Pillar Algorithm in K-Means Method for Identification Health Human Resources Availability Profile in Central Java

M. Nishom[1], Sharfina Febbi Handayani[2], Dairoh[3]

[1,2,3]*Politeknik Harapan Bersama, Tegal, Indonesia*

[1]nishom@poltektegal.ac.id, [2]sharfina.handayani@poltektegal.ac.id, [3]dairoh@poltektegal.ac.id

Abstract - Based on data from the Ministry of Health, the distribution ratio between health workers and patients in Indonesia is still not equal distributed. It influenced by the distribution of health human resources that are not in accordance with the ideal needs of health services. This results need to identify the profile of the availability of health human resources in Indonesia. In this study, an approach will be implemented to identify the profile of health human resources availability using K-Means Clustering with a combination of pillar algorithms in optimizing the selection of the initial cluster centroid. Chi-square analysis is used to determine the disparity in the needs of health human resources with the conditions of the availability of health human resources in the Central Java region. The data collection method used in this research is the observation method, while the scientific method used in this research is the K-Means Clustering method. The results showed that the application has been generated can dynamically determine the health human resource cluster based on the disparity category of health human resource availability in the Central Java region. In addition, the labeling of the Pillar K-Means cluster based on the Chi-square test has a high degree of accuracy, namely 80%.

Keywords: clustering, pillar algorithm, K-Means, health human resource

I. INTRODUCTION

The availability of adequate health human resources is one of the important factors in health development efforts in Indonesia, both in quality and quantity. However, the availability of health human resources in Indonesia is faced with two main problems, namely meeting the needs of health workers that are not in accordance with regional needs and the unequal distribution of health human resource [1]. Based on Indonesian Minister Health Regulation No. 81 / MENKES / SK / I / 2004 concerning Guidelines for the Preparation of Human Resources Planning for Health at the Provincial, District / City Levels and hospitals, namely health need methods, health service demand methods, health service target methods, and ratio methods [2]. One method of compiling the human resource requirements for health is determined using the ratio of personnel to a certain value, such as population, hospital beds and others. According to WHO standards, the maximum ratio of doctors to a population is 1: 1000 [3], but the results of data processing by the Indonesian Ministry of Health and the National Socio-Economic Survey (SUSENAS) of the Central Bureau of Statistics in 2019 show one doctor in an area must serve more than 1000 people [4]. Therefore it is necessary to have a grouping of health personnel profiles which aims to identify disparities in the needs of heath human resources compared to the real conditions of the availability of health human resources in Central Java Indonesia.

Clustering is a method or method commonly used to group data sets into a cluster, looking for and grouping data that have similarities or similarities between one data and another in a dataset. The nature of this method is unsupervised or without direction (meaning that this method is implemented without any training or training and no teacher) and does not require a target or output target. The purpose of clustering is to create clusters that are internally coherent, but distinct from one another. In other words, the data in a cluster must be as similar as possible and in one cluster must be as different as possible from the data in other clusters [5]. One of the most widely used data mining algorithms in research is K-Means[6], this is because the K-Means algorithm has the ability to classify large amounts of data with relatively fast and efficient computing time [7]. The K-Means algorithm is an algorithm that can be used to group data into several clusters based on the level of similarity between the data. This algorithm is popular because it is easy to implement in various types of cases and data attributes [8]. The use of cluster analysis can be used to classify data on the availability of health human resources in an area into a number of clusters according to the level of data similarity. This algorithmic process of idea or flow is quite simple. In the initial stage, first the number of groups or clusters that will be used is determined. Then proceed by selecting the first document or first element in a cluster to be used as the cluster center point. Then iteration or repetition of the steps in determining the distance of the document or object to the centroid, until stability occurs and all object groups have converged [9].

