Medicine Inventory Grouping using Clustering Data Mining

J A M Nugraha*[1]

[1]Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta, Jl. Babarsari No. 43, Yogyakarta 55281

E-mail: joanna.mita@uajy.ac.id[1]

Masuk: 17 Juni 2019, direvisi: 29 Juni 2019, diterima: 1 Agustus 2019

Abstrak. Salah satu faktor utama dalam pelayanan kesehatan adalah persediaan obat yang memadai. Puskesmas adalah salah satu layanan kesehatan yang dikelola di bawah dinas kesehatan kabupaten dan kota untuk melayani pasien setiap harinya. Namun, terdapat kendala dalam proses penyediaan obat di Puskesmas. Puskesmas masih menggunakan teknik persediaan obat secara manual dengan melihat stok obat minimum. Dengan cara ini, banyak obat yang tidak digunakan dan bahkan kurang stoknya. Penerapan penambangan data atau *data mining* dapat digunakan sebagai analisis untuk menentukan persediaan obat sesuai dengan kebutuhan pasien. Dalam metode penambangan data, algoritma *clustering* adalah salah satu metode yang paling populer untuk digunakan, di mana data milik kluster yang sama akan berdekatan satu sama lain dan berjauhan dari data kluster lainnya. Untuk itu, penelitian ini menggunakan menggunakan *clustering* untuk mengelompokkan jenis obat berdasarkan jumlah pemakaian dan permintaan obat. Hasil yang didapatkan berupa informasi jenis obat dengan pemakaian cepat dan jenis obat dengan pemakaian lama setiap bulannya yang diambil dari data tiga tahun. Selain itu, informasi jenis obat dari proses *clustering* dapat digunakan untuk meningkatkan pelayanan pasien yang lebih baik.

Kata kunci: *data mining*; *clustering*; Algoritma K-means; stok; obat

Abstract. One of the main factors in health services is adequate medicine supplies. Puskesmas is one of the health services that is managed under the district and city health offices to serve patients every day. However, there are obstacles in the process of medicine supply at the Puskesmas. Puskesmas still uses medicine supply techniques manually by looking at the minimum medicine stock. In this way, many medicines are unused and even lacking. The application of data mining can be used as an analysis to determine the medicine supply according to the patient's needs. In the data mining method, the clustering algorithm is one of the most popular to use where the data belonging to the same cluster will be close to each other and will be far from the data about another cluster. For this reason, this study used clustering to classify types of medicines based on the number of medicine uses and requests. The results are obtained in the form of information on the type of medicine with rapid use and model of m with extended usage every month taken from three years of data. Also, information on the types of medicines from the clustering process can be used to improve better patient service.

Keywords: data mining; clustering; K-means Algorithm; stock; medicine

1. Introduction

Adequate health care is one of the cornerstones of public health and is a basic need besides food and education. Every day several patients visit the Puskesmas for various health examinations. The

number of patient records always increases from year to year, so a technique is needed to be able to provide the best service [1]. Therefore, productive planning health services to the community needs to be improved as efficient as possible. By managing a good supply of medicines, improving the quality of health services can be one of the supports. Additionally, to improve the quality of health care, medicine have significant role and the cost incurred for medicines purchasing covers large enough budget in health sector early budgeting [1].

The availability and quality of medicines must always be maintained as a guarantee of provided health services quality. Both developed and developing countries have equally large enough budget in overall health cost. Developed countries have a budget of about 10-15%, and the developing countries have a higher budget around 35- 66%. For example, Mali has a budget of 66%, China 45%, Thailand 35%, and Indonesia 39%. With the high spent for medicines purchasing, it is necessary to manage the supply of medicines that suits the patient's needs therefore the rate of medicine eradication due to an excessive supply of medicines with missed expiration dates can be reduced [2].

Pandanaran Health Center is a District / City Health Technical Implementation Unit (UPTD) which is responsible for organizing health empowerment in the region. It is a functional health organization in which the development of public health service will foster the community participation and provides comprehensive and integrated services to the communities in the region. This organization works through the activities required. Puskesmas is an organizational unit in which the independency to carry out operational tasks for health development in the sub-district area is given by the District / City Health Center. Puskesmas Pandanaran has 6 working region covering Mugassari Village, Randusari Village, Barusari Village, Bulustalan Village, Pleburan Village, and Wonodri Village. In this case to avoid the medicine overstock and shortage, Puskesmas requires to perform medicine grouping based on the patients' needs. The data mining clustering method is one of the grouping method which can be implemented.

Some Puskesmas have not implementing excellent services related to the availability of medicine services. Several constraints, among others, are the limited knowledge of the Puskesmas management in the pharmaceutical pharmacy function, the ability of the pharmacy staff, management policies of the institution in charge of the Puskesmas, and the limited knowledge of relevant parties regarding pharmaceutical services [3], [4]. As the results, Puskesmas still provides traditional service in which they focused only for the product in terms of supply and distribution. Some patients suffer less medication due to the medicine shortage caused by the Puskesmas poor medicine supply planning [5].

