INTERNATIONAL JOURNAL OF HEALTH SCIENCE

Link Page: <u>https://ejurnal.politeknikpratama.ac.id/index.php/ijhs</u> Page: <u>https://ejurnal.politeknikpratama.ac.id/index.php</u>

DATA MINING K-MEANS: CLUSTERING HEALTH AND COMPLAINTS RESIDENT IN INDONESIA

Sri Wulandari^a, Husna Sarirah Husin^b, Wahyu Ratri Sukmaningsih^c
 ^{a.c} Sarjana Terapan Manajemen Informasi Kesehatan
 <u>a.sriwulandari@poltekindonusa.ac.id</u>, Politeknik Indonusa Surakarta
 <u>Information Technology, sarirah@unikl.edu.my</u>, University Kuala Lumpur
 <u>c</u> wahyuratri@poltekindonusa.ac.id, Politeknik Indonusa Surakarta

ABSTRACT

This study aims to utilize the Clustering Algorithm in grouping the population Which have complaint health with algorithm K-means in Indonesia. Source data study This collected based on the information documents. The total population of the province have complaints health produced by the Central Bureau of National Statistics. The data used in this study are data for 2013-2017 which consists of 34 provinces. The method used in this research is K-means algorithm. The data will be processed by clustering in 3 clusters, namely level clusters high health complaints, clusters of moderate and low health complaints. Data center for clusters high population level 37.48, Centroid data for clusters of moderate population level 27.08, and Centroid data forlow population level cluster 14.89. So that the acquisition of the assessment is based on the population index owned health complaints with 7 provinces with high levels of health complaints, namely Central Java, in Yogyakarta, Bali, Nusa Southeast West, Nusa Southeast East, Borneo South, Gorontalo, 18 province level complaint moderate health, and 9 other provinces including low levels of health complaints. It can be input to the government to pay more attention to residents in each area that has high health complaints through improving public health services so that the Indonesian population becomeshealthier without exists complaint health.

Keywords: data mining, complaint health, clustering, k-means, resident Indonesia

1. INTRODUCTION

Health complaints are disturbances to physical and mental conditions, including due to accidents, or other things that interfere with daily activities. In general, the main health complaints are that many residents experience are fever, headache, cough, runny nose, diarrhea, asthma/shortness of breath, toothache. PersonWhich suffer disease chronic considered have complaint health although on time survey (One monthlast) concerned does not cure the disease. There are several efforts that we can do in cope problem complaint health like enhancement nutrition, addition facility health, implementationHowever, provide service health free, procurement drug generic, addition amount power medical, and conduct education on the importance of cleanliness and healthy lifestyle. health service quality, including inside it handling complaint public, influenced by availability means and infrastructure including fund. In matter This expected to government for do evaluation handling complaint Forrepair or increase service health. Constitution Number 36 Year 2009 about Healthstates that health efforts are any activity and/or series of activities carried out collectively integrated, integrated And He For look after And increase degrees health public in the form of disease prevention, health promotion, disease treatment, and health restoration by government and/or public (chapter 1), every person entitled get environment Which Healthy for Warning of health degree (article 6), everyone has the right to get information and education about balanced and responsible health (article 7), everyone has the right to obtain information about his personal health data including actions and treatment that he has received or will receive from the staff health (chapter 8).

Data Which used in study it's based documents Which generated by Body CenterNational Statistics through the site <u>https://</u>www.bps.go.id. In this case the researcher raised the topic of the population have health complaints by province where the method used is clustering. Result ofclusters can be used as input for the government so that provinces that are included in the high cluster get attention more. Process clusters are divided inside 3 (three) clusters namely complaint health tall, complaint healthmoderate and low health complaints. Cluster analysis is a statistical analysis that aims to solve it object inside a number of group

Which have characteristic different between group Which One with Which other.