The pillar algorithm is able to optimize the selection of the initial centroid and increase the accuracy of the segmentation process in the K-means algorithm. The pillar algorithm is also able to handle outliers with an outlier detection mechanism. In addition, the execution time of the pillar algorithm also shows better performance than other early centroid optimization algorithms [7-8]. Previous research has proposed a new approach to solve the determination of the optimal starting point for K-Means, namely by optimizing the initial centroid for K-Means by spreading the initial center point in the feature space so that the distance between the centroids can be as far as possible. The results show that the proposed method can improve the initial centroid performance for K-Means by 60.2%, increasing the closeness of the initial centroid, which means that the proposed approach can produce a closer initial centroid than random initialization [10].

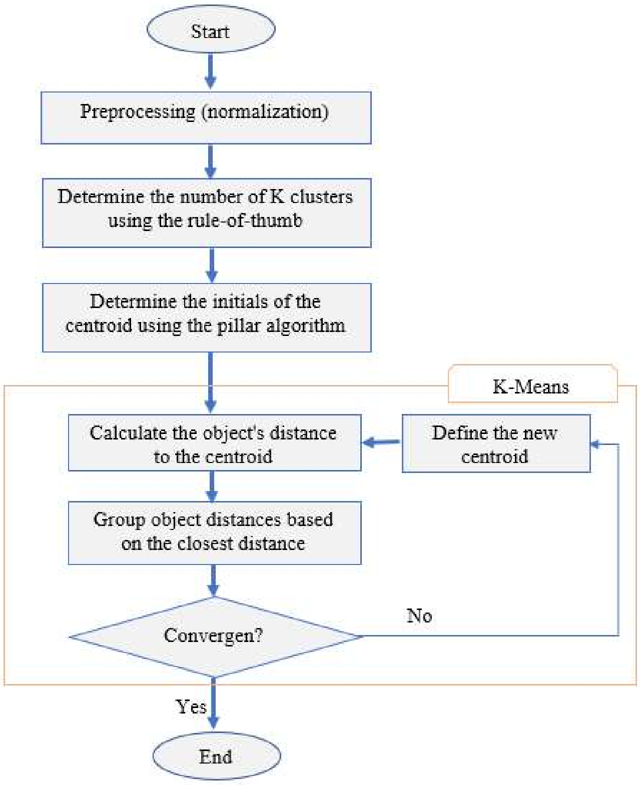

Fig. 1 Flowchart of pillar algorithm and K-Means clustering

II. METHOD

This study uses combination of the pillar algorithm and K-Means algorithm which aims to obtain information related to the results of the centroid determination in the pillar algorithm by grouping data against each cluster center point obtained from the results of the pillar algorithm using the K-Means algorithm (Fig. 1). The Chi-square test method was used to test the level of accuracy and to determine the homogeneity of the cluster. This study used secondary data obtained from Human Resources Development and Empowerment Directorate of the Ministry of Health of the Republic of Indonesia.

First, before entering the clustering process, an analysis of the existing data types is needed, whether the data needs to be normalized or not. The normalization that can be used is the Min-Max normalization [11]. This method is implemented by changing the original value or data to a linear form using the (1):

$$x' = \frac{x - nilai\ min}{nilai\ \max - nilai\ min} \tag{1}$$

where, x = data per column, the value $nilai\ min$ = the smallest value of data per column, the value $nilai\ max$ = the largest value of data per column.

Second, determine the number of clusters (K). This process initializes the initial value K as the number of clusters to be dynamically partitioned [12]. Determination of the amount of K is done using the rule-of-thumb approach using (2):

$$k = \sqrt{\frac{n}{2}} \tag{2}$$

where, n = the number of objects to be grouped and k = the number of clusters.

Third, determine the initial centroid. The simplicity of the K-Means method is widely used because it is easy to implement and has a high level of accuracy so that it is more scalable and efficient, however the K-Means algorithm performs the initial calculation of the centroid randomly so that the accuracy of the results is less than optimal. The results of K-Means calculations are often obtained by experimenting several times and producing different clusters. The determination of the cluster center point randomly causes the K-Means method to not be able to get the best clustering results. In this study, to determine the new centroid is done by calculating the average value of the total object value in the new cluster [13] using (3):

$$C_i = \frac{\sum_{i=1}^{n} x_i \in s_i}{n} \quad (3)$$

where Ci = new centroid to i, si = object to i, xi = value on object i, n = number of data in each group.