Data mining is widely used in various domains including in health domain. It plays an essential role in health domain [6]. Data mining algorithms are used to explore hidden knowledge from growing digital data in which the useful information can be produced later. The aim of this process is to apply the specific data mining methods for pattern discovery and extraction. Among the data mining techniques that have been developed in the last several years, the most popular technique used is the technique clustering. The clustering term, in general, refers to the research methodology concerning in the division of groups with several general characteristics consideration. Nowadays, clustering has become a valid instrument for solving complex problems of computer science and statistics. In particular, this method is often used in data mining and is useful for finding patterns of specific interest from data to support the knowledge discovery process [7], [1].

The application of data mining algorithms can be used as a solution to determine the medicine supplies based on the customer needs. Previous application of data mining in health domain have emerged several reliable early detection systems and various other healthcare-related systems from clinical data and diagnosis [3]. The most popular data mining algorithms is the K-means clustering algorithm. The results of clustering algorithm provides information about medicine types monthly consumption clustering in which this information can be used as the reference for the following year medicine purchasing plan. In addition, the benefit of the data mining process is to produce information that can be used as a recommendation to the Puskesmas in order to improve their service for community.

2. Theoretical Framework

2.1. Data mining

Data mining is also known as Knowledge Discovery in the Database (KDD), which is a process for obtaining efficient, new, and useful patterns that can finally be understood from a large number of data sets. Data mining is a process for extracting or mining useful knowledge from large amounts of data to help make the right decisions. The data mining algorithm model is divided into classification and prediction, clustering analysis, association rules, time patterns, outlier detection, and other categories [8], [9], [10].

Data mining is also used for medical decision-making processes under uncertainty supported by preference, evidence theory, and exploitation of high utility patterns. It later helps decision-makers in the healing process to be implemented using Conditional Preferences Base (CPB). Health professionals have to investigate and understand the uncertainty and allow for clicking identification of characteristics that determine someone under uncertainty and to understand the various forms and representations of uncertainty. Furthermore, it has been determined what is meant by medical incidents in flight and how its management is also a decision based on uncertainty issues. Then, reasoning and ontological reasoning to manage this uncertainty have been formed to support the decision making the process in aircraft [11].

In addition to research in China, data mining is used to analyze research trends and explore general research frameworks by extracting important topics and analyzing important topics for Electronic Health Records (EHR). The results obtained are in the form of knowledge insights that can be used to guide the selection of methods in health knowledge discovery, medical decision support, and public health management [12]. Data mining is also used to identify patterns of disease found in fruit therefore the excessive use of chemicals can be prevented which results in healthy fruit production. In this study, the use of data mining methods shows a result of the accuracy of 89% [13].

2.2. Clustering

Clustering is included in the unsupervised learning category. The purpose is to partition data that does not have labels into the same group. Data belonging to the same cluster will be close to each other and will be far from the data belonging to different groups [7]. Various distance criteria can be used to evaluate how close the data is [14].

To group large data sets, there are three essential elements, namely proximity distance (similarity, difference, or distance measure), function to evaluate the quality of grouping, and the third is the algorithm used for computational grouping [15]. Figure 1 shows the grouping of data from 39 data into 4 clusters, namely cluster 1, cluster 2, cluster 3, and cluster 4. Where the same data types, shown using the same color into 1 cluster and vice versa.

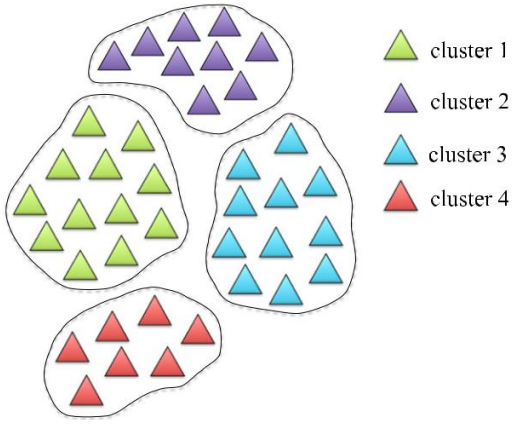

Figure 1. Example of grouping 39 data into 4 clusters

In particular, the measurement of similarity is taken from the proximity value which has large value when point 1 and point 2 are similar. Conversely, a measure of inequality (or measure of distance) is a measure of proximity that gets a small value when point 1 and point 2 are similar. Function to evaluate the quality of the grouping must be able to distinguish between good and lousy grouping. Thus, the algorithm used to calculate groupings is based on the optimization of the evaluation function [7].

