K-MEANS ALGORITHM DATA MINING
IN SALES LEVEL ANALYSIS

Dodi Nofri Yoliadi
Fakultas Ushuluddin Adab dan Dakwah, UIN Mahmud Yunus Batusangkar
Jl. Jenderal Sudirman No.137, Limo Kaum, Kec. Lima Kaum, Kabupaten Tanah Datar, Sumatera Barat 27217
E-mail : dodinofriyoliadi@uinmybatusangkar.ac.id

ABSTRAK- In this research, data collection and processing of sales of electronic goods were carried out with CV. Berkah Elektronik. The data obtained is then aggregated with the K-Means algorithm to gain knowledge about which electronic products are selling well on the market and which are not. In this study, the clustering method was used with Tanagra 1.4.48 software, and the K-Means algorithm was used as an algorithm to draw conclusions about which items were selling well and which were not inputted, namely product prices. goods and sale of goods. And from the results of testing and manual testing with the Tanagra application, the same clusters are produced, namely, products that do not sell well (Cluster_KMeans_1) and products that sell well (Cluster_KMeans_2).

Keywords : Clustering, Sales, K-Means Algorithm, Data Mining, Tanagra.

1. INTRODUCTION

Data mining is a technique used to find information in a database. Data mining usually uses statistical techniques, artificial intelligence, mathematics, and machine learning to extract and identify useful information and aggregate it from large databases [1]. Data mining is the process of extracting data from a very large and unlimited collection of databases. In fact, data mining is a KDD (knowledge discovery in databases) step. Information services are activities of several processes, including data cleaning, integration, data selection, transformation, data mining, pattern assessment, and data presentation [2]. Currently, computerized systems have been widely used in various business companies to obtain important information about marketing activities. One company that implements this computing system is a business that is in the field of selling electronic goods.

The function of data processing for these companies is to obtain precise, accurate, and easy-to-understand information about sales of electronic goods. To obtain this information, various techniques are used, one of which is the clustering technique available in data mining, namely the K-Means clustering algorithm. Among several clustering techniques, the k-means clustering algorithm is the simplest and easiest technique. In this technique, objects are grouped into k groups (clusters). For merging, the value of k must be clustered first. K-Means is a clustering method that is easy, simple, and easy to implement because it has the advantage of being able to group large amounts of data and create clusters quickly, precisely, and accurately. The goal of the K-Means algorithm is to cluster objects in such a way that the distance of each object to the center of the cluster is minimal. Data grouping can be applied to analyze the level of sales of electronic goods in CV. The sooner the level of sales of electronic goods is known, the better because the company can make improvements. Analyzing the sales level of available products by carrying out the K-Means clustering algorithm technique is expected to provide accurate and precise information about the types of electronic goods that are most sold and purchased and are in demand by buyers (customers)..

2. THEORY STUDY

2.1 Data Mining

Data mining is a technique used to find information in a database. Data mining usually uses statistical techniques, artificial intelligence, mathematics, and machine learning to extract and identify useful information and aggregate it from large databases [1]. Data mining refers to the process of extracting information from very large data sets. In fact, data mining is a step towards KDD (knowledge discovery in databases). Information services as a process consist of data cleaning, integration, data selection, transformation, data mining, pattern assessment, and data presentation. [7].

2.2 Data Mining Cycles

According to CRISP-DM, the data mining life cycle is divided into six stages. All sequential steps are adaptive; the next step in the sequence depends on the output of the previous step. 6 in the CRISP-DM stage, namely [1]: 1) Steps in the Business Understanding Phase 2) Steps in the Data Understanding Phase Steps in the Data Processing Phase 4) Steps in the Modeling Phase Steps in the Evaluation Phase 6) Step into Deployment Phase

2.3 Clustering

Clustering is a well-known and widely used technique in data mining. Researchers in the field of data mining are still making various efforts to perfect the cluster model because the development method is

still heuristic. The calculation of the optimal number of clusters and the best clusters is still in progress. Therefore, with the current method, it cannot be guaranteed that the clustering results will be optimal. However, from a practical point of view, the results achieved are usually quite good [2]. The main objective of the clustering method is to group a set of data or objects into groups (clusters) in such a way that each cluster contains as much similar data as possible. In clustering, we try to place similar (closest) objects in a cluster and make the distance between clusters as wide as possible [2].