In this analysis each group is homogeneous between members in the group or object variations in the smallest possible group. Algorithms commonly used in data mining namely algorithm *ID3* [1]–[3], C4.5[1], [4]–[8], K-Means[9], [10][11]–[13], Naive Bayes[14], [15], *Support VectorMachine* [2], *Priori* [16] And a number of algorithm data mining other. Algorithm K-Means can used forstudy various type matter, among them study related [17] who do study about groupingin classifying the number of villages/kelurahan that have health facilities by province with k- method. In order to obtain an assessment based on the Village/Kelurahan Index with Health Facilities with 4 provinces with high level of health facilities namely North Sumatra, West Java, Central Java, East Java, 14 provinces have a moderate level of health facilities, and 16 other provinces have a low level of health facilities. Rohmawati[18] grouped data on Student Learning Assistance (BBM) scholarship applicants, for measurement algorithm performance, this measurement is seen from the cluster results by calculating the purity value (purity measure) from each – each group which in earn.

2. LITERATURE REVIEW

2.1. Data Mining

Data mining is a data processing method to find hidden patterns from data the. The results of data processing with this data mining method can be used to make decisions the future. Data mining is also known as pattern recognition [19][12]. Data mining is large-scale data processing methods therefore data mining has an important role in the field industry, finance, weather, knowledge And technology. kindly general study data mining discuss methods like, clustering, classification, regression, selection variable, and market basket analysis [19].

2.2 Clustering

Clustering / Clustering analysis is the process of dividing data into a set into several groups that have the same data in a group that is larger than the similarity of the data with data in other groups. Potential grouping can be used to determine the inner structure data that can be further used in a wide variety of applications such as classification, image processing, And introduction pattern[17].

2.3 K-Means

K-Means is something algorithm Which used in grouping in a manner petition Which Finally data to in group Which different – different [20].

The K-Means algorithm is a method used to group data based on an iterative search for cluster centers where the location of each data cluster is found to be the minimum distance for each data to the cluster center [8]. The K-Means algorithm is included in the nonhierarchical method which partitions data into one or more clusters, so that data with the same characteristics are grouped in the same cluster and data with different characteristics are grouped into other clusters [3]. The K-Means algorithm is in their research to increase the speed of the clustering method by making centroid determination simpler [11]. K-means has several distinct advantages over other clustering algorithms. That is, K-means is an algorithm that is very simple and powerful, very efficient, and can be used for various types of data. To calculate the center of the latest cluster using the Euclidean distance formula.

$$d(x,y) = \sqrt{\sum_{i=0}^{i} (xi - yi)^2}$$

Where:

d(x,y) = distance between data at point x and y

x = object data point

y = centroid data point

i = number of dataFormula attribute

The formula for grouping data into clusters with the shortest distance

$$C_k = \left(\frac{1}{nk}\right) \sum di$$

Where:

Ck = cluster centroids

nk = amount of data in cluster K

di = sum of the distance values included in each cluster.

3. ANALYSIS AND DISCUSSION

In study This method Which used that is method Data Mining as following:

- a. Collection stage data,
- b. Processing stage data,
- c. Stage clustering and
- d. Stage analysis

3.1 Stage Collection Data

Collection data in study This taken based on documents information Amount resident province Which own complaint health Which generated from Publication Statistics Indonesia And processed from the results of the National Socioeconomic Survey/Susenas, the Central Bureau of National Statistics. Data used in this research is data from year 2013-2017 Which consists from 34 province.

3.2 Stage Processing Data

The data that has been processed will be processed first to be clustered. In the previous stage, the data for each province will be added up for each aspect so that at this stage the value calculation has been obtained which will process on stage clustering.

3.3 Stage Clustering

Partition clustering is an unattended classification and is a process of collecting data objects from One arrange become a number of class. Matter This can done with apply various buy and sell And step-steps regarding the distance of the algorithm, namely the Euclidean Distance [21]. Cluster analysis is a method is used to divide a data set into several groups based on the similarities that have been found previously determined [22]. In determining the cluster based on the data that is already available, it takes a flowchart to make it easier to determine the flow of calculations as a flow to find the results of application of clusters to the data to be processed. The following is a flowchart in determining clusters with K-Mean [15].

