Information System for Analysis of the Need of Doctors Using K-Means Clustering and Chi-Square Test

Nova Christina Sari[1,*], *Retno* Kusumaningrum[2], and *Suryono* Suryono[3]

[1] Magister Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang - Indonesia
[2] Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang – Indonesia
[3] Department of Physics, Faculty of Sciences and Mathematics, Diponegoro University, Semarang - Indonesia

Abstract. Doctor is one of the medical staff who is needed by the community in implementing health in Indonesia. Some provinces in Indonesia have different populations and doctors, an analysis is needed to maximize the number of doctors to the population. In this research, the K-means method was used to divide the cluster into the number of doctors needed and the chi-square method to final testing for the result of clustering. This research will provide results on areas that need to increase the number of doctors or do not need to add additional doctors in Indonesia. Using of the K-means clustering method and chi-square test for doctor analysis, giving results of accuracy which is 78%.

Keywords: **Doctor; K-means clustering; chi-square; analysis.**

1 Introduction

The uneven distribution of medical personnel has a detrimental effect on the community, a lack of quality health services that will make people not get maximum health facilities, such as treatment, examinations, and health services. Doctors are one of the medical personnel that is very much needed in public health.

Data mining seen from data processing provides some algorithms that can be used to extract hidden information from multidimensional data sets [1]. The data mining algorithm that is quite popularly used both in the business, academic, industrial and health research is K-means. K-means clustering algorithm is one method of non-hierarchical clustering data, it carries out data into groups data establish from one cluster or more. Data are grouped in one cluster if data has the same characteristics and other that data are grouped with other clusters so data in one cluster have a small level of variation [2].

Clustering is often used to divide some data into groups that have similar values. The results of data in each group are called clusters, clusters consist of several data that are similar in the distribution of data for each cluster judged based on the value of the attribute and value of the object. Calculations used to determine cluster values usually use distance calculations [3].

Chi-square is of the non-parametric methods to compare many people or responses within a category level with a hypothetical or expected value [4]. Chi-square can be used to assess how close a theoretical distribution, normal and binomial or according to empirical distribution, distribution obtained from data [5].

Using the K-means clustering method is to divide the data into several clusters and to simplify managing data. Chi-square method is used to test the results of each cluster, the results of the chi-square test are used to automatically label each cluster.

2 Methods

The data includes the name of the province, population, number of general practitioners, number of specialists, number of dentists and number of specialist dentists. The information system was created from this data and this research also used some method.

2.1 Data Mining

Data mining is a computer-based information system (CBIS) aimed at scanning large repository data, generating information, and finding knowledge. Data Mining also can be used to find out patterns, regulate hidden information relationships, structure association rules, estimate unknown items and clarify values on objects. Data mining techniques are designed to processes identify and interpret data for understanding, concluding, and designing strategies based on these results [6]. The data mining method is becoming popular in the healthcare field, it has an analytical methodology of efficient for detecting valuable information in health data [7].

* Corresponding author: novachristinasari@gmail.com

2.2 K-means Clustering

The clustering methods for segmented data and the most used is the K-means clustering method, also known as 'Forgy's algorithm' [8]. K-means is often used to process a large of data to be representative data, which is called as cluster centers. Processing K-means is to calculate repeated data until the error does not change and give the final result.

The process of K-means is shown in Fig 1, before calculating the distance of each cluster from the center of cluster to the nearest cluster, is to determine the number K and set the center of the cluster based on value of K. The new cluster center is calculated for the distance until the data give the final result [9].

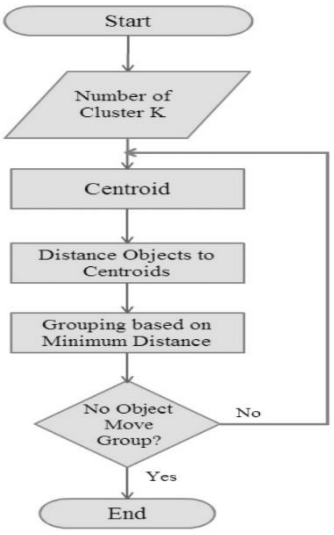

Fig. 1. K-means Clustering Flowchart

Measuring the distance in Euclidean space (space distance) using the formula.