2.4 Cluster Analysis

One of the data mining methods used in this study is clustering, in which the method identifies objects with certain general characteristics and then uses these characteristics as eigenvectors or centroids [6]. Cluster analysis finds a collection of objects in such a way that objects in one group are similar (or related) to another and different (or unrelated) to objects in another group [3]. The goal of cluster analysis is to minimize the distance within a cluster and maximize the distance between clusters [3]. Clustering can be applied to data that is quantitative (numeric), qualitative (categorical), or a combination of both. Data can come from the results of process observations. Each observation can have n measurement variables and n-dimensional vectors grouped as $Zk = [Z1k,..., Znk]T,Zk \ Rn$. The set of observations N is given by $Z = Zk \mid$ denotes $k = 1, 2,...N$ and is represented as an $n \times N[1]$ matrix.

2.5 K-Means Algorithm

The K-Means algorithm is the most popular and widely used clustering algorithm in the industrial world. This algorithm is based on a simple idea. First, determine how many clusters are formed. Every object or first element of a cluster can be chosen to act as the center of the cluster [10]. Several alternatives for implementing K-means have been proposed, with several theories of computational evolution associated. This includes options [12]:
a. Distance space to calculate the distance between the data and the centroid.
b. methods for mapping data back to each cluster;
c. the purpose for which it is used. K-Means belongs to the partition grouping.

All data must be included in a particular cluster, and all data belonging to a certain cluster can be transferred to another cluster in one process step, in the next step. K-Means divides the data into k distinct ranges, where k is a positive integer. The K-Means algorithm is known for its simplicity and ability to classify large amounts of data and outliers very quickly [4].

2.6 K-Means Technique

K-Means takes k input parameters and divides a set of n objects into k clusters, resulting in high intra-

cluster similarity but low inter-cluster similarity. Cluster similarity is measured relative to the average of objects in the cluster, which can be seen as the cluster's center of gravity or center of mass. The K-Means algorithm then repeats the following steps until stability is achieved (obstacles cannot be moved):
a. Determine the center coordinates of each cluster;
b. Determine the distance of each object to its center coordinates;
c. Group these objects according to their minimum distance.

3. RESEARCH METHOD

The research methodology includes steps that must be taken to facilitate dissertation work and can also be used as a guide for researchers conducting research. Research is a series of scientific activities aimed at solving a problem. The task of research is to find explanations and answers to problems and show possible alternatives that can be used to solve them. A frame of reference is needed for research steps to be carried out in a structured manner, creating research methodology steps to maximize the results obtained. This research framework can be seen in the image below:
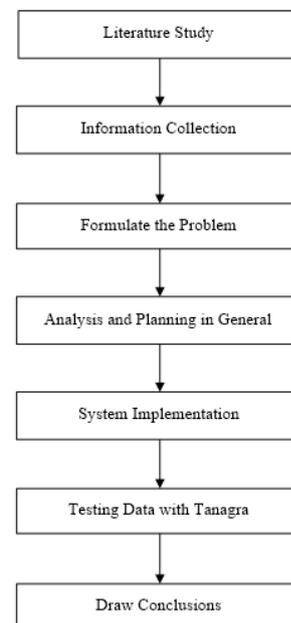

Figure 1. Research Framework

3.1 Literature Study

At this stage, the researcher reviewed the literature on data mining, clustering, and the k-means algorithm. The literature studied came from various sources, such as textbooks, journals, websites, essays, e-books, and academic articles.

3.2 Information Collection

At this stage, data is collected from CVs through direct observation. At Berkah Electronics, the data collected is sales data. After the data has been collected and researched, the next step is to classify

the sales data, and the results of grouping the sales data can be used to decide on problems to be solved and solutions to be sought.

