INDONESIAN JOURNAL OF INFORMATION SYSTEMS (IJIS)
Program Studi Sistem Informasi
Universitas Atma Jaya Yogyakarta
Jl. Babarsari No. 43 Yogyakarta 55281

Yogyakarta, 31 Juli 2019

No : 04/IJIS-E/0719
Lampiran : -
Hal : Penerimaan Naskah Publikasi Jurnal

Kepada Yth:
Bapak/Ibu/Sdr. Joanna Ardhyanti Mita Nugraha
Program Studi Teknik Informatika
Universitas Atma Jaya Yogyakarta
Indonesia

Dengan hormat,
Melalui surat ini kami beritahukan bahwa naskah yang telah Bapak/Ibu kirimkan *kepada Indonesian Journal of Information Systems* (IJIS) yang berjudul:

"Drug Inventory Grouping using Clustering Data Mining"

berdasarkan hasil review dan penyuntingan yang telah dilakukan, dinyatakan **diterima** untuk dipublikasikan pada jurnal kami untuk terbitan Volume 2, Nomor 1, Agustus 2019. Demikian pemberitahuan dari kami, atas perhatian Bapak/Ibu, kami ucapkan terimakasih.

Hormat kami,
Pengelola *Indonesian Journal of Information Systems* (IJIS)

Clara Hetty Primasari, S.T., M.Cs.

Drug Inventory Grouping using Clustering Data Mining

Joanna Ardhyanti Mita Nugraha

Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta, Jl. Babarsari No.44, Daerah Istimewa Yogyakarta, 55281, Indonesia

E-mail: joanna.mita@uajy.ac.id

Abstract. One of the main factors in health services is adequate drug 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 drug supply at the Puskesmas. Puskesmas still use drug supply techniques manually by looking at the minimum drug stock. In this way, many drugs are unused and even lacking. The application of data mining can be used as an analysis to determine the drug 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. The results obtained in the form of information on the type of drug with rapid use and model of drug with extended usage every month taken from 3 years of data. Also, information on the types of drugs from the clustering process can be used to improve better patient service.

Keywords: Data mining; clustering; K-means algorithm; stock; drug

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 improve the best service for patients [1]. Therefore, in order to improve health services in Puskesmas, it needs to be improved by planning health services to the community to be more productive and efficient. By managing a good supply of medicines, it can be one of the supports in improving the quality of health services. Additionally, to improve the quality of health care, medicine has a significant role and the costs incurred for the purchase of drugs has a budget big enough of to the entire early budget in the field of health [1].

The availability and quality of medicines must always be maintained as a guarantee of the quality of health services provided. Both developed and developing countries have budgets that are equally large in overall health costs. Which developed countries have a budget of about 10-15%, and the developing countries have a budget that is higher than about 35- 66%. For example, in Mali has a budget of 66%, in China 45%, in Thailand 35%and Indonesia 39%. With the high budget for purchasing drugs, it is necessary to manage the supply of medicines that can suit the needs of patients so that it can reduce the rate of drug eradication caused by an excessive supply of drugs with missed expiration dates [2].

Pandanaran Health Center is a District / City Health Technical Implementation Unit (UPTD) that is responsible for organizing health empowerment in a work area or functional health organization which is a development of public health that also fosters community participation and provides comprehensive and integrated services to communities in the region it works in the form of necessary activities. The Puskesmas is an organizational unit that is given independence by the District / City Health Service to carry out operational tasks for health development in the sub-district area. The working area of Pandanaran Community Health Center has 6 (six) assisted areas, namely Mugassari Village, Randusari Village, Barusari Village, Bulustalan Village, Pleburan Village, Wonodri Village.

In this case, Puskesmas requires grouping of drugs based on patient needs, to avoid out of stock and overstock. One grouping method that can be used is the data mining clustering method.

Some Puskesmas have not implemented excellent services related to the availability of drug services. Several constraints, among others, cause this is 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, the limited knowledge of relevant parties regarding pharmaceutical services [3], [4]. As a result, Puskesmas services are still traditional in that they only focus on the product, namely in terms of supply and distribution. Some patients experience a lack of medication when they need it. One of the factors that cause out of stock and over-stock of medical supplies in Puskesmas is poor drug supply planning [5]. In this case, it is necessary to group drugs based on patient needs, to avoid out of stock and overstock. One grouping method that can be used is the data mining clustering method.

