Login : Jurnal Teknologi Komputer
ISSN : 2302-9692 (print) | 2723-8695 (online)
Vol. 14, No. 2, December 2020, pp. 214-222
LOGIN http://login.seaninstitute.org/index.php/Login

Identification of Book Needs Level Using the K-Means Method at STMIK Pelita Nusantara Medan Library

Amran Sitohang[1], Sarjon Defit[2], Sumijan[3]
[1]Informatics Engineering Study Program, STMIK Pelita Nusantara, Medan, Indonesia,
[2,3]Faculty of Computer Science, Universitas Putra Indonesia YPTK, Padang, Indonesia.

Article Info

Article history:

Received, Nov 9, 2020
Revised, Nov 28, 2020
Accepted, Dec 19, 2020

Keywords:

Identification,
Book,
Library,
K-Means.

ABSTRACT

Library Sekolah Tinggi Manajemen Ilmu Komputer Pelita Nusantara Medan is one of the facilities that provide book loan services to students. To improve services to the needs of books for Pelita Nusantara STMIK students. This test aims to determine the amount of books needed to borrow, and determine how many students borrow books based on books that are often borrowed simultaneously. The method used is the K-Means method, the K-Means algorithm is a non-hierarchical data clustering method with an effort to partition the available data into one or even more clusters. The data to be used is the amount of book data that is often borrowed, and the number of students who borrow books in 2015-2019. From the results obtained in the overall data process above that cluster 0 is a low-potential book borrower totaling 52 titles, Cluster 1 is a potential borrower of 30 books, and Cluster 2 is a high-potential book borrower totaling 1 book title, therefore knowledge is obtained that borrowing books at STMIK Pelita Nusantara books that are often borrowed will be reviewed for books. K-Means algorithm method has been able to be applied to identify the level of book requirements. Data is processed to obtain the number of books that are often borrowed will be multiplied. The data is processed using Rapid Miner software.

Corresponding Author:

Amran Sitohang,
Informatics Engineering Study Program,
STMIK Pelita Nusantara, Medan, Indonesia,
Jl. Iskandar Muda No.1, Merdeka, Kec. Medan Baru, Kota Medan, Sumatera Utara 20154.
Email: amranryan89@gmail.com

1. INTRODUCTION

Data mining is a technique of processing large amounts of data for grouping. This technique is used in the Knowledge Discovery in Database (KDD) process. In various literatures, theories on data mining have been around for a long time, such as Naive-Bayes and Nearest Neighbor, Decision Trees, association rules, K-Means Clustering and text mining (Fauziah Nur, et al ., 2017).

According to a study entitled "Application of the K-Means Clustering Analysis Algorithm in Human Infectious Diseases (Case Study of Majalengka Regency)", the term data mining has several views, such as knowledge discover or pattern recognition. The two terms actually have their own accuracy, the term knowledge discovery is appropriate because it is used, the main purpose of data mining is to get knowledge that is still hidden in chunks of data. The purpose of designing using the Rapid Miner application is to make it easier for librarians to collect book data. This study discusses and shows the method of grouping data books starting with building data clustering, which is an unsupervised data mining method. There are two types of data clustering that are often used in

the process of grouping data, namely hierarchical (hierarchical) data clustering and non-hierarchical (non-hierarchical) data clustering.

K-means is a non-hierarchical data clustering method that attempts to partition existing data into one or more clusters/groups. This method partitions data into clusters/groups so that data that has the same characteristics are grouped into the same cluster and data that has different characteristics is grouped into other groups. The purpose of this clustering data is to minimize the objective function determined during the clustering process, which generally tries to minimize variations within a cluster and maximize variation between clusters (Ade Bastian, *et al*., 2018) [2].

2. RESEARCH METHOD

The method used must be objective, critical, and able to analyze in order to find the right and correct process. The method must also be logical in order to produce good results. In this study the authors used a literature study by collecting theories obtained from journals and books that discuss problems in accordance with this research. Tables and Figures are presented center, as shown below and cited in the manuscript.