Grouping problems can be classified as Euclidean and Non-Euclidean. Euclidean size is based on the concept of Euclidean space, which is characterized by several dimensions and specific solid points. The average of two or more points can be evaluated in the Euclidean space, and the proximity size can be calculated according to the location of the points in space. The three Euclidean measures that have been used for grouping in many domains are Euclidean distance, Manhattan distance, and Minkowski distance. Euclidean Distance are defined in Equation (1) **[**7**].**

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

2.3. K-means Algorithm

K-means clustering is the most widely used grouping method based on data partitions. The main idea is to collect original data into clusters *k* thus the sample being an attribute of the same is in the cluster are the same. The central processing procedures are as follows: (1) Select sample *k* randomly from the original data. (2) Each sample is taken as the center of the group *k*. (3) Calculate the distance between samples using the Euclidean distance formula. (4) Central samples *k* are calculated separately, and each sample is divided into the nearest center. (5) Clusters that are sampled are clusters from the central sample. (6) Iterations are repeated until the sample group no longer changes. Square error is usually used as a criterion for convergence of functions [16].

2.4. CRISP-DM

Currently, the data mining process model provide a general overview of the life cycle of data mining projects itself. Inside contains the project phase where the focus is on their respective tasks and also the relationship between assignments one and the other. The relationship could exist between the tasks of data mining that depends on the purpose, background, and interests order, and most importantly the data [17].

Figure 2. Phases of the CRISP-DM reference model

Six phases of life cycle data mining project are shown in Figure 2. The sequence of CRISP-DM phase is not rigid and can move back and forth between the different phases if necessary. The results of each phase determine which phase, or specific task of a phase, must be done next. The arrow shows the most important and frequent dependencies between phases. The outer circle in Figure 2 represents the nature of the data mining cycle itself. The process of data mining does not end after the solution is used. Lessons that can be obtained during the process and from the solutions used can trigger new business questions that are often more focused. The subsequent data mining process will get benefit from previous experience [17].

3. Research Methodology

3.1 Stages of Data Mining

The stages of data mining will use the CRISP-DM method. The Medicine Consumption and Demand Report (LPPO) which are stored in excel form will be used for the data mining process. Each data set will be processed every month for three years therefore later each month the types of medicines that are used are the highest. The number of processed medicine data is 9679 data.

3.1.1. Business Understanding Phase. The business understanding phase is the initial phase in the data mining stage. The carried out activities such as business objectives determination, objectives implementation, and the initial research strategy preparation. Figure 3 shows the steps taken in the business understanding phase.

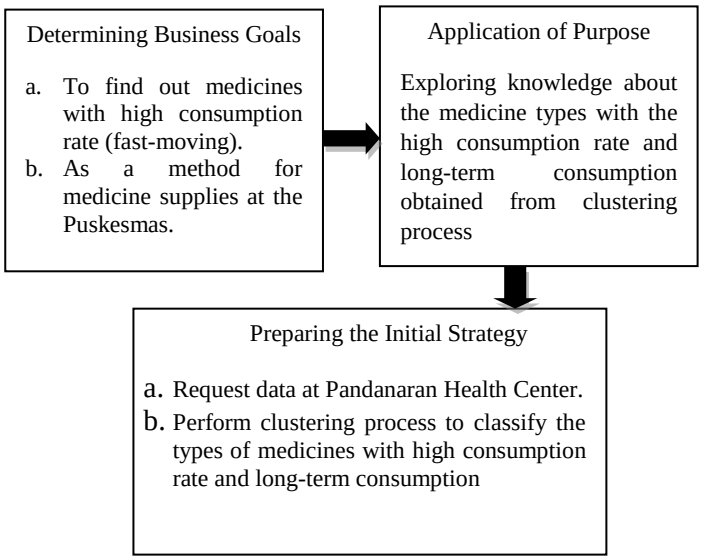

Figure 3. Steps in the Business Understanding Phase

3.1.2 Data Understanding Phase. In this phase, the data is collected and studied for further analysis. In addition, the data quality and completeness are evaluated. The missing values often occur, especially if the data are collected over a long period of time. The missing or empty attributes, spelling values, and whether attributes with different values have the same meaning have to be checked. After the data checking process, it is found that the written format of medicine numbers has different writing styles such as some medicine numbers are written with and without comma, and some are written with blank numbers. Figure 4 shows the steps taken in the Data Understanding phase.