Data mining is widely used in various domains, one of which is the health domain. Data mining plays an essential role in this domain [6]. Data mining algorithms are used to explore hidden knowledge that can later produce useful information from the volume of digital data that is proliferating. The essence of this process is the application of specific data mining methods for pattern discovery and extraction. Among the data mining technique that was developed several years ter a k hir, the most popular technique used is the technique of clustering. The term clustering, in general, is a term that refers to the research methodology about the division of groups that consider several general characteristics. At present, clustering has become a valid instrument for solving complex problems of computer science and statistics. In particular, this is very 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 be able to determine drug supplies based on customer needs. Previous data mining and health domains have emerged several reliable early detection systems and various other healthcare-related systems from clinical data and diagnosis [3]. Also, the most popular use of data mining algorithms is the K-Means clustering algorithm. The results of clustering will provide information in the form of grouping types of drugs with high usage and types of drugs with low usage which is processed every month to be used as reference materials for planning the types of drugs in the following year. In addition, the benefits of the data mining process are to produce information that can be used as a recommendation to the Puskesmas to improve service to the better 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. Put, 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].

Data mining is also used for medical decision-making processes under uncertainty supported by preference, evidence theory, and exploitation of high utility patterns. Which 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 allows 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 has been formed to support the decision making the process in aircraft [10].

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 [11]. Data mining is also used to identify patterns of disease found in fruit so that they can prevent excessive use of chemicals which can enable fruit production to be healthy without chemical residues. In this study, the use of data mining methods showed a result of the accuracy of 89% [12].

2.2. Clustering

Clustering is included in the unsupervised learning category, the purpose of which 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 [13].

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 [14].

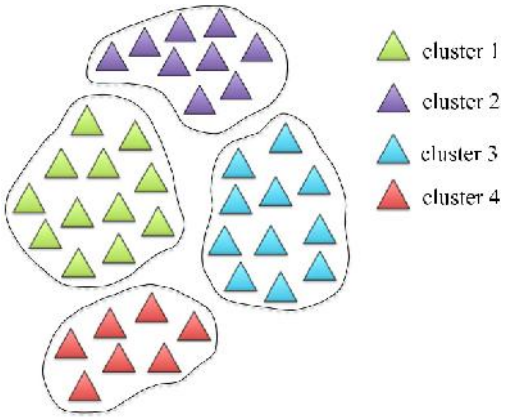


Figure 1. Example of grouping 39 data into 4 clusters

In particular, the lawyer similarity measure is taken of the size of the closeness that has excellent 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 m embedakan between groupings 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 is defined as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad [7]$$

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 so that 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) Pu sick samples 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 [15].

2.4. CRISP-DM

Currently, the data mining process model provides 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 ng on the purpose, background, and interests order, and most importantly, there is the data [16].

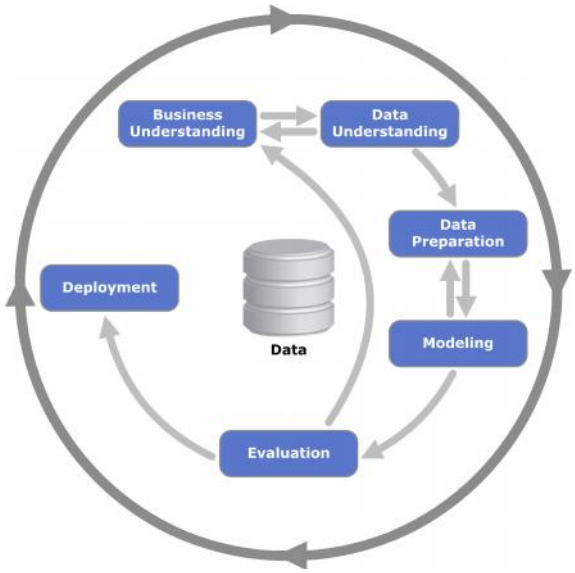


Figure 2. Phases of the CRISP-DM reference model

Project life cycle data mining consists of six phases are shown in Figure 2. The sequence phase of CRISP-DM 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 benefit from previous experience [16].

