INTERNATIONAL MULTIDICIPLINARY JOURNAL
VOLUME 1
NO. 1

Application of the K-Means Clustering Algorithm Analysis on Human Infectious Diseases (Case Study: Pusuk II Simaninggir Village)

Author Name: Jessicha Gratia Sitohang [1], Sorang Pakpahan S.Kom., M.Kom [2]

Affiliation: Catholic University of Santo Thomas[1,2]

Contact Information: jessichasitohang08@gmail.com

Abstract

Infectious diseases pose a serious threat to human health. This research applies data mining techniques to transform large volumes of data into useful information. To address this complex issue, data analysis is essential for understanding the distribution patterns and characteristics of diseases. One method employed is the K-Means clustering algorithm, which effectively groups data based on similar characteristics. This research explores the application of the K-Means clustering algorithm to human infectious disease data in order to identify distribution patterns and relationships between cases. The goal is to analyze the data of six types of infectious diseases in humans, ranked from highest to lowest prevalence. The diseases examined include Tuberculosis (TB), Dengue Hemorrhagic Fever (DHF), Diarrhea, Influenza, Chickenpox, and Measles. The data used in this study was obtained from the recapitulation of human infectious disease records in the population of Pusuk II Simaninggir village from 2018 to 2022. The conclusion of this study is that the K-Means method, along with testing through the RapidMiner application, simplifies data processing, provides accurate final results, and is highly effective for big data analysis

Keywords

Infectious Diseases, K-Means Clustering, RapidMiner, Data Mining,

Introduction

In the current era of Industry 4.0, advancements in information technology have rapidly developed across various fields of life. A vast amount of data is generated by increasingly sophisticated information technology, spanning industries, economics, science, technology,

and many other sectors. The application of information technology in the healthcare sector also generates abundant data regarding human infectious diseases.

Advancements in computer network technology (the internet) have enabled communication and interaction between data that are physically separated. In today's era of technological development, data analysis plays a crucial role in various fields, including healthcare. One important application of data analysis is in understanding human infectious diseases, which can aid in prevention, detection, and management efforts. Furthermore, significant progress in computer science and mathematics allows us to group data more efficiently, one of which is through the use of the K-Means Clustering algorithm.

With the advancement of technology, health activities that were once conducted conventionally are gradually shifting towards digital methods. Barriers that made it difficult to consult doctors about diseases are now being overcome by computer programs. In this context, information data can assist in solving health-related issues by providing advice to readers and finding solutions to various specific problems.

Cluster analysis is a multivariate technique aimed at grouping human infectious diseases based on the frequency of occurrence. In this study, cluster analysis was conducted in Pusuk II Simaninggir Village, Parlilitan Subdistrict, Humbang Hasundutan Regency. Cluster analysis of infectious diseases in humans was performed so that diseases with the most similarities to other objects would be placed in the same cluster. The formed clusters exhibit high internal homogeneity and external heterogeneity

Literature Review

1. **Definition of Analysis**
   Analysis is a set of interrelated activities, processes, and actions aimed at solving problems or breaking down components into more detailed parts, which are then recombined to draw conclusions. One form of analytical activity is summarizing raw data into information that can be communicated to the public. All forms of analysis depict consistent patterns within the data, so the results of the analysis can be studied and interpreted in a concise and meaningful way
2. ***K-Means Clustering***
   The K-Means Clustering method, also known as the K-Means Clustering Algorithm, is a well-known and practical method in data processing. Its purpose is to group data or objects into several clusters, where each cluster will contain data that are closest to the centroid of that cluster. According to Rochcham (2020), the K-Means Algorithm is a data mining method frequently used to identify and analyze similarities in data clustering. According to Thabit et al. (2020), K-Means is an algorithmic method that analyzes data by randomly determining values for the data to be clustered and identifies objects within the same group that share similarities or have relationships, as opposed to those in other groups
3. **Data Mining**
   According to Santoso (2017), Data Mining is a data processing method used to discover new patterns within the data. The utilization of Data Mining is indeed useful as a means to add information across various fields, from business to healthcare. This is also proven after reviewing the definition of data mining according to experts. According to Kurnia et al. (2020), Data Mining is a combination of several disciplines in computer science
4. **Infectious Diseases in Humans**
   An infectious disease is also known as an infection that can spread to humans, caused by biological agents such as viruses, bacteria, fungi, and parasites; not caused by

physical or chemical factors. Transmission can occur directly or through a medium, vector, or disease-carrying animals