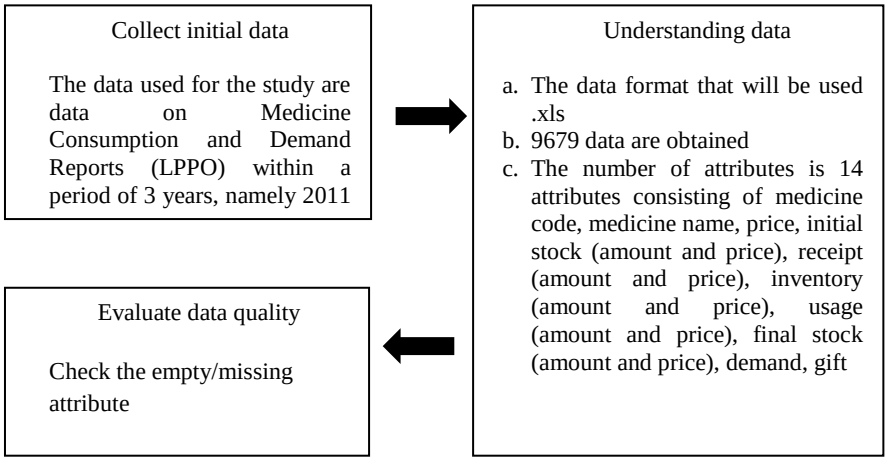

Figure 4. Steps in the Data Understanding Phase

3.1.3 Data Preparation Phase. In the data preparation phase, data sets are set up, which are then ready to be processed by using a modeling tool by selecting a required attribute. The attributes that will be used are taken from LPPO, such as medicine names, usages, and demands. In which later, the types of medicine with high usage and demand are included in the category of high consumption rate (fast-moving) and long term consumption (slow-moving) types of medicine. The selection of the new attribute will be stored in the .xls file, in which it will be ready to be included in the modeling tool.

The monthly separated LPPO will be collected into one year thus later there will be three new datasets that are ready to be processed in the modeling phase. Furthermore, the corrupted data cleaning and repairing are done by deleting the unneeded data, while the data uniformity is done by considering it as the same data but has different value. Data cleaning include: (a) In the new dataset there are still missing values. The missing values will be filled with the value of 0 thus it can be processed in the modeling tool. (b) Medicine number written format uniformity with no comma. Figure 5 shows the steps taken in the Data Preparation phase

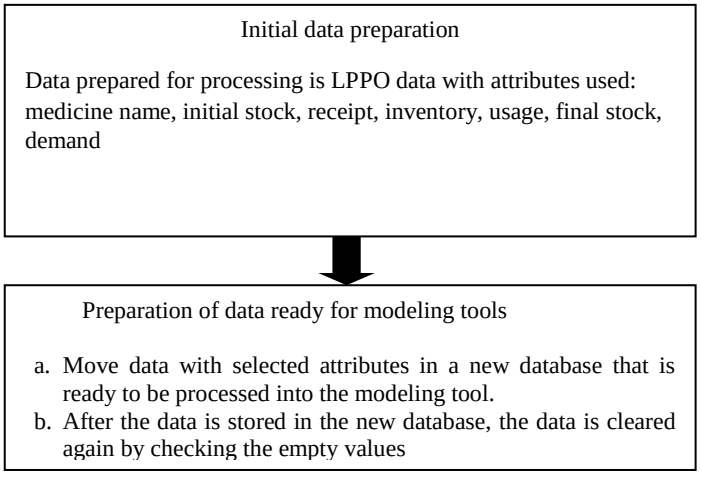

Figure 5. Steps in the Data Preparation Phase

3.1.4 Modeling Phase. This modeling phase is involved directly in the data mining suite method. It includes the selection of data mining methods and its algorithms. It will determine the attributes with optimal value. The steps in modeling are shown in Figure 6.

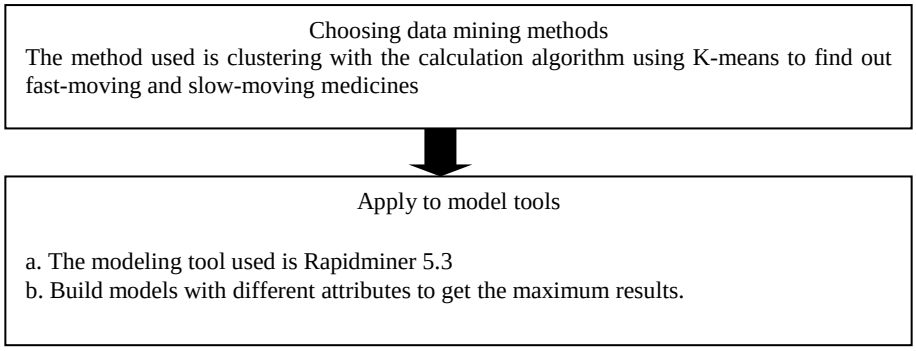

Figure 6. Modeling Steps

3.1.5 Evaluation Phase. In the evaluation phase, an assessment is carried out on the process that has been processed by using the data mining method. This phase provides information on whether the data being processed are suitable with the initial objectives at business understanding phase. Figure 7 shows the steps taken in the Evaluation Phase.