3. Research Methodology

3.1 Stages of Data Mining

The stages of data mining will use the CRIP-DM method. The data used for the data mining process is the Usage and Drug Request Report (LPPO), which is stored in excel form. The data will be processed each set every month for three years so that later each month, the types of drugs that are used are highest. The drug data to be processed amounts to 9679 data.

3.1.1. Business Understanding Phase

The business phase of understanding is the initial phase in the data mining stage. The activities carried out are determining business objectives, implementing objectives, and preparing the initial research strategy. Figure 3 shows the steps taken in the business understanding phase.

3.1.2 Data Understanding Phase

In this phase, collecting and studying data is used to analyze what can be done on the data. In addition to the data understanding phase, there is a stage in evaluating data quality and completeness of data. The missing values often occur, especially if the data is collected over a long period of time. Check for missing or empty attributes, spelling values, and whether attributes with different values have the same meaning. The search results found that writing the format of the number of drugs used coma and some did not use commas, and there were blank values in some LPP O. Figure 4 shows the steps taken 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 using a modeling tool by selecting a needed attribute. The attributes that will be used are taken from L.P. P O, namely, name of drug, usage, and request. Where later, with high usage and demand, indicating that the type of drug is included in the category of types of drugs with fast use (fast-moving) and vice versa. The selection of the new attribute will be stored in the .xls file, which will be ready to be included in the modeling tool.

The collection of separate P.O. L.P.s every month will be collected into one year so that later there will be three new datasets that are ready to be processed in the modeling phase. Furthermore, the data cleaning and repair of data corrupted by deleting data that is not in need and uniformity of data is considered the same but have different values or make it consistent. Data that is cleared include: (a) In the new dataset there are still missing values. The missing value will be filled with a value of 0 so that it can be processed in the modeling tool. (b) Uniformizing the format of writing the number of drugs with no coma. Figure 5 shows the steps taken in the Data Preparation phase.