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj})^2} \quad (1)$$

d is the distance between i and k, i is the cluster central data, k is the data on the attribute, ij is data at the center of the cluster and kj is data on each data to i.

2.3 Analysis of the Need of Doctors

Analysis of standard doctor staff can be calculated using the medical personnel analysis formula. The formula for calculating the standard analysis of doctors is:

$$TM = \frac{P}{K_m} \quad (2)$$

TM is the need for medical personnel, p is the area population and K_m is ratio index.

2.4 Chi-Square

Chi-square is a statistical technique used to test hypotheses if in a population consisting of two or more classes where the data is nominal and the sample is large. To analyze the disparity of physicians, it is necessary to use a chi-square test. The formula of chi-square to do a disparity analysis is:

$$x^2 = \sum_{i=i}^{k} \frac{(F_o - F_h)^2}{F_h} \quad (3)$$

χ^2 is value of chi-square, F_o is value of observed and F_h is value of predicted.

2.5 Research Procedures

Fig 2 shows the research procedure, for a detail explanation about each step can be seen on the subsection.

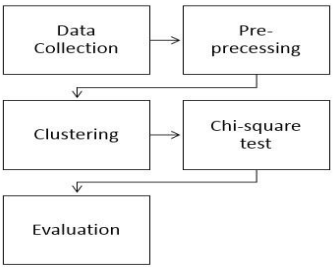

Fig. 2. Research Procedures

2.5.1 Data Collection

Collecting data for this research obtained from:

- Indonesia Central Bureau of Statistics for population data.
- Health Human Resource Information System version 2017, published by the Ministry of Health Republic of Indonesia for the doctor's data.

The research uses the requirement ratio per 100,000 of the total population, the ratio can be seen in Table 1.

Table 1. Requirement Ratio

No	Type of Doctor	Requirement Ratio per 100,000 of Total Population
1	General Practitioner	42
2	Specialist	10
3	Dentist	12
4	Dentis Specialist	10

2.5.2 Pre-processing

Pre-processing needs to normalize data into minimal normalization, data has to range into min to the max because the obtained data has disparity for each type. The disparity data can give influence to the performance of the clustering process [10].

2.5.3 Clustering

The clustering process is using the K-means clustering method. The data will be calculated by the distance of

each data point from the cluster center to the nearest cluster center.

2.5.4 Chi-Square Test

This step is to test the results of each cluster, after using the K-means method. Each cluster was tested using the chi-square method to determine the disparity in the value between real values and the standard value of medical personnel. After computing the data with the chi-square formula, the result from chi-square test will be labeling.

2.5.5 Evaluation

Evaluation is to evaluate the accuracy level of the labeling of Chi-square.

2.6 Framework of Information System

In the information system framework, there are inputs which are databases consisting of the name of the Province, total population, the number of general practitioners, the number of specialists, the number of dentists and the number of dentists specialists. The K-means clustering method, which produces an iteration table containing the results of each cluster, and the chi-square test is calculated last. The results of output the process will be displayed through the chi-square results table which results in being labeled fulfilled or not fulfilled based on each cluster. The system framework can be seen in Fig 3.

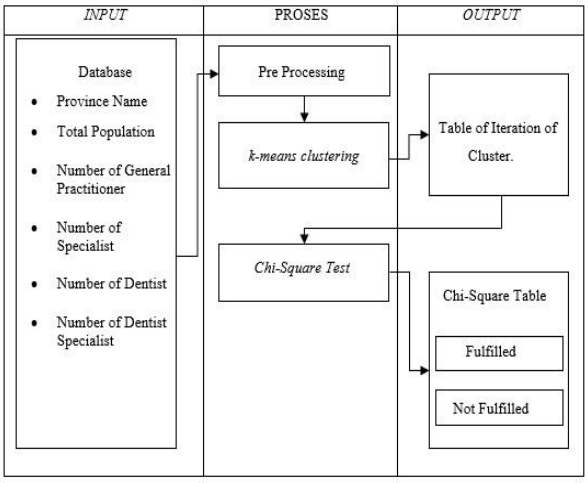

Fig. 3. Framework of Information System for Analysis of the Need of Doctors.