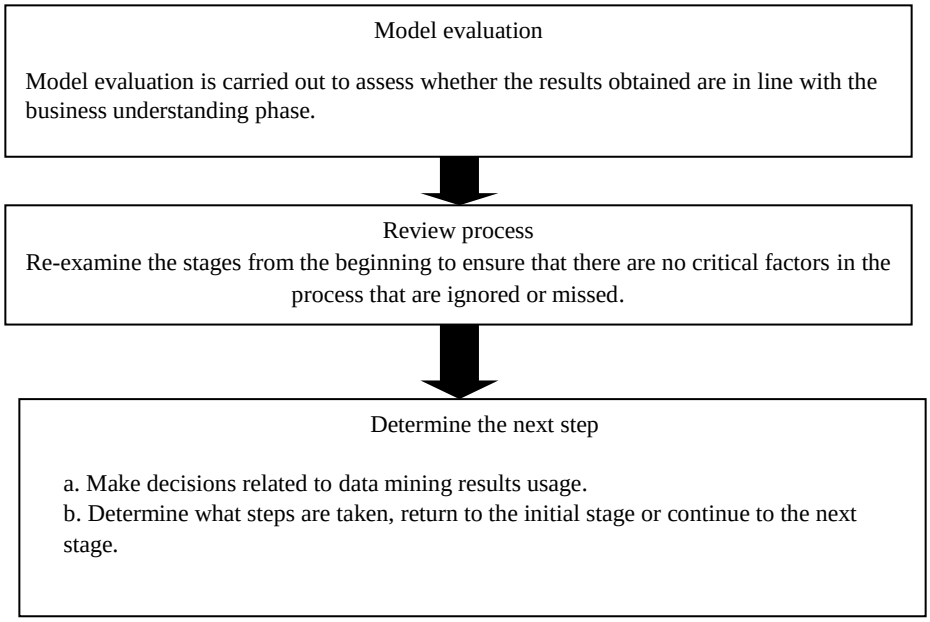

Figure 7. Steps of the Evaluation Phase

3.1.6 Deployment It is a phase of report or presentation preparation of the knowledge obtained from data mining process evaluation. In this case, the application of the clustering method is compatible with the goals/objectives to be achieved in the business understanding phase, in which the types of medicine can be separated into high consumption rate (fast-moving) and long-term consumption (slow-moving).

4. Results and Discussion

4.1. The stages of the clustering method using the K-means Algorithm
The following are stages of the clustering method using the K-means Algorithm:
1. Clusters number determination
In this study, the data will be divided into 2 clusters, namely cluster 0 and cluster 1.

2. Initial cluster center determination
Initial cluster center determination (centroid) is obtained from the data itself rather than by determining a new point using the initial central dimension of the data. The centroid of each cluster is shown in Table 1.

Table 1. The centroid of each cluster

Attribute	Cluster 0	Cluster 1
Use	0	8550
Demand	0	4937

3. Distance calculation with cluster center
Distance calculation uses the Equation (1). The data values are take and cluster center values can be calculated by using the Euclidian Distance formula with each cluster center. For example, the distance of the first medicine data will be calculated with cluster 0.

$$d(1,0) = \sqrt{(219-0)^2 + (289-0)^2}$$
$$= \sqrt{47961-83521}$$
$$= \sqrt{131482}$$
$$= 362,6$$

Based on the calculation results, it is found that the distance of the first medicine with cluster 0 is 362.6. Then the distance of the second medicine will be calculated with cluster 1 with the equation:

$$d(1,1) = \sqrt{(219-8550)^2 + (289-4937)^2}$$
$$= \sqrt{(-8331)^2 + (-4648)^2}$$
$$= \sqrt{69405561-21603904}$$
$$= \sqrt{91009465}$$
$$= 9539,89$$

Based on the calculation results, it is found that the distance of the first medicine with cluster 1 is 9539.89. Based on the distance calculation from the first medicine with cluster 0 and the first medicine with cluster 1, it is found that the closest distance to the center of the cluster is the first medicine, so the first medicine is in cluster 0. The calculation will be continued until the last data in which the closest distance of each data to the cluster center will be known.

4. Data grouping
The distance calculation results indicate that the data is located in one group closer to the cluster center as shown in Table 2.