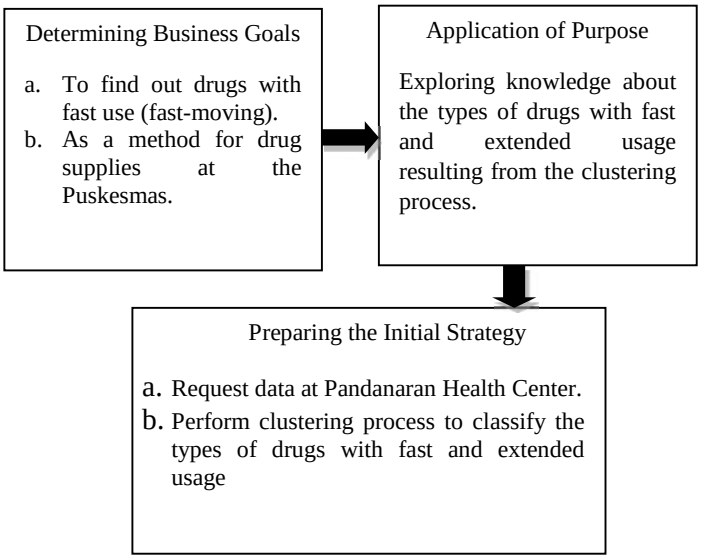


Figure 3. Steps in the Business Understanding Phase

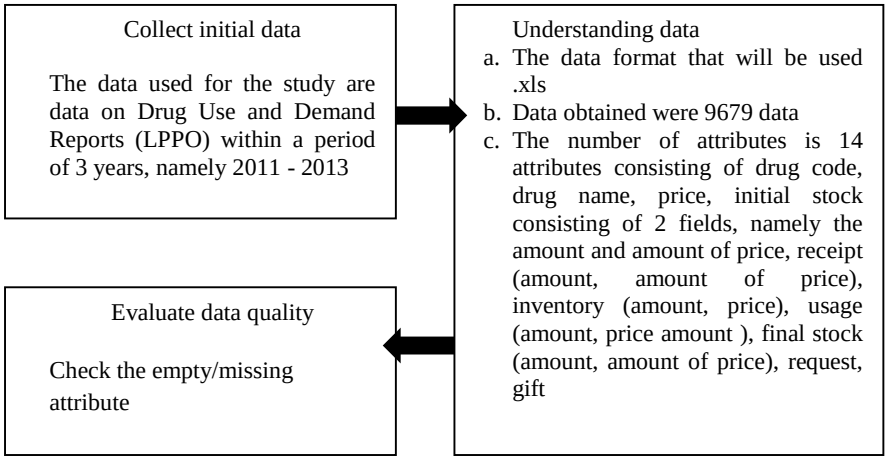


Figure 4. Steps in the Data Understanding Phase

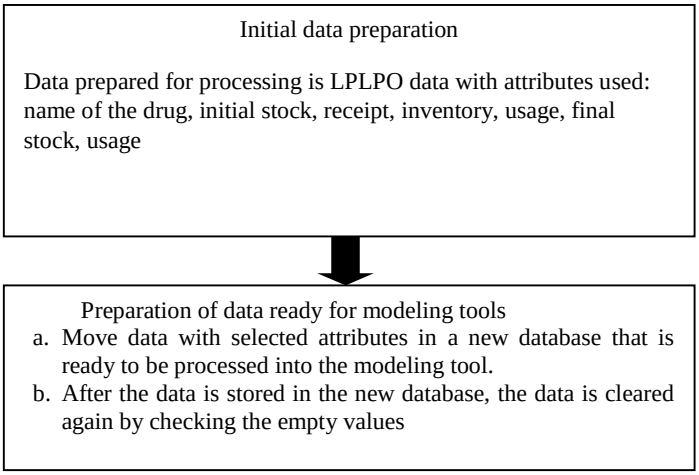


Figure 5. Steps for Data Preparation Phase

3.1.4 Modeling Phase

In the phase of modeling involves directly on the method of data mining suite. The selection of data mining methods that will be used for data processing, the use of algorithms from the data mining method and determine the attributes that have optimal values. The steps in modeling are shown in Figure 6.