Steps to be taken in the completion of this study will be carried out from the beginning to the end pene l itian sequentially until the results to be found, by utilizing the method algorithm *K-Means Clustring*, adapaun of the steps can be seen in Figure 1 below:

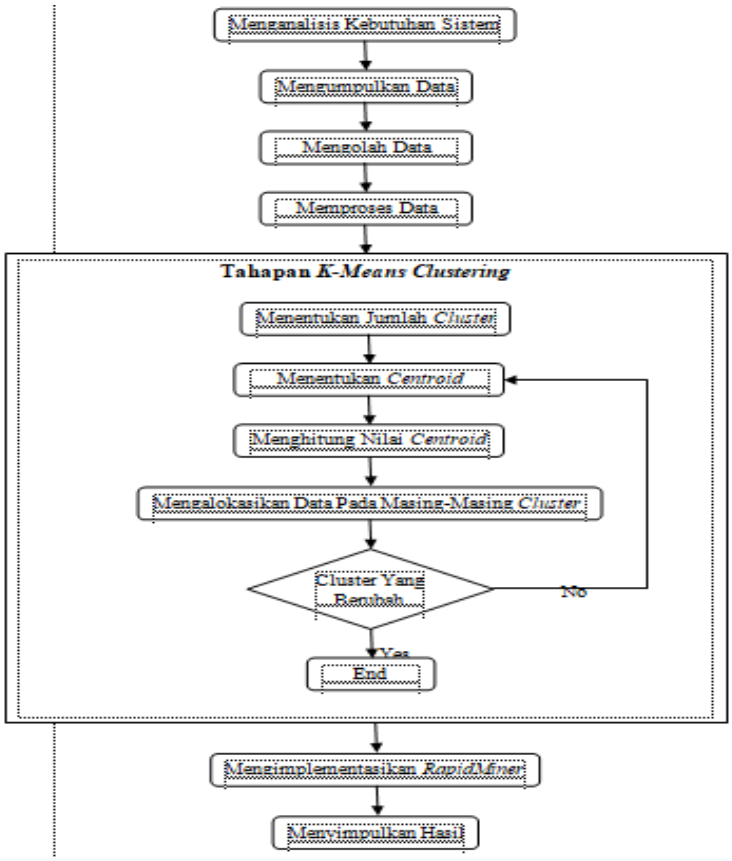

Figure 1. Research Framework

Based on the framework that has been compiled in the research in Figure 1, the steps can be described as follows:

1. Menga Nalis is Requirement System
   The step taken in the System Requirements Analysis is the process of analyzing the problems found by examining the position of determining the problem boundary line. By observing and analyzing the problems that have been determined, it is hoped that these problems can be understood properly.
2. Men gumpu k 's Data

The steps taken in data collection are collecting data from the STMIK Pelita Nusantara Medan library.

3. Processing data
   Before the book data is grouped into clusters that fit the criteria. The raw data will be transformed by initializing the data into numbers which can be processed in grouping.
4. Processing Data
   After the data is processed, the next step is that the data is processed to form a grouping of data into books according to predetermined criteria. The transformed data will be processed using a cluster algorithm, namely the K-means algorithm.
5. Stages of K-Means Cluster ing
   In the k-means clustering stage , there are steps taken in the process of completing this research, including:
   a. Determine the k value of the number of clusters to be formed. At this stage the cluster is formed by 3 clust.
   b. Determine the starting center (centroid) of each cluster . In this study, the starting center point was determined using the range of the title of the book.
   c. Allotment of all data in the nearest cluster . The distance from the object that is the decisive part of the object, wherein the distance between the data with a central cluster m erupakan determinant of the cluster with the data. Where in this process can be calculated the distance from each data to each cluster center . To be able to determine which data will go to which cluster , a certain cluster will determine it. For this process, the Euclidean distance theory formula can be used to calculate the distance from existing data aimed at each cluster center.
   d. Recalculate the center cluster with membership cluster that is, where the center of the cluster is the average of all data in custer particular. However, the data mean is not the only measure that can be used, but there are other alternatives that can be used, for example by using the median of the cluster .
   e. In the next process, if there is another change in the center of the cluster , it must be recalculated, where each data must use the new cluster center , but if no changes are found in the cluster center , the process can be considered complete.
6. Implementation of Rapid Miner
   The implementation process in this study is to calculate the coconut plantation data that has been obtained using the K-Means algorithm calculation with the Rapid Miner application .
7. Summing up the results
   In the last stage, you will get the results of the K-Means algorithm calculation using excel software and the Rapid Miner application . The result of these calculations is a grouping of books based on the results of the respective cluster processes.