Methodology

The K-Means Clustering method, also known as the K-Means Clustering Algorithm, is a well-known and practical method in data processing. Its purpose is to group data or objects into several clusters (groups), with each cluster containing data that is closest to its respective centroid.

1. Determine the number of clusters (K) to be formed.
2. Determine the clustering points randomly based on the clusters.
3. Calculate the distance between the data points and the clustering points using the Euclidean Distance formula:

$$D_{L_i)}(x_2, x_1) = \| x_2 - x_1 \| \sum_{j=1}^{p} |x_{2j} - x_{ij}|$$

4. After the data is grouped based on the closest distance to each clustering point, to determine or calculate the new clustering point, the average value of the data points in each cluster is calculated
5. erform the iterative process until completion. The iteration ends when the values are the same as the previous iteration

Findings

The collected data will be analyzed using the K-Means clustering method. The steps of the analysis will include the following steps

a. Data Processing

1. Data yang akan diolah menggunakan metode K-means ini diperoleh dari catatan medis atau basis data kesehatan yang dikelola oleh Poskesdes Pusuk II Simaninggir. Bentuk data awal sebelum diolah dapat di lihat pada Figure 2

No	Name	Address	Age	Gender	Symptoms	Diagnosis	tment Dur	Immunization Status	Recovery Status	Status
1	Tiwi Sihotang	Sosor	5	Female	High fever, rash	Measles	14 Days	Complete	Recovered	Already
2	Raguel Panjatan	Banera	7	Male	Cough, cold, rash	Measles	10 Days	Complete	Recovered	Already
3	Citra Andini Marbun	Sipang	4	Female	Fever, rash, red eyes	Measles	15 Days	Complete	Recovered	Already
4	Egi Wisno Manullang	Huta Toruan	6	Male	High fever, rash	Measles	15 Days	Complete	Recovered	Already
5	Irfan Mahulae	Lumban Nault	6	Female	Cough, rash, fever	Measles	14 Days	Complete	Recovered	Already
6	Silvi Sihotang	Huta Dolok	6	Female	Cold, rash, fever	Measles	11 Days	Complete	Recovered	Already
7	Celsi Munte	Huta Gonting	6	Female	Cold, rash, cough	Measles	14 Days	Complete	Recovered	Already
8	Sorinda Sintjak	Sosor	6	Female	High fever, rash	Measles	18 Days	Complete	Recovered	Already
9	Anggel Marbun	Banera	7	Female	Cough, cold, rash	Measles	15 Days	Complete	Recovered	Already
10	Putra Situmorang	Sipang	4	Male	Fever, rash, red eyes	Measles	15 Days	Complete	Recovered	Already
11	Mita Butona	Huta Toruan	6	Female	High fever, rash	Measles	17 Days	Complete	Recovered	Already
12	Margaretta Panjatan	Lumban Nault	7	Female	Cough, rash, demam	Measles	14 Days	Complete	Recovered	Already
13	Samuel Situmorang	Huta Dolok	6	Male	Cold, rash, demam	Measles	15 Days	Complete	Recovered	Already
14	Maximus Sihotang	Huta Gonting	6	Male	Fever, rash, cough	Measles	14 Days	Complete	Recovered	Already
15	Putri Munte	Sosor	7	Female	High fever, rash	Measles	14 Days	Complete	Recovered	Already
16	Ray Simbolon	Sipang	6	Male	Fever, red eyes	Measles	14 Days	Complete	Recovered	Already
17	Nur Sihotang	Huta Toruan	5	Female	Cold, rash, fever	Measles	15 Days	Complete	Recovered	Already
18	Deni Nainggolan	Huta Toruan	6	Male	High fever, rash	Measles	9 Days	Complete	Recovered	Already

Figure 1 Initial disease data to be processed

2. After the data is obtained, the next step is to select data that can be calculated and used as indicators, such as age, gender, symptoms of the diseases present in each infectious disease, treatment duration, and others, to perform clustering or grouping on each infectious disease in Pusuk II Simaninggir village. More details can be seen in Figure 3 below.