3.3 Formulate the Problem

After the data collection stage is complete, the next step is to formulate questions about the existing problems. How to make decisions about determining the level of sales of electronic products in the CV Berkah Electronics uses the K-Means algorithm based on sales data.

3.4 Analysis and Planning in General

This step describes analytical methods commonly used in design to determine the best-selling electronic products.

3.5 System Implementation

In this phase, programs created with the Tanagra programming language are run to analyze and evaluate the results of the K-Means algorithm method. At this stage the researcherstage,iewed the literature on data mining, clustering and the kclustering,gorithm. The literature studied was taken from various sources, such as textbooks, journals, websites, books, e-books, and academic books.

3.6 Testing Data with Tanagra

The next step after programming is system testing. In this phase, a series of tests are carried out to ensure that the designed system matches the designed model. The steps in the data mining testing mechanism are as follows:

a. **Selection of information**
    Selection (selection) of data from the operational data set must be done before the data mining phase in KDD begins. Selected information used in the data mining process is stored in files separate from the operational database.

b. **Pretreatment/cleaning**
    Before the data mining process can be carried out, it is necessary to carry out a cleaning process for the data that is the focus of KDD. The cleaning process includes, but is not limited to, removing duplicate data, checking for inconsistent data, and correcting errors in the data, such as B. KDD typos, etc. B. external data or information.

c. **Transformation/Change**
    Coding is the process of transforming the selected data so that it is suitable for the data mining process. The KDD coding process is a creative process and depends heavily on the type or model of data retrieved from the database.

d. **Interpretation/Evaluation**
    The data model created by the data mining process must be presented in a form that is easily understood by interested parties. This step is part of the KDD process called rendering. In this phase, it is examined whether the formula or information found contradicts the facts or hypotheses that existed before.

3.7 Draw conclusions

At the final stage of the research, conclusions will be drawn whose aim is to compare the results of the manual system implementation phase with the K-Means clustering algorithm method.

4. RESULTS AND DISCUSSION

4.1 Data Analysis

CV. Berkah Elektronik has information about the company's sales activities. One of them is the information used to record transactions that occur within the company. consisting of several attributes such as electronic brand, type of goods, name of goods, price of goods, and number of goods sold. Data is used as an attribute to perform data processing. The definition of a best seller (an item that sells in large quantities), while the notion of a no seller is an item that is not selling well (not selling) or has a sales volume below the average.

4.2 K-Means Algorithm for Clustering Analysis

K-Means is included in the partitioning method of data mining; all data must be entered in a certain cluster, and all data belonging to a certain cluster can be transferred to another cluster in one process step, in the next step. K-Means divides the data into K discrete ranges, where K is a positive integer. The K-Means algorithm is known for its simplicity and ability to classify large data sets and outliers very quickly. The following is a flowchart of the K-Means algorithm and explains the steps of the algorithm, assuming the input parameters are the number of records in the data and the number of initialization centers. K = 2 according to research:
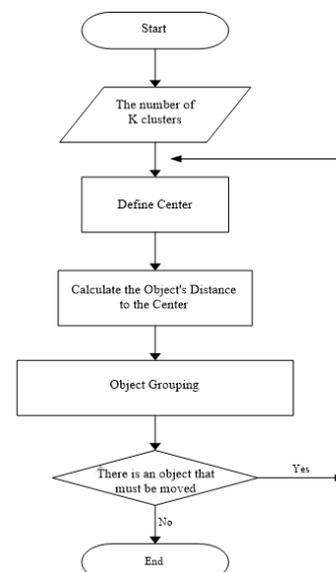

Figure 2. K-Means Process Flowchart

a. **Determination of the number of clusters**
In this first stage, the number of clusters is determined using the data obtained.

b. **Define the center of the first cluster**
In determining the n initial cluster centers, random numbers are generated that represent the input data sequence. The initial center of the cluster is obtained from the data itself, not by defining a new point, namely by randomizing the initial center of the data.