Fourth, calculating the distance between the object and the centroid, to calculate the distance between the object and the centroid can be done using several approaches. This study uses the Euclidean Distance formula. This calculation calculates the quantitative value of the proximity measure, which can produce the distance from the object to the centroid [14]. The following is the Euclidean Distance formula which is used to calculate the distance between objects and the center of the cluster (4).

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)}^2 \quad (4)$$

where, d = distance between x and y, x = cluster center data, y = data on attributes; i = each data, n = amount of data, xi = data at the center of the cluster i, yi = data on each data ith.

Fifth, grouping objects based on the closest distance (the smallest distance from all object distances to the centroid). Before grouping objects, a calculation must first be done to determine the minimum value distance. After obtaining the minimum value, the objects are grouped. The final stage is to test the convergence between the new data group and the data group in the previous process, if the new data group is the same as the previous data group (convergent), the clustering process is complete. If not, then do iteration starting from determining the center of the new cluster.

The cluster homogeneity test can be determined based on the silhouette coefficient value which can be obtained through the following steps. First, calculate the average distance from an object, for example i with all other objects in the cluster using (5) [15]:

$$a_i = \frac{1}{|A|-1} \sum_{j \in A, i \neq j} d(i,j) \quad (5)$$

where | A | = amount of data in cluster A, and i, j = index of document, while d (i, j) = distance between document i and document j.

Second, calculate the average distance from document i with all documents in other clusters, and take the smallest value [16] using (6).

$$\boldsymbol{d(i,C) = \frac{1}{|A|} \sum_{j \in C} d(i,j)} \quad (6)$$

Where, d (i, C) is the average distance of object i with all objects in other cluster C where A ≠ C [17] in (7).

$$b(i) = \min_{C \neq A} d(i,C) \quad (7)$$

Third, calculate the silhouette coefficient with (8) [18]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (8)$$

where a(i) is the average distance between object i and all objects i and all objects in the same cluster (intra-cluster), b(i) is the average distance between object I and all objects in the same clusteron the closest cluster.

This research was conducted in several stages, namely the stages of problem identification, data collection, system analysis and design, system development with the implementation of the pillar algorithm and the K-Means method, system testing. These stages are shown in Fig. 2. The problem identification stage begins by extracting data related to the availability of health human resources in Central Java. The second stage is data collection, which begins with observing and analyzing the condition of the availability of health human resources with the population in Central Java. The analysis and system design phase is carried out by analyzing the collected data, then proceed with designing a system design using UML. After the system design stage is complete, the next step is to build the system by implementing the pillar algorithm in determining the initial centroid and grouping the clusters into the same group. The system development is done using NetBeans IDE version 12.1 and MySQL. After the system development stage is complete, the next stage is the system testing stage using black box testing by testing system functionality and testing the performance and error level of the algorithm.

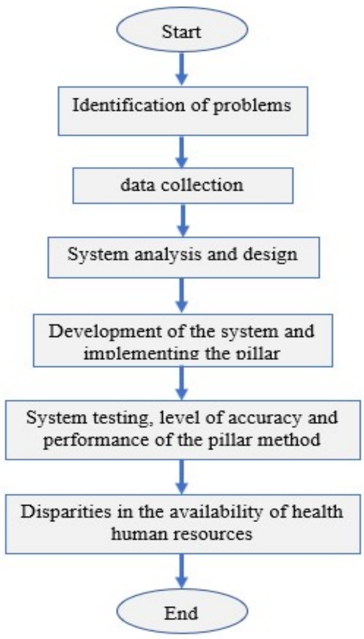

Fig. 2 Research step

III. RESULTS AND DISCUSSION

The results of this study are a desktop-based application that implements the pillar algorithm and the Chi-square-based K-Means Clustering method. This application can inform the disparity of health human resource needs in each district or city in Central Java and can cluster based on the resulting disparity, and can display the accuracy of labeling on the resulting cluster so that the application of the results of this study can be used as decision support in identifying the availability profile of health human resources in the Central Java region is shown in Fig. 3.