3. RESULTS AND DISCUSSION

At t ahapan, the author will do an overview analysis and design of data from the library using the K-Means. After the data is arranged and grouped based on criteria, it will be transformed into groups of data to produce new knowledge. After grouping the data, the author will process the data using Microsoft Office Excel 2007 to assist in analyzing the data processing of book lending which will be tested using Rapid Miner 5.3 Software. Based on the framework discussed and described in chapter 3, the analysis and design stages follow the following flow and rules:

Figure 2. Analysis and Testing Flowchart

Based on the picture above, the first thing to pay attention to is data. The following will explain about analyzing data in the sub-chapters below.

3.1. Data Analysis

Data analysis that is processed in this research is book borrower data. Before carrying out the data mining process with K-Means, it is necessary to pre-process the data using Microsoft Office Excel. Data pre-processing includes data collection, and data cleaning.

3.2. *K-Means* Analysis

At the K-Means analysis stage, the Data Mining processing will be carried out using a predetermined algorithm. This is done by referring to the form of data and the objectives of the results to be achieved from the application of Data Mining. The steps in the cluster using the K-Means algorithm are:

1. **Determine the Number of *Clusters* (K)**
   In the process of implementing the K-Means method, we must first determine how many clusters we want, where the clusters will later become a reference for the results of decision making. This research process for the number of clusters will be divided into 3 clusters where later the 3 clusters will determine the results for research on which books are included in the low, sufficient , and high categories.
2. **Determining the point *Centroid* By *Random***
   The determination of the initial centroid in the calculations in this study was carried out by random or random , in which the researcher took steps, namely taking the lowest number, the highest number, and the highest number regarding the level of need for book borrowers at STMIK Pelita Nusantara . The correlation formula between three objects is to use the euclidean distance formula by determining the distance from each object to each centroid (D). To determine the distance from each object is taken from the book borrower annually, namely 2015 (V), 2016 (W), 2017 (X) , 2018 (Y), 2019 (Z) , with the formula:

$$d_{ij} = \sqrt{(v_i - v_j)^2 + (w_i - w_j)^2 + \ldots + (n_i + n_j)^2} \quad \ldots\ldots 1$$

   Where:
   $d_{i,j}$ = Distance between i and j
   v_i = coordinate v of the object
   v_j = coordinate of v center
   w_i = coordinate w object
   w_j = coordinate of w center
   x_i = the object's x coordinate
   x_j = center x coordinate
   y_i = the object's y coordinate
   y_j = center y coordinate
   n_i = Coordinate of n objects
   n_j = Coordinate of n center
   The amount of data will be divided into three clusters , then the next step is to randomly determine the centroid point.
3. **Calculating the Nearest *Centroid* Distance**
   The formula used to calculate the distance of each data to each centroid is to use the correlation formula between three objects, namely the euclidean distance formula, so that this formula will find the closest distance from each data to each centroid . The process of calculating the distance of each data to the center point of the centroid can be seen in the calculation below:

a. Iteration Process 1 (First)

Iteration 1 (first) is performed to calculate the distance of each data to each centroid using the euclidean distance formula.

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + \ldots + (n_i + n_j)^2} \quad \ldots\ldots 1$$

Where:
$d_{i,j}$ = Distance between i and j
x_i = the object's x coordinate
x_j = center x coordinate
y_i = the object's y coordinate

y_j = center y coordinate
n_i = Coordinate of n objects
n_j = Coordinate of n center

1) Process calculating Cluster (C0) to a category lower.

a) The process of calculating data distance 1 (D1) is taken from the title of the book " Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary " with serial number 1 which has the number of borrowers each year (2015: 3 , 2016: 1 , 2017: 1 0, 2018: 1, 2019: 0), which will be calculated with data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 0, 2017: 0, 2018: 2, 2019: 2) , with the following calculations:

$$D1 = \sqrt{(3-0)^2 + (1-0)^2 + (10-0)^2 + (1-2)^2 + (0-2)^2}$$
$$= \mathbf{11}$$

b) Stages of the data distance calculation process 2 (D2) are taken from the title of the book " Shell Programming in Linux " with serial number 2 which has the number of borrowers each year (2015: 10 , 2016: 3, 2017: 12 , 2018: 5, 2019 : 1), which will be calculated using data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 0, 2017: 0 , 2018: 2, 2019: 2), with the following calculations:

$$D2 = \sqrt{(10-0)^2 + (3-0)^2 + (12-0)^2 + (5-2)^2 + (1-2)^2}$$
$$= \mathbf{16}$$

