Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2022. Vol. 1. No. 2

Grouping Medical Record Data By Type Diseases With K-Means Algorithm

Remonaldi Purba[1], Sumarno[2], Iin Parlina[3], Rafiqa Dewi[4], Ika Purnama Sari[5]

[1,2,5]*STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia*
[3,4]*AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia*
**remonaldipurba@gmail.com*

Abstract

Health is a very valuable thing for human life, because anyone can be affected by health problems without realizing what causes it. People who pay less attention to their health are more likely to get sick. Lack of awareness in protecting and preserving the environment will lead to the rapid spread of disease. Efforts in disease prevention are needed by increasing public awareness about the importance of clean and healthy living behavior. In the application of the k-means algorithm for data processing in finding medical record files in the form of notes and documents about patient identity, examination, treatment, and other service actions given to patients. Clustering is a data analysis method that performs the modeling process without supervision (unsupervised) is also a method that performs data grouping with a partition system. The result is grouping using K-Means Clustering which can help in grouping by type of disease and age, the results are divided into children and toddlers, young and adults, old and elderly.

Keywords: *Medical Records, Types of Diseases, K-Means, Clustering*

1. Introduction

Hospital is a health service institution that provides complete individual health services that provide inpatient, outpatient and emergency services[1] (Regulation of the Minister of Health of the Republic of Indonesia No. 340/MENKES/PER/III/2010). Hospitals have the function and purpose of health service facilities that carry out service activities in the form of outpatient services, inpatient services, emergency services, referral services which include medical record services and medical support and are used for education, training, and research for health workers[2]. Medical records are identity information and medical history by patients in treatment centers (hospitals, health centers). Disease patterns that are often suffered by people in a group of areas can be detected from the information contained in a collection of medical records[3]. In the field of health, medical records are also known as ICD (International Classification Diseases) which is a record of the history of patients undergoing treatment. The creation of medical records in hospitals is a goal in creating good service to patients. In addition, the benefits of medical record data can be used as a guide to analyze or diagnose a disease, so that planning, treatment and medical actions can be immediately carried out on patients.

With the data mining technique, data that initially seems unimportant can be used to find useful information from a very large data set with the aim of getting a decision that is very easy to implement and get good and accurate results[4],[5]. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques that extract and identify useful information and knowledge that is assembled from various large databases[6],[7][8]. Clustering is a data mining method that is unsupervised and a method for finding and grouping data that have similar characteristics between one data and another [9],[10],[11]. In conducting this research, the use of the k-means algorithm[12],[13],[14] is needed to perform data processing in order to make it easier to find out what potential diseases often attack the community and the hospital will certainly know and improve the form of health services to the community.

2. Research Methodology

In this study, the research method is used as the steps used to solve problems regarding the grouping of medical record data based on the type of disease using the K-Means algorithm. Where in this section has a case study of data collected in 2020 in the form of an excel file spreadsheet obtained from the Tuan Rondahaim General Hospital. By conducting research used must have accurate or valid data. The design of this study was carried out by making observations to study the classification of medical record data by knowing what types of diseases are often encountered in medical record data. This research design or model is made in the form of a diagram presented in the Flowchart design in Figure 1 below:

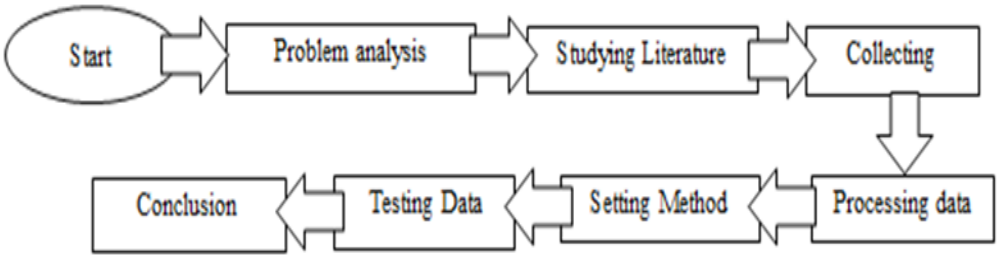

Figure 1. Research design

Figure 1 describes the research design carried out on the application of Data Mining in grouping medical record data to determine the most types of diseases using the K-Means Clustering Method which consists of Problem Analysis is a stage that must understand the purpose of research by knowing what types of diseases exist. And the need for research objectives is to determine the pattern of the disease based on age classification. Studying the literature is understanding and making targets from patient medical record data and focusing on variables or data samples to be taken. After that the medical record data sample will be cleaned with the aim of getting consistent data. Collecting data by using interview techniques to get data from the head of the Medical Record. Processing Data, namely Performing data processing, namely from the form of raw data that is transferred into the form of processed data ready to use using Ms. Excel 2010 (xls). Setting Method ie Establishing a method to solve the problem. In this study, data mining with the k-means clustering method was used. Testing Data, namely doing data testing is done using the RapidMiner application version 5.3. Making a conclusion which is the final stage in making reports on the results of the patient's medical record data that has been processed. The final report containing the grouping of data in the data mining process. Then it can be seen the grouping of types of diseases at what age a person is susceptible to disease and know what potential diseases often occur.

So that the data can be processed by calculating K-Means clustering, the alphabetic data will be initialized in numeric form. The following table of medical records based on the type of disease can be seen in the sample table 1:

Table 1. Medical Record Data Sample

No	Type of disease	Babys - childrens	young s/d mature	Old
1	Abdominal Pain	1	7	5
2	Anemia	6	16	35
3	Anxietas	0	3	26
4	Anxiety	0	5	6
5	Asthma	2	8	4
6	BPH	0	0	37
7	Brocnhitis	13	14	29
8	Cephalgia	0	10	15
9	CHF	0	12	115
10	Cholic Abdomen	1	1	5
11	CKR	0	12	2
12	Cytitis	0	14	2
13	DBD	4	6	5
14	Depression	0	34	0
15	DHF	4	15	8
16	Diabetes	0	6	16
17	Diare	3	3	1
18	Dyspepsia	7	177	188
19	Efusi Pleura	1	17	11
20	Epilepsi	53	66	0
21	Fatigue	0	3	4
22	Febris	27	36	9
23	Anxiety Disorder	0	0	18
24	Panic Disorder	0	0	21
25	Psychotic Disorder	0	4	113
26	Gastritis	1	11	13
27	Gastro Entritis	2	3	2
28	GEA	7	11	7
29	GERD	0	62	132
30	Gouth Astritis	0	6	16
31	Hepatitis	3	3	4
32	Hepoglekimia	0	1	15

33	Hipertensi	1	32	227
34	HIV	0	4	3
35	ISK	2	18	30
36	ISPA	25	12	8
37	Pregnancy	0	80	0
38	Constipation	2	6	8
39	Myalgia	1	30	70
40	Nefrolitiasis	0	2	12
41	OA Gonue	0	1	19
42	PJK	0	1	6
43	Pneumonia	0	17	34
44	PPOK	0	2	204
45	PSMBA	0	4	20
46	Skizifrenia	0	109	22
47	SPOT	0	3	32
48	TBC	12	64	52
49	Thypoid	8	25	19
50	Trauma Kapitis	4	20	6
51	Tumor	0	20	30
52	Vertigo	0	16	23

3. Results And Discussion

At this stage, data analysis is carried out so that calculations are processed using Rapidminer software to see the accuracy of the results obtained from calculations using Microsoft Excel and will be equated with the results exposed in the rapidminer software. Based on the data above, the next step will be processed by entering data into the clustering formula using Microsoft Excel and using the K-Means algorithm. Also processed using rapidminer software to group data into three clusters, namely high cluster, low cluster, and medium cluster. To determine the centroid value of the data, it is necessary to make a provision that the required clustering is 3, the cluster determination is divided into 3 parts, namely high clusters (C1) low clusters (C2) and medium clusters (C3). The cluster point value is determined by determining the largest (maximum) and smallest (minimum) values. The value of the cluster point can be seen in table 2 as follows:

Table 2. Initial Centroid

Cluster	Childrens	Young	Old
Cluster 1	0	12	115
Cluster 2	1	1	5
Cluster 3	53	66	0

3.1. Calculating the Distance of Each Data Centroid (Cluster Center)

After the initial centroid center value data is determined, the next step is to calculate the distance of each data to the cluster center. The process of finding the shortest distance in iteration 1 can be seen in the calculation below:

DAbdominal ,c1 = √([[(1-0)]]^2+[[(7-12)]]^2 [[+(5-115)]]^2)
= 110.118118

DAbdominal ,c2 = √([[(1-1)]]^2+[[(7-1)]]^2 [[+(5-5)]]^2)
= 6

DAbdominal ,c3 = √([[(1-53)]]^2+[[(7-66)]]^2 [[+(5-0)]]^2)
= 78.803553

The table of the shortest distance from the centroid is as follows can be seen in table 3 as follows:

Table 3. Calculation Data of the K-Means Algorithm Results of the 1st Iteration

No	Type Disease	Childrens	Young	Old	C1	C2	C3	Nearest Distance
1	Abdominal Pain	1	7	5	110.11812	6	78.80355	6
2	Anemia	6	16	35	81.437092	33.91165	77.03246	33.91165

3	Anxietas	0	3	26	89.005618	21.11871	86.33655	21.11871
4	Anxiety	0	5	6	109.04128	4.242641	81.03086	4.242641
5	Asthma	2	8	4	111.18003	7.141428	77.33693	7.141428
6	BPH	0	0	37	78.025637	32.03123	92.37965	32.03123
7	Brocnhitis	13	14	29	87.800911	29.8161	71.72866	29.8161
8	Cephalgia	0	10	15	100.31949	13.49074	78.54935	13.49074
9	CHF	0	12	115	10	110.5532	137.659	10
10	Cholic Abdomen	1	1	5	110.00909	0	83.39065	0
....	…….	…..	…..	…..	……	…….	……..	……..
51	Tumor	0	20	30	86.884981	31.41656	76.32169	31.41656
52	Vertigo	0	16	23	93.059121	23.45208	76.40681	23.45208

3.2. Determining the Cluster Center or Grouping

Based on the distance of the data to the center of the cluster. The data that has the smallest distance from the centroid will be a member of the group. Seen in table 4 where the position of the data with each cluster in the 1st iteration.

Table 4. Results of clustering in the 1st iteration

C1	C2	C3
	1	
	1	
	1	
	1	
	1	
	1	
	1	
	1	
1		
	1	
…	…….	…….
	1	
7	**40**	**5**

The K-Means process will continue to iterate until the data grouping is the same as the previous iteration data grouping. In other words, the process will continue to iterate until the data for the last iteration is the same as before.

3.3. Calculate the new center point using

The results of each member in each cluster. Calculation of the center point on c1, c2, and c3. To determine C1 as follows:

$$DCh,c1 = \frac{0+7+0+0+1+1+0}{7} = 1.2857$$

$$DYg,c1 = \frac{\begin{matrix}1+6+0+0+2+0+13+0+1+0+0+4+0+4+0+3+1+0+0+0+1\\+2+7+0+3+0+0+2+25+2+0+0+0+0+0+\\0+8+4+0+0\end{matrix}}{40} = 2.225$$

$$DOld,c1 = \frac{53+27+0+0+12}{5} = 18.4$$

To determine C2 as follows:

$$DCh,c2 = \frac{12+117+36+4+62+32+2}{6} = 37$$

$$D_{Yg,c2} = \frac{\begin{matrix}7+16+3+5+8+0+14+10+1+12+14+6+34+15+6+3+17+3+0+\\ +0+11+3+11+3+1+4+18+12+6+2+1+1+17+4+\\ 3+25+20+20+16\end{matrix}}{40}$$

$$= 8.95$$

$$D_{Old,c2} = \frac{66+36+80+109+64}{5}$$

$$= 71$$

To determine C3 is as follows:

$$D_{Ch,c3} = \frac{115+188+0+113+132+227+204}{7}$$

$$= 149.86$$

$$D_{Yg,c3} = \frac{\begin{matrix}5+35+26+6+4+37+29+15+5+2+2+5+0+8+16+1+11+4+18+\\ +13+13+2+7+16+4+15+3+30+8+8+12++19+6+34+\\ 20+32+19+6+30+23\end{matrix}}{40}$$

$$= 13.7$$

$$D_{Old,c3} = \frac{0+19+0+22+52}{5}$$

$$= 16.6$$

Then the new data centroid for the 2nd iteration is as follows:

Table 5. New Centroid 1st iteration

Cluster	Childrens	Young	Old
Cluster 1	1.2857	37	149.56
Cluster 2	2.225	8.95	13.7
Clustet 3	18.4	71	16.6

Then, steps 4 to 6 are repeated. If the centroid values are not the same or not balanced and the data position is still changing, the iteration process continues in the next iteration. However, if the centroid value is the same as the previous iteration, it is optimal and the position of the data cluster does not change again, then the iteration process stops or is complete. So that the results obtained in the 4th iteration are as follows:

Table 6. Results of the 4th Iteration

No	Type Disease	Childrens	Young	Old	C1	C2	C3	Nearest Distance
1	Abdominal Pain	1	7	5	161.208457	10.46219221	67.32993391	10.46219221
2	Anemia	6	16	35	130.153115	21.29110435	59.3069979	21.29110435
3	Anxietas	0	3	26	141.609204	12.86302709	71.06982482	12.86302709
4	Anxiety	0	5	6	160.633641	10.36852189	69.33195511	10.36852189
5	Asthma	2	8	4	162.001543	11.1953345	66.30776727	11.1953345
6	BPH	0	0	37	131.819953	23.96020893	76.1296263	23.96020893
7	Brocnhitis	13	14	29	136.824096	18.18251378	58.58259127	18.18251378
8	Cephalgia	0	10	15	150.826059	2.26185817	63.73476288	2.26185817
9	CHF	0	12	115	54.8315603	99.95874223	116.1986231	54.83156026
10	Cholic Abdomen	1	1	5	162.475126	13.22945356	73.05696408	13.22945356
....	…….	…..	…..	…..	……	…….	……..	……..
51	Tumor	0	20	30	134.406721	18.38262442	55.84908236	18.38262442
52	Vertigo	0	16	23	141.914881	10.48780488	58.34826476	10.48780488

cluster of each data based on the distance of the data to the center of the cluster. The data that has the smallest distance from the centroid will be a member of the group. Seen in table 7 where the position of the data with each cluster in the 4th iteration.

Table 7. Clusters in the 4th iteration

Cluster		
C1	**C2**	**C3**
	1	
	1	
	1	
	1	
	1	
	1	
	1	
	1	
1		
	1	
.....	.....	.....
	1	
6	**42**	**4**

Manual calculations on the data above get the final results where in the 3rd iteration and 4th iteration the data grouping performed on 3 clusters gets the same results. The results of the two iterations are C1 = 6, C2 = 42 and C3 = 4 at the data position of each cluster, the iteration process stops until the 4th iteration.

Conclusion

The results obtained from the K-Means Clustering method which is implemented into RapidMiner have the same validation value, which produces several clusters 1 there are 6 of them, namely the most common disease (high) CHF, Dyspepsia, Psychotic Disorders, GERD, Hypertension, COPD, for cluster 2 diseases that are rare (low) namely Abdominal Pain, Anemia, Anxiety, Anxiety, Asthma, BPH, Bronchitis, Cephalgia, Cholic Abdomen, CKR, Cytitis, DHF, Depression, DHF, Diabetes, Diarrhea, Pleural Effusion, Fatigue, Disorders Anxiety, Panic Disorder, Gastritis, Gastro Entritis, GEA, Gout Athritis, Hepatitis, Hepoglekimia, HUV, UTI, ARI, Constipation, Myalgia, Nephrolithiasis, OA Gonue, CHD, Pneunomia, PSMBA, SPOT, Typhoid, Trauma Capatic, Tumor, Vertigo . As for cluster 3 diseases that often occur (moderately) namely Epilepsy, Pregnancy, Schizophrenia, TBC. The results obtained from the research can be input to the Government in making policies and in the future it can be of more concern to reduce the level of disease spread in the region.