Table 2. Data distance and the closest distance to cluster center

Medicine name	Use	Demand	Distance to cluster 0	Distance to cluster 1	Closest distance
Amoxicillin Sir Krg 125mg / 5ml	219	289	362.60	9539.89	Cluster 0
Antalgin tablets 500mg	8550	945	8602.07	3992.00	Cluster 1
Dekstrimetorfan hbr sir	102	163	192.28	9703.60	Cluster 0
Diazepam tablets 2 mg	1671	1050	1973,51	7901.23	Cluster 0
10mg tab belladon extract	0	0	0.00	9873.02	Cluster 0
Glyceryl is a 100mg tab	3367	0	3367.00	7158.03	Cluster 0
Ibuprofen tablet 200mg	811	366	889.76	8988.11	Cluster 0
Isosorbide dinitrate tab 5mg	141	137	196.60	9682,52	Cluster 0
Folic acid	315	100	330.49	9550.49	Cluster 0
Klorfeniramin maleate tb 4mg	7191	4937	8722,64	1359,00	Cluster 1

5. New cluster center determination

In order to get a new cluster, its center can be calculated from the average value of cluster members and cluster centers. The new cluster center is used to do the next iteration if there is no convergence results obtained. The calculation example of the new cluster center in cluster 0 is done by looking at the data that located at closest distance to cluster 0 or the data that included in cluster 0 is divided by the amount of data entered in cluster 0, for example:

$$= \left(\frac{219+102+1671+0+3367+811+141+315}{8}, \frac{289+163+1050+0+0+366+137+100}{8} \right)$$

$$= \left(\frac{6626}{8}, \frac{2105}{8} \right)$$

$$= (828.25, 263.13)$$

Based on the calculation results, the new centroid is obtained in cluster 0, namely (828.25, 263.13). Then the new center in cluster 1 is calculated and yields new centroid, namely:

$$= \left(\frac{8550+7191}{2}, \frac{945+4937}{2} \right)$$

$$= \left(\frac{15741}{2}, \frac{5882}{2} \right)$$

$$= (7870.5, 2941)$$

The new centroid in cluster 1 is (7870.5, 2941). Therefore, in the next calculation the new centroid is used. The centroid of each cluster is shown in Table 3.

Table 3. The centroid of each cluster

Attribute	Cluster 0	Cluster 1
Use	828,25	7870,5
Demand	263,13	2941

The iteration will be repeated consecutively to find out whether the data is moved. Calculations will be performed as it is performed at the second stage, in which the data distance with each cluster is known by using the new centroid.

4.2. Apply to model tools

The modeling tool used is Rapidminer 5.3, which can be used to facilitate the calculation of the K-means algorithm and the C4.5 decision tree algorithm. The Rapidminer can also be used to calculate the accuracy of the processed data. The clustering process divides the data into 2 clusters. The divisions of 2 clusters can be seen in Figure 8, in which it is circled in red. K = 2 means the cluster division into two clusters, according to the expected process. After the cluster division, the process is executed by selecting the Run button on the taskbar.

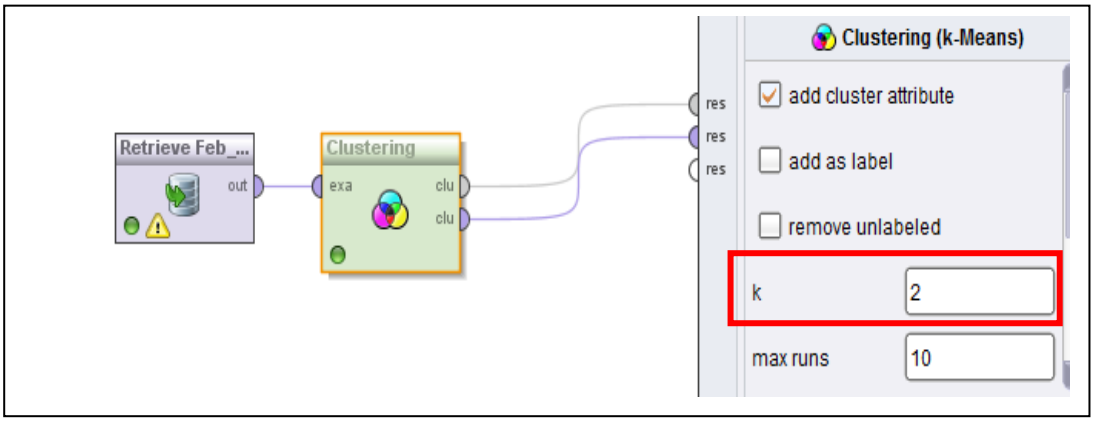

Figure 8. The K-means clustering process uses Rapidminer

The results obtained from the clustering process in January are divided into two clusters, namely cluster 0 and cluster 1 with item numbers of 11 and 471 respectively as shown in Figure 9. After the results are obtained, the data are analyzed based on the attributes used thus the obtained results with cluster 0 is categorized as fast-moving medicines while the results with cluster is categorized as slow-moving medicines.