c. **Calculate the object's distance to the cluster center**
To measure the distance between data and cluster centers, Euclidean distance is used, which is an algorithm for calculating the distance between data and cluster centers. Take the data value and cluster center value, then calculate the Euclidean distance to each cluster center.

d. **Grouping Data Objects**
The calculated distance is compared, and the shortest distance between the data and the cluster center is selected; this distance indicates that the data is in the same group as the nearest cluster center. The data grouping algorithm is as follows: a) Take the value of the distance from the center of each cluster to the data. b) Find the smallest distance value. c) Grouping the data based on the center of the cluster with the closest gap.

e. **Establishment ff A New Cluster Center**
To get a new cluster center, it can be calculated from the average of cluster members and cluster centers. The new cluster center is used to carry out the next iteration when the results obtained have not converged, stopping the iteration when the maximum number of iterations entered by the user is reached or the results obtained have converged (the new cluster center is the same as the old cluster center). The algorithm used to determine the cluster center is: Find the number of members in each cluster, and then calculate the new center using Formula 2.1.

f. **Calculate the Distance To the Center Of The Cluster**
Calculate the Euclidean distance from all data to the new centers (C1 and C2) as in step 2. After we get the calculation results, compare these results.

4.3 Interpretation/Evaluation
Interpret patterns resulting from data mining, and then check whether the patterns or data found contradict the facts or hypotheses that existed before.

4.4 The Clustering Process Uses The K-Means Algorithm
In this phase, the main process is carried out with CV brand-based merchandise sales data. Electronics Blessings, that is, segmentation or grouping of goods sales data seen in sales reports, which is the K-Means algorithmic grouping method. The collection of sales data obtained from sales reports is used as an application of the K-Means algorithm in selling goods. Experiments were carried out with the following parameters:
1. total clusters: 2;
2. total records: 37;
3. total attributes: 5

a. **Determination of the Number of Clusters**
Determine the number of clusters obtained from existing sales data. The number of clusters in this case is three, namely brand, type, and type.

b. **Initial Cluster Center Determination**
The initial center of the cluster or centroid is obtained randomly or randomly based on existing sales data; for the initial determination of the cluster, it is assumed:
1. Cluster center 1: (25, 6)
2. Cluster center 2: (40, 2)

c. **Calculation of Object Distance to Cluster Center**
To measure the distance between the data and the cluster center, Euclidean distance can be used. Of the 37 data points prepared to serve as samples, the selection of the initial cluster centers was carried out, namely C1 = (25, 6), and C2 = (40, 2). Then it will calculate the distance from the remaining sample data to the cluster center, for example with M.

4.5 System Implementation
Testing the results of the analysis is very important to find out and verify whether the results are correct or not. Test the accuracy of the results of manual data processing. The stages of testing carried out in determining the level of sales of electronic goods at CV. Berkah Elektronik in testing sales data samples using one of the Tanagra 1.4.48 software With the following steps:

Figure 3. Database View

After determining the database to be used, we will start implementing the Tanagra 1.4.48 application.
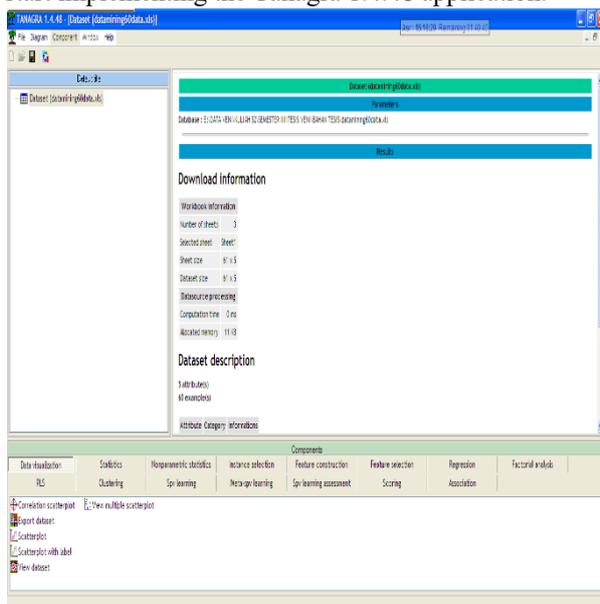
Figure 4. Dataset Display