The initial stage in this study is to find the optimal center point of the cluster or centroid using the pillar algorithm. This algorithm is inspired by the placement of pillars in a building, where the pillars must be placed at each corner of the building that is furthest away so that the mass of the building is centered on each pillar [6]. This algorithm is able to find the centroid separately as far as possible between the initial centroids in one data distribution, and can avoid selecting the outlier data as the initial centroid [7]. The implementation of the pillar algorithm in the K-Means method begins by calculating the average value of the total objects. After the average value is obtained, then the object with the highest average value is taken, a number of clusters, as shown in Table I. In this study, a rule-of-year approach was used to determine the number of clusters and produced four clusters.

After the initial centroid candidate is obtained, it is followed by calculating the average distance value of the total object value in the initial centroid candidate using the Euclidean distance. Furthermore, the dmax value (the largest value of the distance matric) is calculated to determine the new initial centroid candidate. The centroid data generated from the selection process using the pillar algorithm as shown in Table II.

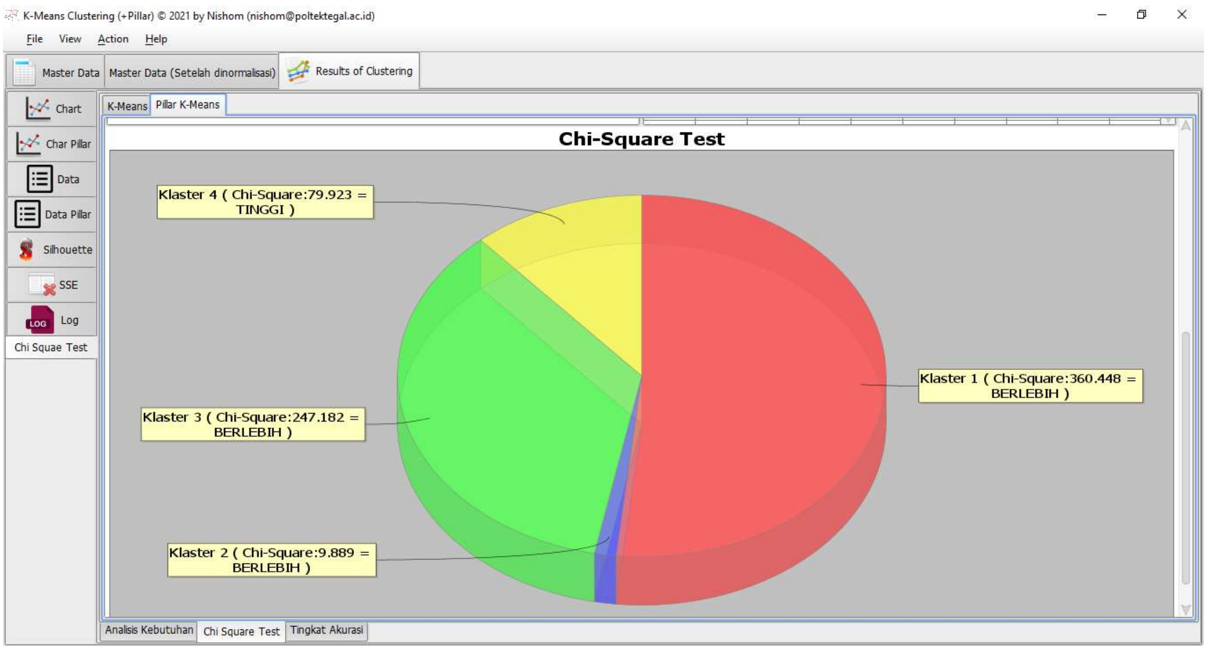

Fig. 3 Pillar – K Means application

TABLE I
CANDIDATE OF INITIAL CENTROID

ID	Population	Hospital	Public Health Center	Maternity Hospital	Clinic	Medical Staff(s)
33	1	1	0,941176	0,57971	0,57971	0,494872
2	0,928425	0,769231	1	0,5	1	0,737179
29	0,997037	0,384615	0,970588	0,166667	0,304348	1
1	0,948574	0,346154	0,970588	0,333333	0,376812	0,802564

*) Initial centroid candidate data has been normalized using the min-max method in the early stages, to minimize outliers.