Number	Name	Age	Gender	Cough	Fever	Loss of Appetite	Weight Loss	Night Sweats	Fatigue	Exhaustion
1	Martin Sintipik	32	Male	Dry Cough	NO	NO	NO	NO	NO	NO
2	Pakrun Sintang	35	Male	Phlegm Cough	NO	NO	NO	NO	NO	NO
3	Budi Munte	30	Male	Phlegm Cough	NO	NO	NO	NO	NO	NO
4	Tita Sotang	34	Male	Phlegm Cough	NO	NO	NO	NO	NO	NO
5	Santo Mahulue	50	Male	Phlegm Cough	NO	NO	NO	NO	NO	NO
6	Frans Munte	48	Male	Dry Cough	NO	NO	NO	NO	NO	NO
7	Amelia Sintiringo	42	Female	No Symptoms	YES	YES	NO	NO	NO	NO
8	Santoso Simorangkir	44	Female	Phlegm Cough	NO	NO	NO	NO	NO	NO
9	Tita Sintang	38	Male	Dry Cough	NO	NO	NO	NO	NO	NO
10	Tina Sintang	34	Female	Phlegm Cough	NO	NO	NO	NO	NO	NO
11	Benediktus Buaton	29	Male	Phlegm Cough	NO	NO	NO	NO	NO	NO
12	Nando Munte	33	Male	Dry Cough	NO	NO	NO	NO	NO	NO
13	Parasant Sintang	45	Male	No Symptoms	NO	NO	NO	NO	NO	NO
14	Kamaru Paijat	46	Male	Phlegm Cough	NO	NO	NO	NO	NO	NO
15	Rostani Manulang	40	Female	Dry Cough	NO	NO	NO	NO	NO	NO
16	Juhrana Sibuea	29	Female	No Symptoms	NO	NO	NO	NO	NO	NO
17	Santo Mahulue	42	Female	Phlegm Cough	NO	NO	NO	NO	NO	NO
18	Merina Ningrat	44	Female	Dry Cough	NO	NO	NO	NO	NO	NO
19	Julius Huta	32	Male	No Symptoms	NO	NO	NO	NO	NO	NO
20	Ferina Manihuruk	31	Female	No Symptoms	NO	NO	NO	NO	NO	NO
21	Putri Martan	28	Male	No Symptoms	NO	NO	NO	NO	NO	NO
22	Ramy Sintang	29	Female	Dry Cough	NO	NO	NO	NO	NO	NO

Findings Figure 2 Data used as indicators for grouping

3. After the indicators are obtained, the next step is to convert the data from text to numerical values so that it can be implemented into the K-means method. The conversion of text data from the disease recap data at the Pusuk II Simaninggir health post can be seen in the figure below.

Number	Name	Age	Gender	Cough	Fever	Loss of Appetite	Weight Loss	Night Sweats	Fatigue	Exhaustion
1	Martin Sintipik	30	Male	1	0	0	1	1	1	0
2	Pakrun Sintang	35	Male	1	0	0	1	0	1	0
3	Budi Munte	31	Male	1	0	0	1	1	0	0
4	Tita Sotang	29	Male	1	0	0	1	0	0	0
5	Santo Mahulue	50	Male	1	0	0	1	0	0	1
6	Frans Munte	42	Male	0	0	0	0	1	1	0
7	Amelia Sintiringo	42	Female	0	1	1	0	0	0	0
8	Santoso Simorangkir	29	Female	1	0	0	0	0	0	1
9	Tita Sintang	34	Male	1	0	0	0	1	0	1
10	Tina Sintang	29	Female	0	0	0	0	1	0	0
11	Benediktus Buaton	28	Male	0	0	0	0	0	0	1
12	Nando Munte	44	Male	0	0	0	0	0	1	0
13	Parasant Sintang	38	Male	0	0	0	0	0	1	0
14	Kamaru Paijat	44	Male	0	0	0	0	0	0	1
15	Rostani Manulang	41	Female	0	0	0	0	0	0	0
16	Juhrana Sibuea	31	Female	0	0	0	0	0	0	0
17	Santo Mahulue	28	Female	0	0	0	0	0	0	1
18	Merina Ningrat	46	Female	1	0	0	1	0	1	1
19	Julius Huta	45	Male	0	0	0	0	1	0	0
20	Ferina Manihuruk	29	Female	0	0	0	0	0	1	0
21	Putri Martan	31	Male	1	0	0	1	1	1	1
22	Ramy Sintang	29	Female	0	0	0	0	0	0	0