3.4 Stage Analysis

At this stage data analysis is carried out on the number of residents who have health complaints. data that obtained processed using the calculation of the weight of each index. In the previous stage, it has determined will clustered to in 3 group ie group level complaint health tall, group level complainthealth medium and group level complaint health low. On it is this stage will analyzed result.

In *clustering*, the data obtained will be calculated first based on the average results Number of population experiencing health complaints by Province. The data is then accumulated And taken mark average his, that is:



Figure 1. Mark average after alternative data accumulated

59

After being accumulated and looking for the average value in Figures 1 above, the value of each variable will be obtained. Then the data will enter the clustering stage by applying the K-means algorithm to group data into three groups.

3.5 Centroids Data

In application algorithm *K-means* generated mark point middle or *center of mass* from data Which got with The condition that the desired clusterization is 3. Determination of this cluster point is done by taking the largest (maximum) value for the high cluster (C1), the average (average) value for the medium cluster (C2) and the smallest (minimum) for clusters low (C3). Mark point the can is known on table following.

3.6 Clustering Data

By using these *centroids*, the data that has been obtained can be clustered into 3 clusters. Cluster process by taking the shortest distance from each processed data. From the amount data residents who experience complaint health obtained grouping on iteration 1 For 3 group the. Group residenthigh (C1) namely Central Java, DI Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, Kalimantan South And Gorontalo. Group moderate population (C2) ie 22 Province And group low population (C3) ie5 province other. Process search distance short, grouping data on iteration 1 And Top grouping candepicted on the table And picture following:

PROVINCE	AVERAGE	E C1 C2 C3		DISTANCE SHORTEST	
ACEH	27,61	9.87	0.53	12.72	0.53
SUMATERA NORTH	22.86	14,62	4,23	7.97	4,23
SUMATERA WEST	29,67	7,81	2.59	14.78	2.59
RIAU	27.02	10.46	0.06	12,13	0.06
JAMBI	22,26	15,21	4.82	7,38	4.82
SUMATERA SOUTH	26.8	10.67	0.28	11.92	0.28
BENGKULU	27,69	9.79	0.6	12.8	0.6
LAMPUNG	28.98	8.5	1.9	14.09	1.9
KEP, BANGKA BELITUNG	28.59	8.88	1.51	13.71	1.51
KEP, RIAU	22,42	15.06	4.66	7,53	4.66
DKI JAKARTA	30,22	7,26	3,14	15,33	3,14
JAVA WEST	28,34	9,14	1.26	13.45	1.26
Central Java	33,30	4,18	6,21	18,41	4,18
YOGYAKARTA	37,48	0	10,40	22.59	0
JAVA EAST	30,52	6.96	3,44	15,63	3,44
BANTEN	28.94	8.54	1.86	14.05	1.86
BALI	33,30	4,18	6,22	18,41	4,18
NUSA SOUTHEAST WEST	34,45	3.03	7,37	19.56	3.03
NUSA SOUTHEAST	34,64	2.83	7.56	19.76	2.83
KALIMANTAN WEST	25,22	12,26	1.86	10.33	1.86
CENTRAL KALIMANTAN	25,41	12.07	1.68	10.52	1.68
KALIMANTAN SOUTH	35.04	2.43	7.96	20,16	2.43
KALIMANTAN EAST	21.59	15.88	5,49	6,71	5,49
KALIMANTAN NORTH	14.89	22.59	12,19	0	0
SULAWESI NORTH	26.02	11.46	1.07	11,13	1.07
CENTRAL SULAWESI	29,17	8,31	2.08	14,28	2.08
SULAWESI SOUTH	24.68	12.8	2.4	9.79	2.4
SULAWESI SOUTHEAST	26.75	10.72	0.33	11.87	0.33
GORONTALO	35,84	1.64	8.76	20.95	1.64
SULAWESI WEST	28.85	8,63	1.77	13.96	1.77
MALUKU	19.05	18,42	8.03	4,17	4,17
MALUKU NORTH	16.04	21.44	11.04	1.15	1.15
PAPUA WEST	20	17,48	7.09	5,11	5,11