TABLE II
PILLAR ALGORITHM CENTROID

ID	Population	Hospital	Public Health Center	Maternity ospital	Clinic	Medical Staff(s)
1	0.948574	0.346154	0,970588	0.333333	0.376812	0.802564
2	0.928425	0.769231	1	0.5	1	0.737179
29	0.997037	0.384615	0,970588	0.166667	0.304348	1
33	1	1	0.941176	1	0.57971	0.494872

After the centroid is determined by the pillar algorithm, the next process is the normal calculation of the K-Means algorithm, which is calculating the distance of the object to the centroid, determining the distance of the object closest to the centroid, and checking the convergence of the cluster members until convergent results are found. The results of clustering of all data on health human resources in the Central Java resulted in 4 (four) clusters. The determination of the number of clusters is determined using a rule-of-thumb approach to avoid using or randomly determining the initial centroid, this is because randomly determining the initial centroid does not produce a definite cluster, fluctuating performance, and changing accuracy of the method. The clustering results are shown in Fig. 4 and Table III.

Based on Table III, there are 5 area that need further attention regarding the need for health human resources. This is due to the fact that these regions have a high number of disparities (the high number of normative needs for health human resources with the availability of health human resources). This study uses the pillar algorithm to determine the initial centroid to get good results, this is indicated by the number of stable iterations and the relatively fast execution time as well as the increased percentage of method accuracy (to 80%) when compared to the K-Means method which does not use an algorithm. pillar as a determinant of initial centroid (with a percentage of 77.14%). The cluster homogeneity test was carried out using the silhouette coefficient and the homogeneity test results showed that the average silhouette value for all data was 0.87, which means that the clustering results had a high (very good or high) structure. The details of the silhouette values for each data or region are shown in Table IV.