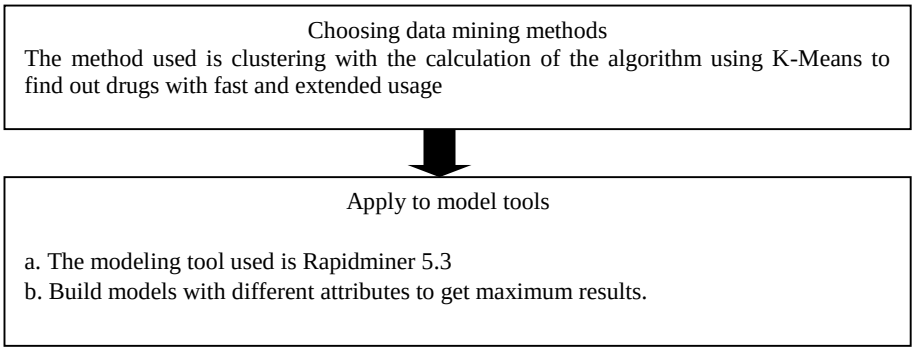


Figure 6. Modeling Steps

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

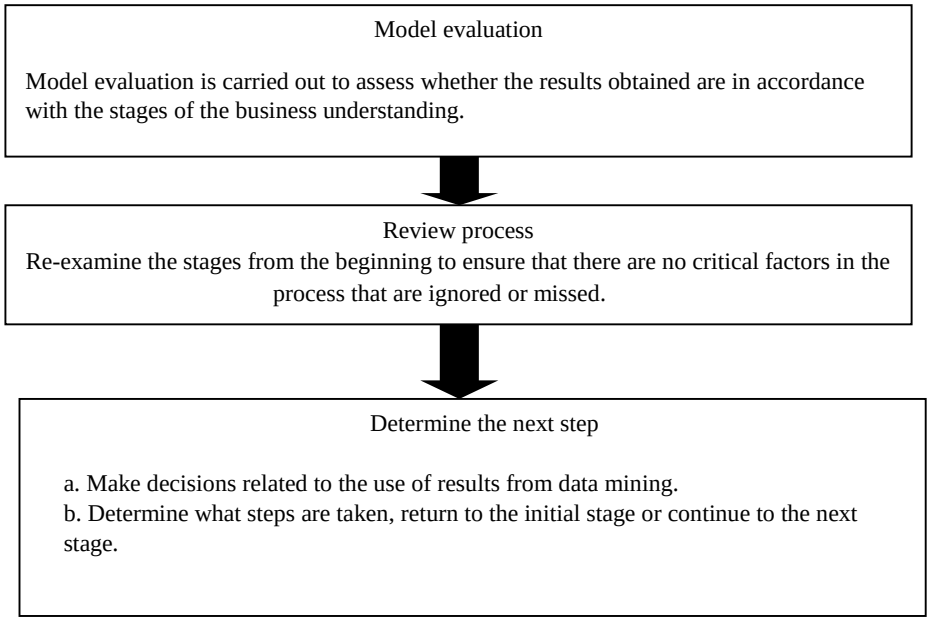


Figure 7. Steps of the Evaluation Phase

3.1.6 Deployment
It is the phase of preparing a report or presentation of the knowledge gained from evaluating the data mining process. Namely, the application of clustering methods by the goals/objectives to be achieved in the business phase of understanding, namely to be able to find out the types of drugs with fast use (fast-moving) and the types of drugs with prolonged use (slow-moving).

4. Results and Discussion

4.1. The stages of the clustering method use the K-means algorithm :
1. Determination of the number of clusters
In this study, the data will be divided into 2 clusters, namely cluster 0 and cluster 1.
2. Determination of the initial cluster center
Determination of the center of the initial cluster (centroid) is obtained from the data itself rather than by determining a new point, that is by the initial central dimension of the data.

Table 1. The centroid of each cluster

Attribute	Cluster 0	Cluster 1
Use	0	8550
Demand	0	4937

3. Calculation of distance with the center of the cluster
Calculation of distance using the Euclidian Distance Formula.

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

Take data values and cluster center values then calculate using the Euclidian Distance formula with each cluster center. For example, the distance of the first drug data will be calculated with cluster 0.

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

From the results of the calculation, it was found that the distance of the first drug with cluster 0 was 362.6. Then the distance of the second drug will be calculated with cluster 1 with the equation

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

From the results of the calculation, it was found that the distance of the first drug with cluster 1 was 9539.89. Based on the calculation of the distance from the first drug with cluster 0 and the first drug with cluster 1, it was found that the closest distance to the center of the cluster was the first drug. So the first drug is in cluster 0. The calculation will continue until the last data. So that it will be known the distance of each data with the nearest cluster center.

4. Grouping data
The distance from the calculation will be done and the closest distance between the data and the cluster center, this distance indicates that the data is in one group with the nearest cluster center.

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

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. Determination of the new cluster center
To get a new cluster 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 the results obtained have not been converged. Example calculation of the new cluster center in cluster 0 is by looking at data that has the closest distance to cluster 0 or data included in cluster 0 divided by the amount of data entered in cluster 0, for example:

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

$$= \left(\frac{6626}{8}, \frac{2105}{8}\right)$$
$$= (828.25, 263.13)$$

From the results of calculations, the new centroid is obtained in cluster 0, namely (828.25, 263.13). Then also calculated the new centroid in cluster 1, namely:

$$= \left(\frac{8550 + 7191}{2}, \frac{945 + 4937}{2}\right)$$
$$= \left(\frac{15741}{2}, \frac{5882}{2}\right)$$
$$= (7870.5, 2941)$$

And the new centroid in cluster 1 is (7870.5, 2941). So that in the next calculation, the centroid used is the new centroid, namely:

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 done again to find out whether the data is moved or not. Calculations will be carried out like the second stage, which is knowing the distance of data with each cluster 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. Also, Rapidminer can also calculate the accuracy of the data that has been processed. The clustering process will divide the data into 2 clusters. The distribution of 2 clusters can be seen in Figure 10, which is circled in red. K = 2 is the division of clusters into 2, according to the expected process. After the division of clusters, the process is executed by selecting the Run button on the taskbar.