After the dataset is displayed, add DEFINE STATUS and enter the price and amount inputs into Define Attribute Statuses. In the same way, enter the target brand, type, and type into Define Attribute Statuses.
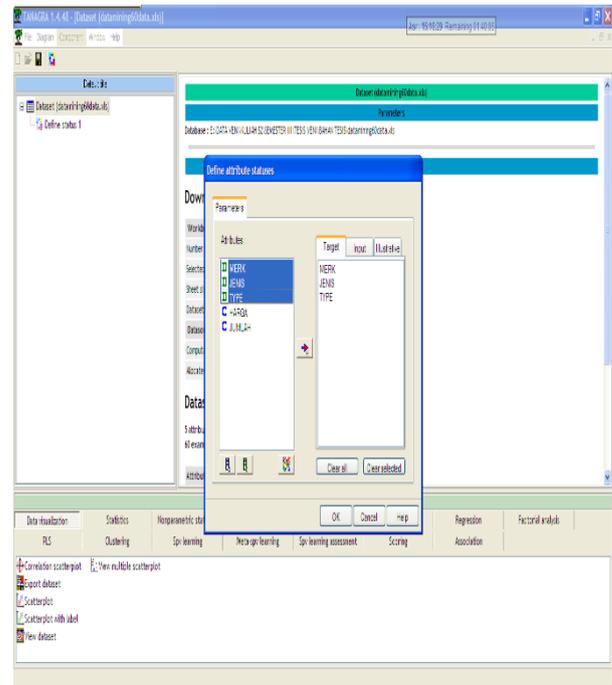

Figure 5. Display Define Status 1 Target

In the next stage, we will see Clustering on the menu at the bottom, then select K-Means, then drag towards Define Status 1, then K-Means under Define Parameter Parameters, then display the K-Means Parameters screen to display the K results for K-Means in the Data Visualization menu, Scatterplot;
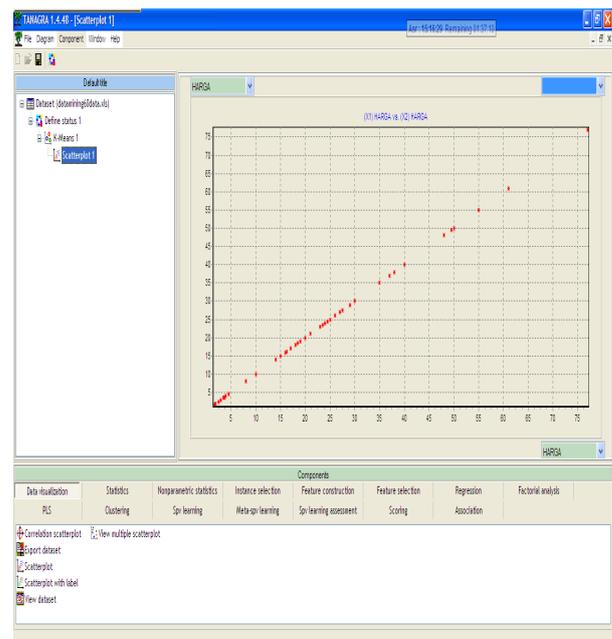
Figure 6. View Scatterplot display

The next step is the Data Visualization menu. On the menu will be Export Dataset 1 and place it under K-Means 1. To see the results of Export Dataset 1 obtained by Export Dataset 1, select parameters and then set where to store the results or output from the