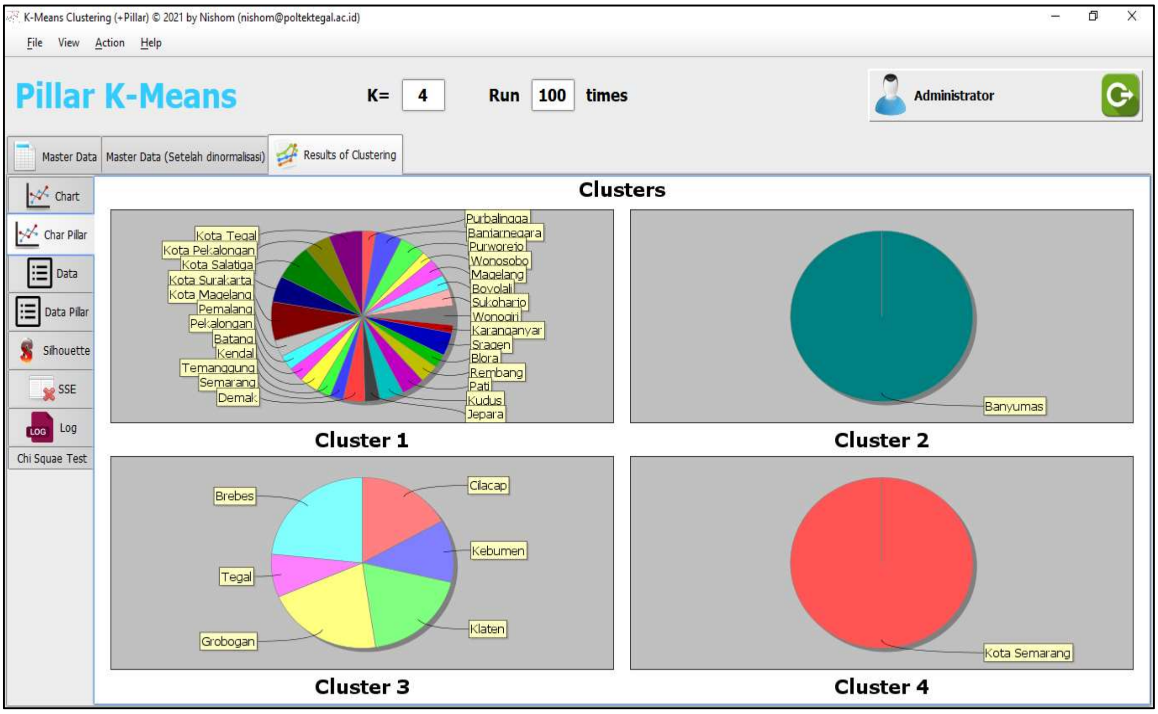

Fig. 4 Result clustering of health human resources in Central Java

TABLE III
CLUSTERING OF HEALTH HUMAN RESOURCES PROFILE

Num	Region	Cluster	Disparity	Num	Region	Cluster	Disparity
1	Purbalingga	1	EXCESS	19	Kendal	1	EXCESS
2	Banjarbegara	1	EXCESS	20	Batang	1	EXCESS
3	Purworejo	1	EXCESS	21	Pekalongan	1	EXCESS
4	Wonosobo	1	EXCESS	22	Pemalang	1	HIGH
5	Magelang	1	HIGH	23	Kota Magelang	1	EXCESS
6	Boyolali	1	REASONABLE	24	Kota Surakarta	1	EXCESS
7	Sukoharjo	1	EXCESS	25	Kota Salatiga	1	EXCESS
8	Wonogiri	1	EXCESS	26	Kota Pekalongan	1	EXCESS
9	Karanganyar	1	EXCESS	27	Kota Tegal	1	EXCESS
10	Sragen	1	HIGH	28	Banyumas	2	EXCESS
11	Blora	1	EXCESS	29	Cilacap	3	EXCESS
12	Rembang	1	EXCESS	30	Kebumen	3	EXCESS
13	Pati	1	EXCESS	31	Klaten	3	EXCESS
14	Kudus	1	REASONABLE	32	Grobogan	3	EXCESS
15	Jepara	1	HIGH	33	Tegal	3	EXCESS
16	Demak	1	EXCESS	34	Brebes	3	EXCESS
17	Semarang	1	REASONABLE	35	Kota Semarang	4	HIGH
18	Temanggung	1	EXCESS				

TABLE IV
SILHOUETTE RESULT HOMOGENEITY TESTING

No	Region	Silhouette	No	Region	Silhouette
1	Cilacap	0.802258056	19	Kudus	0.966568035
2	Banyumas	0	20	Jepara	0.94659801
3	Purbalingga	0.941197365	21	Demak	0.963013015
4	Banjarnegara	0.969337838	22	Semarang	0.93984605
5	Kebumen	0.718003762	23	Temanggung	0.94230083
6	Purworejo	0.969072805	24	Kendal	0.958774377
7	Wonosobo	0.926844024	25	Batang	0.946653909
8	Magelang	0.958657029	26	Pekalongan	0.950330283
9	Boyolali	0.94880627	27	Pemalang	0.954827479
10	Klaten	0.824614488	28	Tegal	0.590519317
11	Sukoharjo	0.955210962	29	Brebes	0.857163779
12	Wonogiri	0.961526308	30	Kota Magelang	0.98061415
13	Karanganyar	0.900525588	31	Kota Surakarta	0.971180558
14	Sragen	0.967583323	32	Kota Salatiga	0.97893515
15	Grobogan	0.840048348	33	Kota Semarang	0
16	Blora	0.944052312	34	Kota Pekalongan	0.968938714
17	Rembang	0.958810544	35	Kota Tegal	0.976137843
18	Pati	0.964942903			