c) Stages of the data distance calculation process 3 (D3) are taken from the name of the book " Architecture and Thermal Comfort " with serial number 3 which has a number of borrowers each year (2015: 6 , 2016: 4, 2017: 8 , 2018: 4, 2019 : 0), which will be calculated with data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 0, 2017: 0 , 2018: 2, 2019: 2), with the following calculations:

$$D3 = \sqrt{(6-0)^2 + (4-0)^2 + (8-0)^2 + (4-2)^2 + (0-2)^2}$$
$$= \mathbf{11}$$

a) Process calculating Cluster (C1) for the category enough .
The process of calculating data distance 1 (D1) is taken from the name of the book " Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary " with serial number 1 which has the number of borrowers each year (2015: 3 , 2016: 1 , 2017: 1 0, 2018: 1, 2019: 0), which will be calculated using data centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 1, 2017: 9 , 2018: 1, 2019: 2), with the following calculations:

$$D1 = \sqrt{(3-0)^2 + (1-1)^2 + (10-9)^2 + (1-1)^2 + (0-2)^2}$$
$$= \mathbf{4}$$

b) The process of calculating data distance 2 (D2) is taken from the name of the book " Shell Programming in Linux " with serial number 2 which has a number of borrowers each year (2015: 10 , 2016: 3, 2017: 12 , 2018: 5, 2019 : 1), which will be calculated using data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 1, 2017: 9 , 2018: 1, 2019: 2), with the following calculations:

$$D2 = \sqrt{(10-0)^2 + (3-1)^2 + (12-9)^2 + (5-1)^2 + (1-2)^2}$$
$$= \mathbf{11}$$

c) Stages of the data distance calculation process 3 (D3) are taken from the name of the book " Architecture and Thermal Comfort " with serial number 3 which has the number of borrowers each year (2015: 6 , 2016: 4, 2017: 8 , 2018: 4, 2019 : 0), which will be calculated with data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 1, 2017: 36 , 2018: 1, 2019: 6), with the following calculations:

$$D3 = \sqrt{(6-0)^2 + (4-1)^2 + (8-9)^2 + (4-1)^2 + (0-2)^2}$$
$$= \mathbf{8}$$

2) Process calculating Cluster (C2) to a category higher.

a) The process of calculating data distance 1 (D1) is taken from the name of the book " Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary " with serial number 1 which has the number of borrowers each year (2015: 3 , 2016: 1 , 2017: 1 0, 2018: 1, 2019: 0), which will be calculated using data centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 1, 2017: 36 , 2018: 1, 2019: 6), with the following calculations:

$$D1 = \sqrt{(3-0)^2 + (1-1)^2 + (10-26)^2 + (1-1)^2 + (0-6)^2}$$
$$= \mathbf{27}$$

b) The process of calculating data distance 2 (D2) is taken from the name of the book " Shell Programming in Linux " with serial number 2 which has a number of borrowers each year (2015: 10 , 2016: 3, 2017: 12 , 2018: 5, 2019 : 1), which will be calculated using data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 1, 2017: 36 , 2018: 1, 2019: 6), with the following calculations:

$$D2 = \sqrt{(10-0)^2 + (3-1)^2 + (12-36)^2 + (5-1)^2 + (1-6)^2}$$
$$= \mathbf{27}$$

c) Stages of the data distance calculation process 3 (D3) are taken from the name of the book " Architecture and Thermal Comfort " with serial number 3 which has the number of borrowers each year (2015: 6 , 2016: 4, 2017: 8 , 2018: 4, 2019 : 0), which will be calculated with data from centroid 1 which has data on the number of borrowers each year (2015: 0 , 2016: 1, 2017: 36 , 2018: 1, 2019: 6), with the following calculations:

$$D3 = \sqrt{(6-0)^2 + (4-1)^2 + (8-36)^2 + (4-1)^2 + (0-6)^2}$$
$$= \mathbf{30}$$

4. Grouping Data for Each Cluster

Performing calculations in the first iteration is carried out, then the distance from each data C0, C1, and C2 can be compared. In the results of clustering of each data, the calculation process is obtained in iteration 1. The results of this clustering are 49 for C0, 33 for C1, and 1 for C2.