Table 1. calculation Distance Center Group Iteration 1

International Journal Of Health Science Vol.1, No.3, November 2021, pp. 57-63

PROVINCE	AVERAGE	C1	C2	C3	DISTANCE SHORTEST
PAPUA	17,68	19.79	9.4	2.8	2.8

PROVINCE	C1	C2	C3	PROVINCE	C1	C2	C3
ACEH		X		NUSA SOUTHEAST WEST	X		
SUMATERA NORTH		X		NUSA SOUTHEAST EAST	X		
SUMATERA WEST		X		KALIMANTAN WEST		Х	
RIAU		X		KALIMANTAN MIDDLE		Х	
JAMBI		X		KALIMANTAN SOUTH	X		
SUMATERA SOUTH		X		KALIMANTAN EAST		Х	
BENGKULU		X		KALIMANTAN NORTH			Х
LAMPUNG		X		SULAWESI NORTH		Х	
KEP, BBANG BELITUNG		X		SULAWESI MIDDLE		Х	
KEP, RIAU		X		SULAWESI SOUTH		Х	
DKI JAKARTA		X		SULAWESI SOUTHEAST		Х	
JAVA WEST		X		GORONTALO	X		
JAVA MIDDLE	X			SULAWESI WEST		Х	
IN YOGYAKARTA	X			MALUKU			Х
JAVA EAST		X		MALUKU NORTH			Х
BANTEN		X		PAPUA WEST			Х
BALI	X			PAPUA			Х

Table 2. Iteration 1 grouping results

K-means process will continue to iterate until the data grouping is the same as the data grouping iteration previously. In other words, the process will continue to iterate until the data in the last iteration are equal previous iteration. After getting the midpoint or *centroid value*, the same process is carried out by searching closest distance. The shortest distance search process, data grouping in the last iteration and data clustering can depicted on table following:

 Table 3. Centroids Data Iteration 4

PROVINCE	AVERAGE	C1	C2	C3	DISTANCE			
SUMATERA SOUTH	26.8	10.67	0.28	11.92	0.28			
SULAWESI SOUTHEAST	26.75	10.72	0.33	11.87	0.33			
ACEH	27,61	9.87	0.53	12.72	0.53			

PROVINCE	AVERAGE	C1	C2	C3	DISTANCE SHORTEST
YOGYAKARTA	37,48	0	10,40	22.59	0
KALIMANTAN NORTH	14.89	22.59	12,19	0	0
RIAU	27.02	10.46	0.06	12,13	0.06
BENGKULU	27,69	9.79	0.6	12.8	0.6
SULAWESI NORTH	26.02	11.46	1.07	11,13	1.07
MALUKU NORTH	16.04	21.44	11.04	1.15	1.15
JAVA WEST	28,34	9,14	1.26	13.45	1.26
KEP, BANGKA BELITUNG	28.59	8.88	1.51	13.71	1.51
GORONTALO	35,84	1.64	8.76	20.95	1.64
CENTRAL KALIMANTAN	25,41	12.07	1.68	10.52	1.68
SULAWESI WEST	28.85	8,63	1.77	13.96	1.77
BANTEN	28.94	8.54	1.86	14.05	1.86
KALIMANTAN WEST	25,22	12,26	1.86	10.33	1.86
LAMPUNG	28.98	8.5	1.9	14.09	1.9
CENTRAL SULAWESI	29,17	8,31	2.08	14,28	2.08
SULAWESI SOUTH	24.68	12.8	2.4	9.79	2.4
KALIMANTAN SOUTH	35.04	2.43	7.96	20,16	2.43