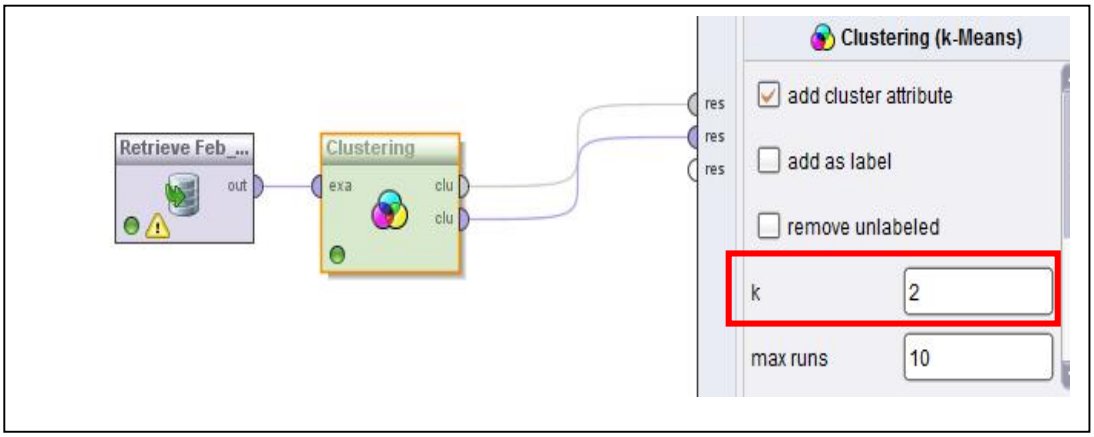


Figure 9. The K-Means Clustering Process uses Rapidminer

The results obtained from the clustering process in January have divided 2 clusters, namely cluster 0 and cluster 1, which are worth 11 and 471. After obtaining the results, the data is analyzed based on the attributes used. So that the results obtained that cluster 0 is a drug with rapid use and cluster 1 is a drug with prolonged use.

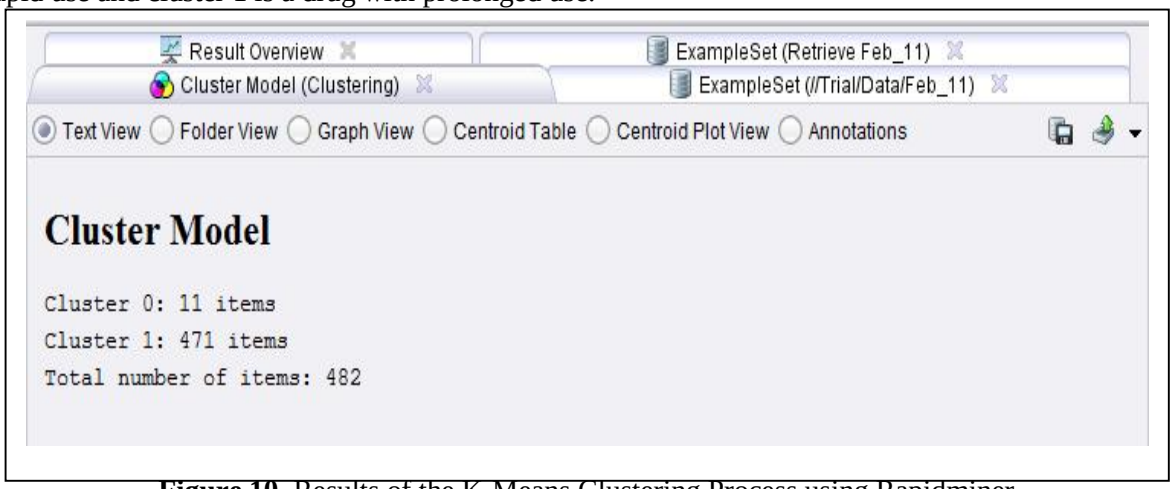


Figure 10. Results of the K-Means Clustering Process using Rapidminer

Data will be processed in a clustering manner every month so that each data will be known whether the type of drug includes clusters with rapid use or clusters with extended usage. Then each cluster that has been divided is named as a label, which is a fast and long label.