3 Results and Discussion

3.1 Results

The data that has been collected, it will be divided into several clusters. The process uses the K-means clustering method. Table 2 shows the result of K-means clustering, which consists of 5 clusters. The results of cluster 1 consist of 3 provinces, cluster 2 consists of 4 provinces, cluster 3 consists of 13 provinces, cluster 4 consists of 3 provinces and cluster 5 consists of 12 provinces.

Each cluster produces a different centroid value, it will calculate the centroid of each cluster. After calculating the last centroid of each cluster, the results of calculations last centroid are processed by the chi-square test method. The results of the centroid process for each cluster are in Table 3.

Table 2. Clustering Result

Cluster ID	The Number of Population	HHR Availability			
		General Practitioner	Specialist	Dentist	Dentist Specialist
1	36.583.687	32.052	9.572	8.899	1.134
2	31.473.388	12.364	3.241	2.825	136
3	22.713.023	7.724	1.681	1.355	66
4	120.473.636	43.271	12.251	10.436	1.324
5	47.461.252	21.421	5.469	4.854	404

Table 3. Last Centroid from Clustering Process

Cluster ID	Province Name
1	Sumatera Utara
	DKI Jakarta
	Banten
2	Riau
	Sumatera selatan
	Lampung
	Sulawesi Selatan
3	Bengkulu
	Kepulauan Bangka Belitung
	Kepulauan Riau
	Klaimantan Tengah
	Kalimantan Utara
	Sulawesi Utara
	Sulawesi Tengah
	Sulawesi Tenggara
	Gorontalo
	Sulawesi Barat
	Maluku
	Maluku Utara
	Papua Barat
4	Jawa Barat
	Jawa Tengah
	Jawa Timur
5	Aceh
	Sumatera Barat
	Riau
	Jambi
	DI Yogyakarta
	Bali
	Nusa Tenggara Barat
	Nusa Tenggara Timur
	Kalimantan Barat
	Kalimanatan Selatan
	Kalimantan Timur
	Papua

Chi-square test is used to calculate the last result of the clustering, the result of chi-square test to give the labels automatically to identify the disparity status between the value of available and requirements. The result can be seen in Table 4.

Table 4. Chi-Square Test

	Cluster ID	General Practitioner	Specialist	Dentist	Dentist Specialist
F_o – Availability	1	32.052	9.572	8.899	1.134
	2	12.364	3.241	2.825	136
	3	7.724	1.681	1.355	66
	4	43.271	12.251	10.436	1.324
	5	21.421	5.469	4.854	404
F_o – Requirement	1	15.365	3.804	4.536	3.804
	2	13.219	3.273	3.902	3.273
	3	9.539	2.362	2.816	2.362
	4	50.598	12.528	14.938	12.528
	5	19.933	4.936	5.885	4.936
Expected - Value (F_h)	1	23.708	6.688	6.718	2.469
	2	12.791	3.257	3.364	1.705
	3	8.632	2.022	2.086	1.214
	4	46.934	12.390	12.687	6.926
	5	20.677	5.202	5.369	2.670
Chi-Square Value - (Availability)	1	2.936,31	1.243,40	708,40	722,02
	2	14,27	0,08	86,28	1.443,36
	3	95,44	57,35	255,95	1.085,59
	4	285,95	1,55	399,36	4.531,32
	5	26,76	13,67	49,48	1.922,96
Chi-Square - Requirement	1	2.936,31	1.243,40	708,40	722,02
	2	14,27	0,08	86,28	1.443,36
	3	95,44	57,35	255,95	1.085,59
	4	285,95	1,55	399,36	4.531,32
	5	26,76	13,67	49,48	1.922,96
Chi-Square	1	5.872,62	2.486,80	1.416,80	393,43
	2	28,55	0,16	172,57	70,88
	3	190,88	114,71	511,89	88,57
	4	571,90	3,11	798,72	5,62
	5	53,52	27,34	98,95	5,68

Using 2 categories for availability and requirement, the degree of freedom is (2-1) = 1, and fault tolerance 0.5, which is 3,841 in the Chi-square table. The automatic labeling process was based on the rule of disparity label. The disparity label rules contain word fulfilled if chi-square value < 3,841 and else not fulfilled. The label for each cluster is in Table 5.