Table 4. calculation Distance Center Group Iteration

PROVINCE	AVERAGE	C1	C2	C3	DISTANCE SHORTEST
SUMATERA WEST	29,67	7,81	2.59	14.78	2.59
PAPUA	17,68	19.79	9.4	2.8	2.8
NUSA SOUTHEAST	34,64	2.83	7.56	19.76	2.83
NUSA SOUTHEAST WEST	34,45	3.03	7,37	19.56	3.03
KEP, RIAU	22,42	15.06	4.66	7,53	4.66
JAMBI	22,26	15,21	4.82	7,38	4.82
DKI JAKARTA	30,22	7,26	3,14	15,33	3,14
JAVA EAST	30,52	6.96	3,44	15,63	3,44
MALUKU	19.05	18,42	8.03	4,17	4,17
Central Java	33,30	4,18	6,21	18,41	4,18
BALI	33,30	4,18	6,22	18,41	4,18
SUMATERA NORTH	22.86	14,62	4,23	7.97	4,23
PAPUA WEST	20	17,48	7.09	5,11	5,11
KALIMANTAN EAST	21.59	15.88	5,49	6,71	5,49

3.7 Analysis Data

On iteration 4, grouping data Which done to 4 group with iteration 1 obtained results WhichThe same. Of the 34 known province-based population data, 7 provinces are high-level clusters 18 province group level medium and 9 province other including clusters level low.

4. CONCLUSION

To carry out an assessment of the results of population health complaints by province, you can apply the K-Means clustering method. The data is processed to obtain values from residents with complaints health by province. The data is processed using Ms. Excel to determine the inner centroid value 3 group that is group level complaint health tall, group level complaint health currently And group levellow health complaints. Cetroid data for clusters of high health complaints 37.48, Cetroid data for group level complaint health currently 27.08 And Cetroid data For group level complaint health low 14.89. In order to obtain an assessment based on the population index that has health complaints of 7 provinces with high levels of health complaints, namely Central Java, in Yogyakarta, Bali, West Nusa Tenggara, Nusa East Southeast, South Kalimantan, Gorontalo, 18 provinces with moderate health complaints, and 9 provinces others include a low level of health complaints. The results obtained from research can be input to government, province Which become attention more on resident Which own complaint health tallbased on group Which has done.

5. REFERENCE

- P. PDAM and K. Kendal, "A Comparative Analysis of ID3 and C4 Algorithms. 5 For Recipient Classification Grant Installation Air Drink," vol. 1, NO. 3, matter. 234–244, 1978.
- [2] D. Ispriyanti And A hoyyi, "ANALYSIS CLASSIFICATION TIME STUDIES STUDENT STUDY PROGRAM STATISTICS UNDIP with METHOD SUPPORT VECTOR MACHINE (SVM) And ID3 (ITERATIVEDICHOTOMISER 3)," *Media Stats.*, vol. 9, NO. 1, matter. 15–29, 2016.
- [3] P Soepomo, "APPLICATION DATA MINING FOR CLASSIFICATION PREDICTION DISEASE ISPA (Infection Channel Respiration I) WITH ALGORITHM DECISION TREE (ID3)," vol. 2, 2014.
- [4] P. Algorithm, C. For, K. Predicate, M. Faculty, K. Dan, and I. University, "Implementation of the c4.5 algorithm For classification predicate graduates student faculty communication And informatics university muhammadiyah surakarta" NO. February, 2016.
- [5] RH Pambudi and BD Setiawan, "Application of the C4 Algorithm. 5 To Predict Graduation Grades Student School Currently Based on Factor external," *J. Development Technol. inf. And Knowledge computer.*, vol.2, NO. 7, matter. 2637–2643, 2018.
- [6] S. Case and S. Pringsewu, "ANALYSIS OF STUDENT SATISFACTION WITH SERVICES ACADEMIC USE METHOD ALGORITHM C4 . 5", vol. 2, NO. 1, matter. 1–11, 2016.
- [7] L. Swastina, "Application Algorithm C4.5 For hole Major Student," *Gems actual.*, vol. 2, NO.1, P 6,

2013.