The accuracy level of labeling for each region in the cluster was tested using the chi-square test. Chi-square test is used to label each cluster in general or each region regarding the level of disparity (difference between the value of availability and the value of needs) of health human resources in the Central Java region. This study uses 2 categories, namely availability and need, then the value of the degrees of freedom is (2-1) = 1. Based on the value of degree of freedom= 1 and error tolerance (alpha level) 0.05, the Chi-Square value is 3.841 as in (9).

$$label = \begin{cases} REASONABLE, if\ X^2 < 3.841 \\ High, if\ X^2 \geq 3.841\ and\ F_o < F_h \\ Excess, if\ X^2 \geq 3.841\ and\ F_h < F_o \end{cases} \quad (9)$$

where $X^2 =$ the value of chi square, $F_o =$ Value of the current number of health human resources, $F_h =$ Value The expected number of health human resources. *REASONABLE = The number of health human resources is adequate; *High = The number of health human resources needs to be increased, because it is still not sufficient * Excess = The number of health human resources exceeds the number of normative human resources needs. The details of the value of chi square for each data or region are shown in Table V.

TABLE V
SILHOUETTE RESULT HOMOGE

Num	Region	Medical Staff(s) (Availability)	Medical Staff(s) (Needs)	Medical Staff(s) (Expected)	Chi-Square (Availability)	Chi-Square (Needs)	Chi-Square (Expected)
1	Cilacap	2034	1727	1880	12.615	12.452	25.066
2	Banyumas	1881	1693	1787	4.945	4.945	9.889
3	Purbalingga	1197	933	1065	16.361	16.361	32.721
4	Banjarnegara	1616	923	1269	94.885	94.339	189.224
5	Kebumen	1530	1197	1363	20.461	20.217	40.679
6	Purworejo	1196	718	957	59.688	59.688	119.375
7	Wonosobo	893	790	841	3.215	3.093	6.308
8	Magelang	997	1290	1143	18.649	18.906	37.555
9	Boyolali	1056	984	1020	1.271	1.271	2.541
10	Klaten	1276	1174	1225	2.123	2.123	4.247
11	Sukoharjo	1025	891	958	4.686	4.686	9.372
12	Wonogiri	1100	959	1029	4.899	4.762	9.661
13	Karanganyar	1007	886	946	3.933	3.805	7.739
14	Sragen	713	890	801	9.668	9.889	19.557
15	Grobogan	1793	1377	1585	27.296	27.296	54.592
16	Blora	1495	865	1180	84.089	84.089	168.178
17	Rembang	831	638	734	12.819	12.556	25.375
18	Pati	1582	1259	1420	18.482	18.254	36.736
19	Kudus	893	871	882	0.137	0.137	0.274
20	Jepara	872	1257	1064	34.647	35.008	69.655
21	Demak	1322	1162	1242	5.153	5.153	10.306
22	Semarang	1070	1053	1061	0.076	0.060	0.137
23	Temanggung	919	772	845	6.480	6.307	12.787
24	Kendal	1488	971	1229	54.582	54.161	108.743
25	Batang	1113	768	940	31.839	31.472	63.312
26	Pekalongan	1415	897	1156	58.029	58.029	116.057
27	Pemalang	1006	1302	1154	18.981	18.981	37.962
28	Tegal	1799	1440	1619	20.012	19.791	39.803
29	Brebes	2496	1809	2152	54.989	54.670	109.658
30	Kota Magelang	156	122	139	2.079	2.079	4.158
31	Kota Surakarta	651	519	585	7.446	7.446	14.892
32	Kota Salatiga	264	194	229	5.349	5.349	10.699
33	Kota Semarang	1314	1814	1564	39.962	39.962	79.923
34	Kota Pekalongan	484	307	395	20.053	19.605	39.658
35	Kota Tegal	318	249	283	4.329	4.085	8.413

IV. CONCLUSION

The implementation of the pillar algorithm and the chi-square-based K-Means method was successfully applied to the clustering application of the availability of health human resources in Central Java. The results of the chi-square test show that of the 4 (four) clusters produced, there are 5 (five) clusters that are district / city areas that have HIGH disparity rates. The evaluation results show that the level of accuracy in the disparity labeling of health human resources is 80%. Based on the labeling, it can be seen that the regencies or cities with HIGH disparity status are Magelang, Sragen, Jepara, Pemalang, and Semarang City.