Acknowledgement

Acknowledgments to the supervisors and examiners who are lecturers at AMIK and STIKOM Tunas Bangsa so that this research can be arranged as one of the requirements for completing Bachelor's education (S1) at STIKOM Tunas Bangsa. I hope this research can be a reference for other research related to the methods and algorithms used. I hope for constructive suggestions for the readers for the perfection of this research in the future.

References

[1] I. Conference, S. Science, and M. Program, "(IN ISNTALASI EMERGENCY (IGD) (Study of Policy Implementation based on Regional Regulation Number 02 of 2014 concerning minimum service standards at Meloy Sangatta Hospital , East Kutai Regency) Hartati," pp. 1–5, 2020.

[2] A. L. Hartzler, L. Tuzzio, C. Hsu, and E. H. Wagner, "Roles and Functions of Community Health Workers in Primary Care," *Ann. Fam. Med.*, vol. 16, no. 3, pp. 240–245, May 2018, doi: 10.1370/afm.2208.

[3] N. Lunt, R. Smith, M. Exworthy, T. Stephen, D. Horsfall, and R. Mannion, "Medical Tourism : Treatments , Markets and Health System Implications : scoping review," *Dir. Employment, Labour Soc. Aff.*, pp. 1–55, 2011.

[4] A. A. Aprilia Lestari, "Increasing Accuracy of C4 . 5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 32–38, 2020.

[5] A. B. U. Nájera and J. de la Calleja Mora, "Brief review of educational applications using data mining and machine learning," *Rev. Electron. Investig. Educ.*, vol. 19, no. 4, pp. 84–96, 2017, doi: 10.24320/redie.2017.19.4.1305.

[6] A. Twin, "Data Mining Data mining," *Min. Massive Datasets*, vol. 2, no. January 2013, pp. 5–20, 2005.

[7] V. Marriboyina and L. C. Reddy, "A Review on Data mining from Past to the Future," *Int. J. Comput. Appl.*, vol. 15, Feb. 2011, doi: 10.5120/1961-2623.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.
[9] A. M. H. Pardede *et al.*, "Smart Health Model with A Linear Integer Programming Approach," 2019, doi: 10.1088/1742-6596/1361/1/012069.
[10] N. A. Khairani and E. Sutoyo, "Application of K-Means Clustering Algorithm for Determination of Fire-Prone Areas Utilizing Hotspots in West Kalimantan Province," *Int. J. Adv. Data Inf. Syst.*, vol. 1, no. 1, pp. 9–16, 2020, doi: 10.25008/ijadis.v1i1.13.
[11] M. Omran, A. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intell. Data Anal.*, vol. 11, pp. 583–605, Nov. 2007, doi: 10.3233/IDA-2007-11602.
[12] S. S. Nagari and L. Inayati, "Implementation of Clustering Using K-Means Method To Determine Nutritional Status," *J. Biometrika dan Kependud.*, vol. 9, no. 1, p. 62, 2020, doi: 10.20473/jbk.v9i1.2020.62-68.
[13] I. D. Nirmala and P. D. Atika, "Implementation of K-Means Algorithm As a Clustering Method for Selecting Achievement Students Based on Academic Grade," *J. Pilar Nusa Mandiri*, no. Ningrum 2009, pp. 199–204, 2020, doi: 10.33480/pilar.v16i2.1575.
[14] A. M. H. Pardede, N. Novriyenni, and L. A. N. Kadim, "Emergency patient health service simulation as a supporter of smart health care," 2020, doi: 10.1088/1757-899X/725/1/012084.