- [8] FF Harryanto, S. Hansun, UM Nusantara, G. Serpong, and C. Employees, "Application of the C4 Algorithm. 5 For predict Reception Candidate Employee New at PT WISE," vol. 3, NO. 2, matter. 95–103, 2017.
- [9] GF Mandias *et al.*, "Application of the K-Means Algorithm for Analysis of Student Academic Achievement Faculty Knowledge Computer University Klabat Application from K-Means Algorithm For Academic Achievement Analysis from Faculty from Computer Science University Clabat," p. 230–239, 2017.
- [10] I. Print, I. Online, H. Sulastri, and AI Gufroni, "Journal of Technology and Information Systems DATA DEEP MINING GROUPING Sufferers," vol. 2, matter. 299–305, 2017.
- [11] A P Windarto, "Application Data Mining On Export Fruits According to Country Objective Using K-Means Clustering," *Techno.COM*, vol. 16, NO. 4, matter. 348–357, 2017.
- [12] AP Windarto, "Implementation of Rice Import Data Mining Based on Main Country of Origin Using Algorithm Use K-Means Clustering Method," *int. J. artificial Intell. Res.*, vol. 1, no. 2, p. 26–33, 2017.
- [13] MG Sadewo, AP Windarto, and D. Hartama, "IMPLEMENTATION OF DATAMINING IN POPULATIONS MEAT CHICKEN RACE MEAT IN INDONESIA BASED ON PROVINCE USE K-MEANS CLUSTERING," *InfoTekJar (Journal Nas. Tell. And Technol. network)*, vol. 2, NO. 1, matter. 60–67, 2017.
- [14] H. Muhammad, R. Sudrajat, and R. Rosadi, "Application of Data Mining in Analyzing Nursing User Media Social With Naive Bayes Classifier : Studies Case," J. Tell. Univ. Padjadjaran, vol.1, NO. 1, P 5258, 2016.
- [15] M. Ridwan, H. Suyono, and M. Sarosa, "Application of Data Mining for Academic Performance Evaluation Student Use Algorithm Naive Bayes Classifier," *Eeccis*, vol. 7, NO. 1, matter. 59–64, 2013.
- [16] H. Kurniawan And J. S. Information, "Application datamining For predict backup level studentuse a priori algorithm in aunt darmajaya city Lampung" vol. 2, NO. 1, matter. 79–93, 2016.
- [17] M. G. Sadewo, A P Windarto, And S. R. Andani, "Utilization Algorithm Grouping In Grouping Amount Village / Ward Which Own Means Health," vol. I, matter. 124–131, 2017.
- [18] Mohammed jajuli nurul rohmawati, sofi challenge, "Implementation Algorithm K-Means In ClusteringStudent Applicants Scholarship," *Jitter 2015*, vol. I, NO. 2, matter. 62–68, 2015.
- [19] JO Ong, "Implementation of the K-means clustering algorithm to determine the president's marketing strategyUniversity," J. Science. Tech. eng., vol. vol.12, NO, NO. June, matter. 10–20, 2013.
- [20] A K. Wardhani, "Implementation Algorithms K-Means For Disease Grouping Patient on Public health center Kajen Pekalongan," *J. Change.*, vol. 14, NO. 1, matter. 30–37, 2016.
- [21] B. Venkateswarlu And PG S. V. P Raju, "Mine Blood Donor Information through Upgraded K-MethodClustering."
- [22] e. Sugiharti and M. A Muslim, "On line clustering of lecturer performance computer science department from Semarang country University use K-Means Algorithm," J. Theor. Application inf. Technol., vol. 83,NO. 1, matter. 64–71, 2016