Table 5. General Label for Each Cluster

Cluster ID	General Practitioner	Specialist	Dentist	Dentist Specialist
1	Not Fulfilled	Not Fulfilled	Not Fulfilled	Not Fulfilled
2	Not Fulfilled	Fulfilled	Not Fulfilled	Not Fulfilled
3	Not Fulfilled	Not Fulfilled	Not Fulfilled	Not Fulfilled
4	Not Fulfilled	Fulfilled	Not Fulfilled	Not Fulfilled
5	Not Fulfilled	Not Fulfilled	Not Fulfilled	Not Fulfilled

Based on Table 2 and Table 5, two clusters have a good disparity label, fulfilled status for the specialist, is cluster 2 and cluster 4. The evaluation process was done to measure the accuracy of the labeling process. The processed labeling is based on the ratio between the number of error result and the number of accuracy result. The result shows that the average of labeling accuracy is about 78%.

Table 6. Result of Performance Evaluation

No	Cluster ID	Error Result	Accuracy Result
1	1	25%	75%
2	2	25%	75%
3	3	17%	83%
4	4	33%	67%
5	5	8%	92%
Average		22%	78%

Table 6 shows the average of labeling accuracy is about 78% and the average error result is 22%. The error result from cluster 5 is the smallest than other clusters.

3.2 Discussion

The development of the industrial revolution 4.0 influences in the health sector, therefore this research can help to distribute the doctors evenly in Indonesia. If the number of doctors is fulfilled for each province in Indonesia then the development of technology in the health sector will increase. It can make it easier for doctors to analyze and treat patients and the community can get services from health facilities that are better than before.

4 Conclusion

The K-means method is used to cluster data in clusters. From the results of the application that has been used, the K-means clustering method used for the group of the doctor's data. The K-means clustering method has an accuracy of 78%, judging from the results of 5 clusters, provinces that require the addition of doctors are in cluster 5 which has the smallest number of errors, which is 8%. In clusters 2 and 4 have fulfilled status for specialists, which means that provinces in cluster 2 and cluster 4 have good disparities for specialist.

References

1. P-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Pearson, New York (2005)
2. M.E. Celebi, H.A. Kingravi, P.A. Vela, *A comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm*, Expert System With Application **40**, 200-210 (2013).
3. D.T. Nguyen, G.T. Nguyen, V.T.N Lam, *An Approach to Data Mining in Healthcare: Improved*

K-means Algorithm, Journal of Industrial and Intelligent Information **1,** 14-18 *(*2013).

4. C.J.Spencer, C. Yakymchuk, M. Ghaznavi, *Visualising Data Distributions with Kernel Density Estimation and Reduced Chi-Squared Statistic*, Journal of Geoscience Frontiers **8,** 1247-1252 (2017)
5. C.S. Withers, S. Nadarajah, *Expansion for the Distribution of Asymptotically Chi-square Statistic*, Statistical Methodology **12,** 16-30 (2013).
6. I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, San Fransisco (2005).
7. S. R, S. R, A.N. R, *A Survey of Health Care Prediction Using Data Mining*, Journal of Innovative Research in Science, Engineering and Technology **05,** 14538-14543 (2016)
8. A.K. Jain, *Data Clustering: 50 Years beyond K-Means.* Pattern Recogn. Lett **31,** 651–666 (2010)
9. C. Chin-Hsing, H. Wen-Tzeng, T. Tan-Hsu, C. Cheng-Chun, C. Yuan-Jen, *Using K-Nearest Neighbor Classification to Diagnose Abnormal Lung Sounds,* Jurnal Sensor **15,** 13132-13158 (2015)
10. R. Kusumaningrum, Farikhin, *An Automatic Labeling of K-means Clusters based on ChiSquare Value*, Journal of Physics: Conference Series **801** (2017)