ACKNOWLEDGMENT

This research supported by Politeknik Harapan Bersama in 2020 through an institutional research fund scheme, contract number: 035.16/P3M.PHB/XI/2020.

REFERENCES

[1] Dinas Kesehatan Provinsi Jawa Tengah, "Buku Data Dasar Puskesmas & Rumah Sakit Tahun 2019," Available: https://dinkesjatengprov.go.id/v2018/wp-content/uploads/2020/09/Buku-Data-Dasar-Puskesmas-dan-RS-2019.pdf, 2020. [Online]. [Accessed: 11-Nov-2020.

[2] A. Kurniati and F. Efendi, *Kajian Sumber Daya Manusia Kesehatan di Indonesia*, Jakarta: Salemba Medika, 2012.

[3] WHO, "Global Health Workforce statistics database," https://www.who.int/gho/health_workforce/physicians_density/en, 2020. [Online]. [Accessed: 23-Nov-2020].

[4] Anindhita Maharrani, "Distribusi tenaga kesehatan tak kunjung merata," https://lokadata.id/artikel/distribusi-tenaga-kesehatan-tak-kunjung-merata, 2020. [Online]. [Accessed: 18-Nov-2020].

[5] C. D. Manning, *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009.

[6] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[7] A. R. Barakbah and Y. Kiyoki, "A pillar algorithm for K-means optimization by distance maximization for initial centroid designation," *IEEE Symposium on Computational Intelligence and Data Mining*, pp. 61–68, 2009.

[8] B. B. Bhusare and S. M. Bansode, "Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm,"*International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 3, no. 4, pp. 1317–1322, 2014.

[9] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 7, pp. 292–295, 2010.

[10] A. R. Barakbah and A. Helen, "Optimized K-means : an algorithm of initial centroids optimization for K-means," *Semin. Soft Comput. Intell. Syst. Inf. Technol.*, 2005.

[11] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 114–122, 2021.

[12] T. Alfina, B. Santosa, and A. R. Barakbah, "Tahta Alfina, Budi Santosa, dan Ali Ridho Barakbah Jurusan Teknik Industri, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember (ITS) Jl. Arief Rahman Hakim, Surabaya 60111," *J. Tek. POMITS*, vol. 1, no. 1, pp. 1–5, 2012.

[13] A. Hadi, "Segmentasi Pelanggan Internet Service Provider (ISP) Berbasis Pillar K-Means," *J. Ilm. Teknol. Inf. Asia*, vol. 13, no. 2, pp. 151, 2019.

[14] S. K. Dini and A. Fauzan, "Clustering Provinces in Indonesia based on Community Welfare Indicators," *EKSAKTA J. Sci. Data Anal.*, vol. 1, no. 1, pp. 56–63, 2020.

[15] A. R. Barakbah and Y. Kiyoki, "A pillar algorithm for k-means optimization by distance maximization for initial centroid designation," *2009 IEEE Symp. Comput. Intell. Data Mining, CIDM 2009 - Proc.*, pp. 61–68, 2009.

[16] H. Frigui, "Clustering: Algorithms and applications," *2008 1st Int. Work. Image Process. Theory, Tools Appl. IPTA*. 2008.

[17] I. H. Rifa, H. Pratiwi, and R. Respatiwulan, "Clustering of Earthquake Risk in Indonesia Using K-Medoids and K-Means Algorithms," *Media Stat.*, vol. 13, no. 2, pp. 194–205, 2020.

[18] I. Wahyudin, T. Djatna, and W. A. Kusuma, "Cluster Analysis for SME Risk Analysis Documents Based on Pillar K-Means," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 14, no. 2, pp. 674–683, 2016.