Table 4. Add label fields

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

4.3. Evaluation

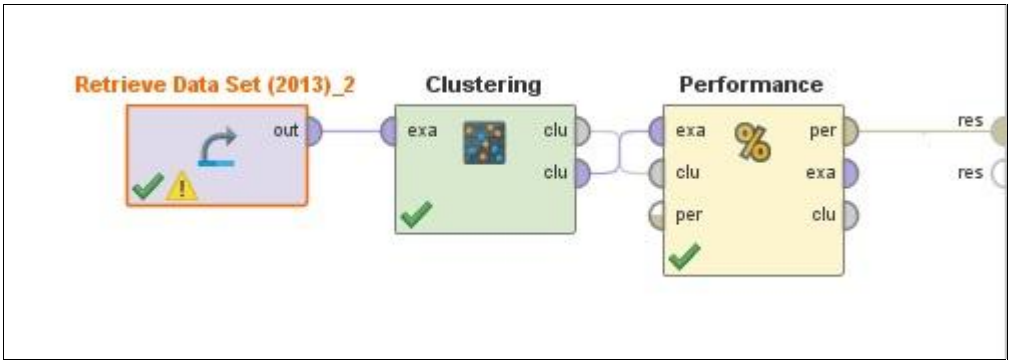
Evaluation is the phase of interpretation of the results of data mining. Evaluation is carried out in-depth with the aim that the results in the modeling phase are by the goals to be achieved in the stage of the business understanding.

4.3.1 Evaluation of the model

At this stage, the clustering process is evaluated using the Davies Bouldin Index (DBI) and produces a value of -0.320. With the results of the evaluation, it can be concluded that the clustering process that has been done is quite good. 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.

Also, the evaluation phase is also reviewed, whether the results obtained from the clustering process have met the objectives set in the stage of the business understanding. At the stage of business understanding, the goal was determined to find out drugs with rapid use so that they could be used for inventory control methods at Pandanaran Health Center. And the results of the data mining process can cluster the drug into 2 clusters, namely fast usage and long usage. After the results obtained are

the same as those in the business understanding, a checking process will be carried out which serves to ensure that all stages have been carried out in the data processing processor that no critical factors have been missed.



Gambar 10. Hasil proses evaluasi menggunakan Rapidminer

4.3.2. Process Review

At this stage, it will be ensured that all-important stages/factors that have been carried out by processing the data have not been missed. Thus the next process can be carried out in the data processing process because it has met the objectives of data mining.

4.3.3. Determine the next step

Stage it has two options, that is, back in the early stages or proceed to the final stage. Because 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 results that have been obtained by doing analysis.

4.4. Deployment

This phase is the phase of application of data mining methods that have been formulated in the business understanding phase. The goal achieved is to be able to find out information on the types of drugs with high/fast usage based on the data obtained in the LPPO document. So that later, with the information obtained from the data mining method, this can be used by the Puskesmas as a reference in determining the drug inventory the following year.

Thus, the objectives that have been prepared at the stage of understanding the business have been achieved, namely to classify the types of drugs based on the patient's needs. So that the supply of drugs can be really maximized later. In this study, a small DBI value was obtained, which indicated that the resulting clustering process was quite good compared to other studies.

5. Conclusion

From the results of the study, it can be concluded that the results of the data mining clustering method can be used to classify the types of drug data according to the patient's needs. The results of this clustering are used as pre-processing for the next process, which is for determining drug supplies. So that later, when there will be the procurement of drugs in the coming year, officers can see a list of drugs with rapid use, which is the result of data 3 years earlier.

From the process of extracting data, the results are in the form of drugs with rapid use and type of drug with extended usage every month from the date of the last three years which can be used as a reference for drug supplies in the following year. So that for further research, a process of determining drug supplies will be carried out based on the results of the clustering that has been carried out. In addition, the types of drugs classified as high usage need to get more control so that there is no drug shortage when the patient needs it. And vice versa, it should be noted for the types of drugs with low use in order to be able to manage supplies according to the needs of patients, so as to minimize the process of drug destruction.

The results of the information from the clustering process can also be used as a recommendation from the Puskesmas to disseminate diseases that often appear in certain months, by analyzing the function of the type of drug.

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] S. Mita, Y. Yamazoe, T. Kamataki, and R. Kato, *DataMining异常检测*, vol. 14, no. 3. 1981.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Melton *et al.*, "Clustering Analysis", *Data Mining: Concepts and Techniques*, Second Edition, San Fransisco, USA, 1999, bab 7, hal. 383-464.

[15] 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.

[16] 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.