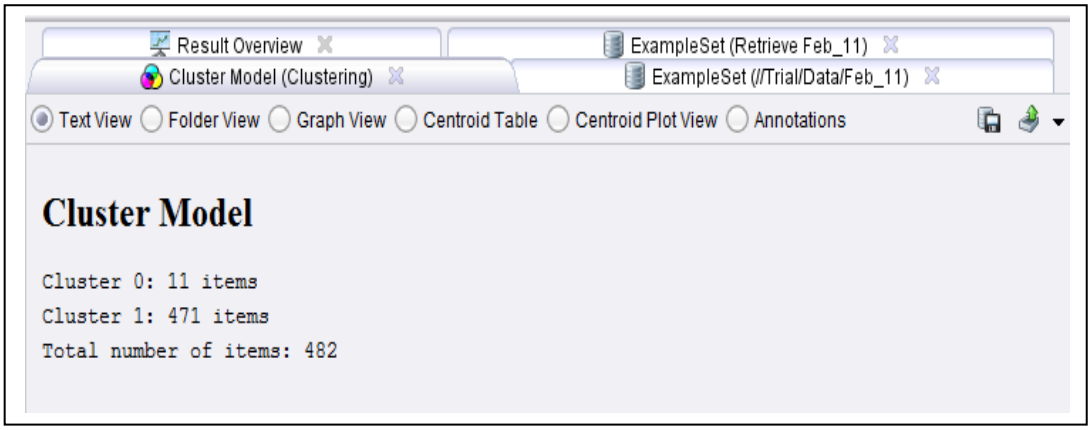

Figure 9. Results of the K-Means clustering process using Rapidminer

Data will be processed in a clustering manner monthly thus each datum will be known whether the type of medicine belongs to the fast-moving cluster or slow-moving cluster. Then the cluster is labeled as label as it is shown in Table 4.

Table 4. Add label fields

Medicine name	Label	First stock	Reception	Stock	Use	Last stock
Mercury dental use	Slow	8	0	8	0	8
1ml disposable syringe	Slow	800	0	800	0	800
2.5ml disposable syringe	Slow	2334	0	2334	69	2265
2.5 ml syringe	Slow	166	0	166	0	166
3ml disposable syringe	Slow	0	0	0	0	0
5ml disposable syringe	Slow	992	0	992	30	962
Albendazole 400mg tablets	Slow	1360	0	1360	0	1360
Allopurinol tablets 100 mg	Slow	1466	300	1766	1265	501
Amoxicillin capsules 500mg	Fast	82145	0	82145	5946	76199

4.3. Evaluation

Evaluation is the phase of data mining results interpretation. Evaluation is performed in-depth with the aim that the results in the modeling phase are suitable with the business understanding phase objectives.

4.3.1 Evaluation of the model. In this stage, the clustering process is evaluated by using the Davies Bouldin Index (DBI) and yields a value of -0.320. Based on the evaluation results, it can be concluded that performed clustering process is satisfactory. This can be seen from the acquisition of the DBI value. If the DBI value is close to the value 0 indicates the better the cluster obtained. The evaluation process is shown in Figure 10.

The evaluation phase is reviewed whether the obtained results suit with the business understanding phase objectives. In the business understanding phase, the objective is to determine which medicines are categorized as fast-moving medicines thus it can be used for inventory control methods at Pandanaran Health Center. The results of the data mining process divides the medicine clusters into two clusters, namely fast-moving and slow-moving. After the obtained results show compatibility to the business phase objectives, a checking process is performed to ensure that there is no critical factor missed at the all performed stages in data mining processing processor.

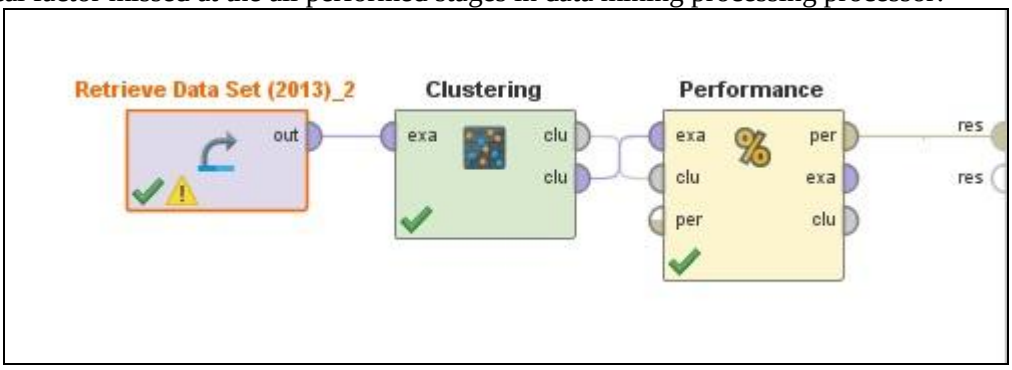

Figure 10. The results of the evaluation process using Rapidminer

4.3.2. Process Review. In this stage, it will be ensured that all-important performed stages of factors by processing the data are not missed. Thus the next process can be performed in the data processing process because it is met the objectives of data mining.

*4.3.3. Determine the next step***.** This stage has two options, back in the early stages or proceeded to the final stage. As in the previous stage has met the objectives and no step is missed, the next step is towards the final stage to determine the distribution of the obtained results by performing analysis.