Figure 3: Results after being converted into numerical data

b. Analysis using the K-Means Method

1. After all the infectious disease data from the Pusuk II Simaninggir village, obtained from the health post data recap, has been converted into numerical form, the data can be grouped using the K-Means Clustering algorithm. To cluster these data into several clusters, several steps need to be followed, namely: Determine

the desired number of clusters. In this study, the data will be grouped into four clusters. Determine the cluster centroids randomly, as this will affect the results obtained in subsequent iterations. The random cluster points used for the first iteration are patient data number 7 for C1, patient data number 16 for C2, patient data number 2 for C3, and patient data number 8 for C4. The purpose of determining these random cluster points is to calculate the nearest centroid for the first iteration.

2. To determine which cluster is closest to the data, the distance between each data point and the centroid of each cluster needs to be calculated. To calculate the distance from the first patient data to the center, the first cluster calculates the distance of each data point to the centroid value. The following formula can be used for the first iteration data.

$$D_{L_i)}(x_{2,}x_1) = \| x_2 - x_1 \| \sum_{j=1}^{p} |x_{2j} - x_{ij}|$$

3. Using the same method, the calculation can be performed for the next cluster and the next patient, using the cluster as per the random cluster data provided.
4. The next step is to determine the cluster location by comparing the four clusters. The minimum value is the one selected; once the smallest value (minimum) is found, the data can be assigned to that cluster.
5. The next step is to determine the new centroid value for the next iteration, or iteration 2. This value is determined by the data that belong to the cluster.
6. For each indicator of each disease that belongs to cluster 1, the values are summed and divided by the number of disease data in cluster 1.
7. This process continues for each indicator in C1, C2, C3, and C4, and the patient data that belong to each cluster are then divided by the number of patients in the cluster. This continues until each indicator for all four clusters is completed.
8. To find the centroid value for the next iteration, repeat the steps as in the first iteration. Once the new centroid values are found, repeat the distance calculation step from the previous step, and then find the minimum value. The final data is when the above steps are repeated with the same process until the values for a cluster remain exactly the same as the previous data, or in other words, there is no change in its position within the cluster. At that point, the centroid calculation stops at that iteration.
9. From the results of the data tested and manually calculated using the K-Means Algorithm, it can be concluded that the clustering calculation for the tuberculosis (TBC) disease data is completed at the 5th iteration, because the centroid points for each cluster no longer change and no data move from one cluster to another, compared to the previous iteration, which was the 4th iteration.

After the steps above are performed, the following table shows the number of patients in each cluster from the tuberculosis (TBC) disease data in the Pusuk II Simaninggir village.