a) Cluster members (C0) consist of 49 book titles with serial numbers (5, 8, 11, 14, 15, 18, 19, 21, 23, 25, 26, 28, 29, 30, 35, 37, 39, 40, 41, 44, 45, 46, 48, 50, 52, 53, 54, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 71, 72, 74, 75, 76, 77, 80, 83)

b) Cluster members (C1) consist of 33 book titles with serial numbers: (1, 2, 3, 4, 6, 7, 9, 10, 12, 13, 16, 17, 20, 22, 24, 27, 31, 32, 33, 34, 36, 38, 42, 43, 47, 49, 51, 70, 73, 78, 79, 81, 82)

c) Cluster members (C2) consist of 1 book title with serial number: (62).

a. Iteration Process 2 (Second)

The second iteration process is carried out after we get a grouping from the previous calculation results, where the new centroid point we have to calculate the average of the data that is in the same centroid . In this second iteration process, how to determine the new centroid can be seen as follows:

1) Process calculating Cluster (C0) numbered 49 title of the book with serial number: (5, 8, 11, 14, 15, 18, 19, 21, 23, 25, 26, 28, 29, 30, 35, 37, 39, 40, 41, 44, 45, 46, 48, 50, 52, 53, 54, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 71, 72, 74, 75, 76, 77, 80, 83)

$$C0(V2015) = \frac{(4+0+0+4+0+0+0+0+0+0+0+0+0+0+0+8+7+1+0+0+4+2+0+2+0+0+0+0+2+0+0+0+0+0+0+0+3+3+6+0+0+3+0+1+0+0+3+4+0+0)}{49}$$
$$= \mathbf{1}$$

$$C0(W2016) = \frac{(1+0+0+1+0+0+0+0+0+0+0+0+0+0+0+4+5+0+0+0+3+0+0+0+0+3+0+0+0+0+2+1+0+1+0+0+2+2+0+0+0+3+0+1+0+0+1+0+0+0)}{49}$$
$$= \mathbf{1}$$

C0(X2017)=(4+4+0+3+2+3+4+0+0+4+0+0+1+2+4+0+1+0+0+4+4+0+4+0+3+0+0+1+0+3+1+1+1+0+1+1+2+0+0+2+4+0+1+0+0+2+1+4+4)

49

= **2**

C0(Y2018)=(2+1+0+4+1+0+6+0+0+0+0+0+0+0+0+7+8+7+0+2+7+0+11+7+5+0+5+8+0+6+4+7+8+2+3+11+7+0+1+0+3+0+8+0+0+1+11+4+4)

49

= **3**

C0(Z2019)=(0+8+0+2+2+0+12+0+0+0+0+0+2+0+4+1+13+5+0+3+4+0+4+4+7+0+1+7+0+13+6+5+4+2+12+3+7+0+5+1+5+0+9+0+0+1+11+8+9)

49

= **4**

2) Process calculating Cluster (C1) numbered 33 title of the book with serial number: (1, 2, 3, 4, 6, 7, 9, 10, 12, 13, 16, 17, 20, 22, 24, 27, 31, 32, 33, 34, 36, 38, 42, 43, 47, 49, 51, 70, 73, 78, 79, 81, 82)

C1(V2015)=(3+10+6+5+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+10+2+0+3+2+6+0+1+4+6+1+1)

33

= **2**

C1(W2016)=(1+3+4+4+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+5+2+1+3+1+1+0+6+5+2+0+2)

33

= **1**

C1(X2017)=(10+12+8+6+8+10+13+16+13+6+7+14+6+6+8+10+11+10+10+5+14+10+9+9+6+7+5+8+7+8+10+9+14)

33

= **9**

C1(Y2018)=(1+5+4+4+1+0+0+3+0+1+0+3+0+1+0+2+0+5+1+1+0+1+11+1+9+0+0+3+0+5+4+0+2)

33

= **2**

C1(Z2019)=(0+1+0+2+0+0+2+7+3+1+8+1+1+0+1+3+4+3+0+2+3+4+17+2+4+2+2+2+5+7+2+2+3)

33

= **3**

3) Process calculating Cluster (C2) numbered 1 title of the book with serial number: (62)

C2(V2015)=(0)

1

= **0**

C2(W2016)=(1)

1

= **1**

C2(X2017)=(36)

1

= **36**

C2(Y2018)=(1)

1

= **1**

C2(Z2019)=(6)

1

= **6**

When the above calculation process steps have been completed, a new centroid will be obtained which will be used for the third iteration process. The new centroid can be seen in table 1. below:

Table 1. New Centroid

Centroid	2015	2016	2017	2018	2019
C0	1	1	2	4	4
C1	2	1	10	2	2
C2	0	1	36	1	6

4. CONCLUSION

Based on the results of the research and the results of the discussion in the previous chapter, that the results of the Identification of the Level of Book Needs in the STMIK Pelita Nusantara library , it can be concluded that: The K-Means algorithm method has been able to be applied to identify the level of book needs. The results of the implementation used in classifying book borrowers from the data used are 83 data, that for the low book lending category there are 52, for the sufficient category it is 30, and for the high category it is 1. Testing with m enggunakan software RapidMiner can mengh acyl right cluster 1, 2, and 3 show the results of the cluster category of Low, Self, and High. This is known because the results of testing manually and using the application produce the same thing in the calculation. From the results obtained in the overall process of data cluster 0 is a book borrower with low potential totaling 52 book titles, Cluster 1 is a potential borrower totaling 30 books, and Cluster 2 is a high potential book borrower totaling 1 book title, therefore knowledge is obtained. that books borrowed at STMIK Pelita Nusantara, books that are often borrowed will be updated.

REFERENCES

[1] Fauziyah, N . , " Knowledge Management Implementation in Library Information Systems (Case Study at the National Library of the Republic of Indonesia) " . *JIPI (Journal of Library and Information Science)* , *4* (1), 96-105 , 2019

[2] Gunawan, S , *et al* , " Implementation of K-Means, Suffix Tree and Dewey Decimal Classification for Shelving Library Books " . *Journal of Algorithms, Logic and Computing* , *2* (1) , 2019.

[3] Gustientiedina, G , *et al* , " Application of the K-Means Algorithm for Clustering Drug Data " . *National Journal of Technology and Information Systems* , *5* (1), 17-24 , 2019.

[4] Kusuma, AW, & Ellyana, R. L , " Application of Compressed Image in Image Segmentation Using K-Means Algorithm " . *Journal of Applied Information Technology* , *2* (1), 65-74 , 2018.

[5] Mahmuda, F , *et al* , " Clustering Library Visitor Profiles Using the K-Means Algorithm " . *Journal Of Applied Informatics And Computing* , *1* (1), 14-21. 2017.

[6] Maulida, L , " The Application of Datamining in Grouping Tourist Visits to Leading Attractions in Prov. Dki Jakarta with K-Means " . *JISKA (Journal of Informatics Sunan Kalijaga)* , *2* (3), 167-174 , 2018.

[7] Priyatman, H, *et al* , " Clustering Using the K-Means Clustering Algorithm to Predict Student Graduation Time " . *JEPIN (Journal of Education and Informatics Research)* , *5* (1), 62-66 , 2019.

[8] Purba, W , *et al* , " The Effect Of Mining Data K-Means Clustering Toward Students Profile Potential Drop Out Model. In *Journal Of Physics: Conference Series* (Vol. 1007, No. 1, P. 012049). IOP Publishing , April 2018.

[9] Putra, R. R, " Implementation of Data Mining Selection of Potential Customers Using the K Means Algorithm " . *INTECOMS: Journal Of Information Technology And Computer Science* , *1* (1), 72-77 , 2018 .

[10] Rosm , R , *et al* , "The Implementation of the K-Means Method in Mapping Student Groups Through Lecture Activity Data " . *IT Journal Research And Development* , *3* (1), 22-31 , 2018.

[11] Rustam, S , *et al* , " Optimization of K-Means Clustering for Identification of Endemic Areas of Infectious Diseases Using Particle Swarm Optimization Algorithm in Semarang City " . *ILKOM Scientific Journal* , *10* (3), 251-259 , 2018.

[12] Subianto , M., & Fitriana, A. R , "The Pattern of Borrowing Books at the Syiah Kuala University Library Using the Eclat Algorithm " . *Periodic Library and Information Science* , *14* (1), 35-44 , 2018.

[13] Windarto, A. P , " Implementation Of Data Mining On Rice Imports By Major Country Of Origin Using Algorithm Using K-Means Clustering Method " . *International Journal Of Artificial Intelligence Research* , *1* (2), 26-33 , 2017.

[14] Zuha, K , "The Effectiveness of the Application of Administrative Sanctions in Improving Discipline of Users in the Library of the State University of Padang " . *JIPI (Journal of Library and Information Science)* , *4* (1), 52-67 , 2019.