4.4. Deployment

This is the application phase of formulated data mining methods in the business understanding phase. The objective is to determine the information about the fast-moving medicine types based on the data obtained in the LPPO document. Therefore, this data mining information can be used as the reference for Puskesmas to determine its medicine inventory in the following years.

Thus, the formulated objectives in the business understanding phase is completed by classifying the types of medicine based on the patients' needs. Therefore, the medicines supply to the Puskesmas can be optimized. In this study, a small DBI value is obtained, in which it indicates the clustering process result is more satisfactory compared to other studies.

5. Conclusion

Based on the results, it can be concluded that the data mining clustering method results can be used to classify the types of medicine data according to the patients' needs. The results of this clustering process are used as pre-processing for the next process, in which the medicines supply is determined.

In the next following year, the three earlier years fast-moving medicines data can be used as reference by the medical officer for medicine purchasing plan.

Based on the results, the fast-moving and slow-moving medicine clusters can be accessed monthly in the three earlier years. In the further research, a process of determining medicine supplies will be performed based on the clustering process results. In addition, the fast-moving medicine type supplies need to be controlled precisely to prevent the shortage when the patients need. While the slow-moving medicine type supplies can also be controlled to prevent the medicine eradication
This clustering process results can also be used as a reference for Puskesmas strategy in disseminating the common spread diseases by analyzing the cure effectiveness of certain types of medicine.

6. References

[1] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," *Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.
[2] E.N. Zuliani, "Pengendalian Persediaan Obat Antibiotik dengan Analisis ABC Indeks Kritis di RSUD Pasar Rebo tahun 2008" U. Indonesia and Y. S. R. I. Rahayu, Fakultas kesehatan masyarakat, Depok, Indonesia, 2011.
[3] M. Cohen, "A systemic approach to understanding mental health and services," *Soc. Sci. Med.*, vol. 191, pp. 1–8, 2017.
[4] Y. Zhang, Z. Zhou, and Y. Si, "When more is less: What explains the overuse of health care services in China?", *Soc. Sci. Med*, vol. 232. Elsevier Ltd, 2019.
[5] C. Peng, P. Goswami, and G. Bai, "Linking health web services as resource graph by semantic REST resource tagging," *Procedia Comput. Sci.*, vol. 141, pp. 319–326, 2018.
[6] M. Ilayaraja and T. Meyyappan, "Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets," *Procedia Comput. Sci.*, vol. 70, pp. 586–592, 2015.
[7] Jan Wira Gotama Putra, "Pengenalan Pembelajaran Mesin dan Deep Learning," edisi 1.2. March, 2018.
[8] L. Yang, H. Xu, and Y. Jiang, "Applied research of data mining technology in hospital staff appraisal," *Procedia Comput. Sci.*, vol. 131, pp. 1282–1288, 2018.
[9] C. C. Aggarwal, "An Introduction to Data Mining" in *DataMining*, vol. 1, New York, USA, 2015, hal 3-5.
[10] G. Lukhayu Pritalia, "Penerapan Algoritma C4.5 untuk Penentuan Ketersediaan Barang E-commerce," *Indones. J. Inf. Syst.*, vol. 1, no. 1, pp. 47–56, 2018.
[11] A. Sene, B. Kamsu-Foguem, and P. Rumeau, "Data mining for decision support with uncertainty on the airplane," *Data Knowl. Eng.*, vol. 117, no. April, pp. 18–36, 2018.
[12] J. Chen, W. Wei, C. Guo, L. Tang, and L. Sun, "Textual analysis and visualization of research trends in data mining for electronic health records," *Heal. Policy Technol.*, vol. 6, no. 4, pp. 389–400, 2017.
[13] N. M. Rémy, T. T. Martial, and T. D. Clémentin, "The prediction of good physicians for prospective diagnosis using data mining," *Informatics Med. Unlocked*, vol. 12, no. August, pp. 120–127, 2018.
[14] R. Srinivasan, S. Manivannan, N. Ethiraj, S. Prasanna Devi, and S. Vinu Kiran, "Modelling an Optimized Warranty Analysis Methodology for Fleet Industry Using Data Mining Clustering Methodologies," *Procedia Comput. Sci.*, vol. 87, pp. 240–245, 2016.
[15] J. Melton *et al.*, "Clustering Analysis", *Data Mining : Concepts and Techniques*, Second Edition, San Fransisco, USA, 1999, bab 7, hal. 383-464.
[16] H. Yu, G. Wen, J. Gan, W. Zheng, and C. Lei, "Self-paced Learning for K-means Clustering Algorithm," *Pattern Recognition Letters*, Elsevier B.V., 2018.
[17] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model," *Procedia CIRP*, vol. 79, pp. 403–408, 2019.