Table 1. Number of Patients in Each Cluster

Clusters	Number of patients
C1	18 patients

C2	41 patients
C3	50 patients
C4	46 patients

The manual calculation above using Excel is the grouping of patients for tuberculosis (TBC). Similarly, for other infectious diseases such as chickenpox, measles, dengue, diarrhea, and influenza, the calculations are performed in the same way as for tuberculosis mentioned above. By using the same method for the other infectious diseases, the grouping results are as follows

Table 2. Grouping Results for Chickenpox

Number	Clusters	Number of patients	weight (%)
1	C1	27 patients	30%
2	C2	27 patients	30%
3	C3	22 patients	24,45%
4	C4	14 patients	15,55%

Table 3. Grouping Results for Measles

Number	Clusters	Number of patients	weight (%)
1	C1	11	17.74%
2	C2	14	22.58%
3	C3	19	30.64%
4	C4	18	29.04%

Table 4. Grouping Results for Dengue Fever (DBD)

Number	Clusters	Number of patients	Weight(%)
1	C1	15	9.38%
2	C2	48	30 %

3	C3	42	26.25%
4	C4	55	34.37%

Table 5. Grouping results for diarrheal disease

No	Clusters	Number of patients	Weight (%)
1	C1	30	15.15%
2	C2	50	25.25%
3	C3	42	21.21%
4	C4	76	38.39%

Table 6. Grouping results for influenza disease

No	Clusters	Number of patients	Weight (%)
1	C1	14	10.37%
2	C2	25	18.51%
3	C3	43	31.85%
4	C4	53	39.27%

Table 7. Grouping results for TB disease

Number	Clusters	Number of patients	Weight (%)
1	C1	18	11,61%
2	C2	41	26,45%
3	C3	50	32,25%
4	C4	46	29,69%

c. Implementasi Model

The results of the data visualization depicting the number of cases from various diseases, such as chickenpox, measles, dengue fever, diarrhea, influenza, and tuberculosis, are presented in the form of a histogram to facilitate analysis and understanding of the trends in the spread of these diseases

Figure 4. Visualization of disease in the form of a histogram diagram

Conclusion

Based on the results of this study, it can be concluded that data clustering using the K-Means Clustering method and the testing of the RapidMiner application provide ease in processing complex and large data. This method has proven to be effective in identifying significant patterns in the data, producing accurate final results, and improving efficiency in the data processing process. Furthermore, the RapidMiner application allows users to manage and analyze data more quickly and systematically, making it a highly suitable solution for processing large data in various fields

References

1. Bastian, A., Sujadi, H., & Febrianto, G. (2018). Application of the K-Means Clustering Algorithm in Human Infectious Diseases (Case Study of Majalengka Regency). *Journal of Information System*, 14(1), 26–32.
2. Dhuhita. (2015). Application of the K-Means Clustering Algorithm in Infectious Diseases: Case Study of Majalengka Regency.
3. Ediyanto, Mara, N., & Satyahadewi, N. (2013). Classification of Characteristics Using the K-Means Cluster Analysis Method. *Scientific Bulletin of Mathematics, Statistics, and Its Application (Bimaster)*, 02(2), 133–136.
4. Murti. (2017). Application of the K-Means Clustering Method to Group Fruit Production Potential in the Special Region of Yogyakarta Province.
5. Purba, N., Poningsih, P., & Tambunan, H. S. (2021). Application of the K-Means Clustering Algorithm in the Spread of Acute Respiratory Infections (ARI) in Riau Province. *Journal of Information System Research (JOSH)*, 2(3), 220–226.
6. Sugianto, C. A., Rahayu, A. H., & Gusman, A. (2020). K-Means Algorithm for Clustering Patient Diseases at Cigugur Tengah Health Center. *Journal of Information Technology*, 2(2), 39–44. https://doi.org/10.47292/joint.v2i2.30
7. Tampubolon, S. S. (2011). Application of the K-Means Clustering Method to Group Fruit Production Potential in the Special Region of Yogyakarta Province.
8. Hendra, T., & Putra, A. S**.** (2019). A Comparative Study of Clustering Algorithms for Medical Data Analysis: A Case of Infectious Diseases. *Journal of Computational Science and Technology*, 5(4), 135-143.
9. Jaya, S., & Widodo, A. (2016). Utilization of K-Means Clustering for Early Detection of Infectious Disease Patterns in Urban Areas. *Journal of Medical Informatics*, 10(1), 48-55.
10. Rangga, L., & Fitri, L. (2018). Application of Data Mining Techniques in Health Sector: The Use of K-Means Clustering for Disease Classification. *Health Informatics Journal*, 3(2), 87-92.