter y.

Table 2. Data on preprocessing of infant immunization in gandapura district

No.	Village	X	Y
1	Cot Puuk	0.125	0.35
2	Cot Teubee	0.257575758	0.333333333

3	Cot Tufah	0.208333333	0.35
4	Ie Rhop	0.464285714	0.736111111
5	Lingka Kuta	0	0.574074074
6	Paloh Kaye	0.277777778	0.25
7	Paloh Me	0.5625	0.833333333
8	Pante Sikumbang	0.583333333	0.738095238
9	Paya Baro	0.291666667	0.333333333
10	Pulo Gisa	0.354166667	0.375
11	Tanjong Raya	0.083333333	0.361111111
12	Ujong Bayu	0.083333333	0.305555556

B. Euclidean Distance Value

From the data in table (2) it is used to calculate the Euclidean distance. The results of the calculation of the euclidean distance can be seen in table (3) below:

Table 3. Euclidean distance value

d	1	2	…	9	10	11	12
1		0,1336193	…	0,1674979	0,2305263	0,0431227	0,0609214
2			…	0,0340909	0,1051947	0,1764427	0,1764427
3				0,0849837	0,1479607	0,1254929	0,1326662
4				0,4382092	0,3775281	0,5345556	0,5748937
5				0,3781872	0,4062813	0,2286868	0,2811523
6				0,0844828	0,1464932	0,2239516	0,2022253
7				0,5686393	0,5034602	0,6727515	0,7128465
8				0,4989004	0,4293664	0,6261925	0,6611283
9					0,0751157	0,210177	0,210177
10						0,2711892	0,2795948
11							0,0555556
12							

C. Average Linkage Value

From the search results in table (3), the smallest value is the 18th data, namely the distance variable d(2,9) with a value of 0.034090909. In the next iteration, the distance variable d(2,9) is calculated for the average linkage with all other distance variables. After getting the average linkage result for the distance variable d(2,9) , steps are taken to get the smallest value until the last iteration. The results of the average linkage calculation can be seen in table (4) below:

Table 4. Average linkage value

Iteration	Result

1st $(d_{(1)}, d_{(2)}, d_{(3)}, d_{(4)}, d_{(5)}, d_{(6)}, d_{(7)}, d_{(8)}, d_{(9)}, d_{(10)}, d_{(11)}, d_{(12)})$	0.034090909
2nd $(d_{(1)}, d_{(2,9)}, d_{(3)}, d_{(4)}, d_{(5)}, d_{(6)}, d_{(7)}, d_{(8)}, d_{(10)}, d_{(11)}, d_{(12)})$	0.043122707
3rd $(d_{(1,11)}, d_{(2,9)}, d_{(3)}, d_{(4)}, d_{(5)}, d_{(6)}, d_{(7)}, d_{(8)}, d_{(10)}, d_{(12)})$	0.058238489
4th $(d_{(1,11,12)}, d_{(2,9)}, d_{(3)}, d_{(4)}, d_{(5)}, d_{(6)}, d_{(7)}, d_{(8)}, d_{(10)})$	0.068485069
5th $(d_{(1,11,12)}, d_{(2,9,3)}, d_{(4)}, d_{(5)}, d_{(6)}, d_{(7)}, d_{(8)}, d_{(10)})$	0.097325912
6th $(d_{(1,11,12)}, d_{(2,9,3,6)}, d_{(4)}, d_{(5)}, d_{(7)}, d_{(8)}, d_{(10)})$	0.097490115
7th $(d_{(1,11,12)}, d_{(2,9,3,6)}, d_{(4)}, d_{(5)}, d_{(7,8)}, d_{(10)})$	0.118691048
8th $(d_{(1,11,12)}, d_{(2,9,3,6,10)}, d_{(4)}, d_{(5)}, d_{(7,8)})$	0.128630206
9th $(d_{(1,11,12)}, d_{(2,9,3,6,10)}, d_{(5)}, d_{(7,8,4)})$	0.187062096
10th $(d_{(1,11,12,2,9,3,6,10)}, d_{(5)}, d_{(7,8,4)})$	0.329530615
11th $(d_{(1,11,12,2,9,3,6,10,5)}, d_{(7,8,4)})$	0.55510094
12th $(d_{(1,11,12,2,9,3,6,10,5,7,8,4)})$	0.55510094

D. Output

From the results of the average linkage in table (4), the output in the form of village names and information on immunization levels is displayed. In this study, the required grouping is 4 clusters. The iteration that has 4 clusters is in the 9th iteration. The output results can be seen in table (5) below:

Table 5. Clustering result

Cluster	Village	Result
1	1. Desa Lingka Kuta	Immunization rates for boys are very low, while girls are doing well.
2	1. Desa Pante Sikumbang 2. Desa Palohme 3. Desa Ie Rhop	Immunization rates for male and female infants have started to improve.

3	1. Desa Pulo Gisa 2. Desa Paya Baro 3. Desa Cot Teubee 4. Desa Cot Tufah 5. Desa Paloh Kayee Kunyet	Immunization rates for boys and girls are very low.
4	1. Desa Ujong Bayu 2. Desa Tanjong Raya 3. Desa Cot Puuk	Immunization rates for boys and girls are very low.

4. CONCLUSION

The final result of this study is a monitoring system for infants undergoing routine immunization 0-12 months with the implementation of the Agglomerative Hierarchical Clustering algorithm, based on the number of attendances in immunization vaccines, namely BCG (1 dose), DPT (3 doses), polio (4 doses), hepatitis B (3 doses), and measles (1 dose) had good results. The results obtained for one group consisted of 3 villages, namely Pante Sikumbang Village, Palohme Village and Ie Rhop Village. It has an immunization rate for boys of 54% and an immunization rate for girls of 77%. Thus awareness of immunization in this group "has started well".

REFERENCES

[1] R. R. Putra and C. Wadisman, "Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means," INTECOMS: Journal of Information Technology and Computer Science, vol. 1, pp. 72-77, 2018.

[2] Z. Arifin, S. Santosa, and M. A. Soeleman, "Klasterisasi Genre Cerpen Kompas Menggunakan Agglomerative Hierarchical Clustering-Single Linkage," Jurnal Cyberku, vol. 13, pp. 2-2, 2017.

[3] Q. Nafisah and N. E. Chandra, "Analisis Cluster Average Linkage Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur," Zeta-Math Journal, vol. 3, pp. 31-36, 2017.

[4] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," J. Inform, vol. 4, 2019.