process that has been carried out starting from start to finish with the result as shown below:
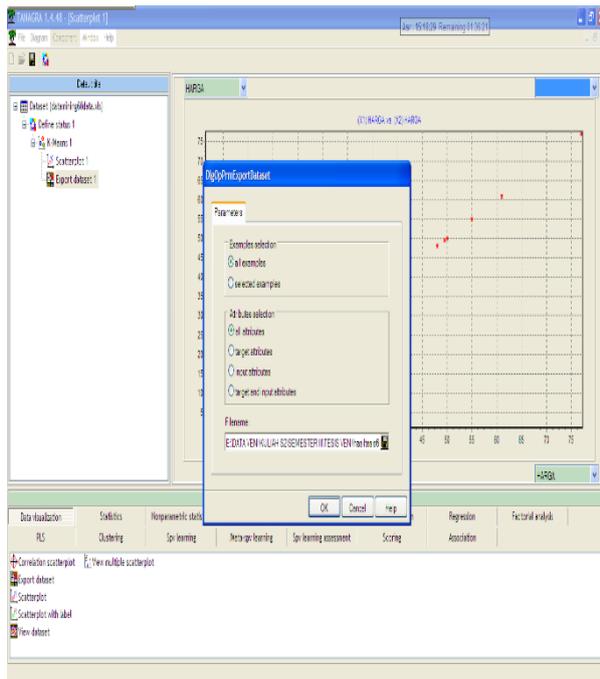
Figure 7. DIg Op Prm Export Dataset view

After determining the storage location, if you want to see where the output results are stored, you can do it easily by viewing Export Dataset 1, as shown below:
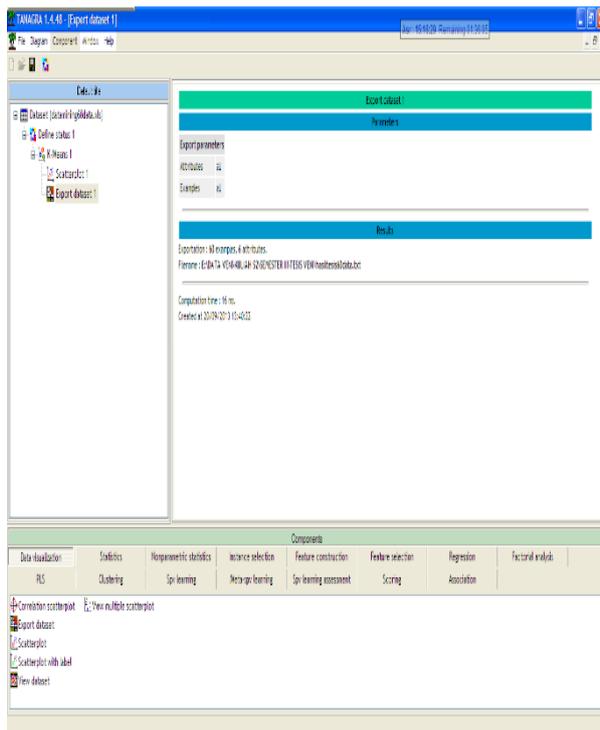
Figure 8. View Export Dataset 1

In the next stage, we will see the output of the processed data.

Figure 9. Display Output for 60 Records

Figure 10. Display Output for 60 Records (Continued)

Where the output produced in Figure 10. is
1. If Cluster_KMeans_1 or c_kmeans_1 = then sales level = sellout
2. If Cluster_KMeans_2 or c_kmeans_2 = not selling, then sales level = not selling.

4.6 Discussion of Test Results

From the stages of testing that have been carried out by following all the processes manually and with the Tanagra 1.4.48 software, it can be concluded that the test results are the same as the centroid score and the clustering results are produced by Tanagra 1.4. 48 software is the same as using manual calculations. Likewise testing using all data (60 records). the results of the manual calculation of 60 data records that produce three iterations of the manual calculation of Euclidean Distance one and Euclidean Distance two, it

is obtained that the number of Cluster 1 is 12 data records, and for Cluster 2 there are 28 data records. After manual calculations, tests were carried out using Tanagra 1.4.48 software to compare the results obtained. In testing the Tanagra 1.4.48 software from 60 records of sales data, the results for Cluster 1 were 12 data, and for Cluster 2 there were 28 data. So it can be said that manual calculations are tested using all data (60 records). on Tanagra 1.4.48 software has the same result. The results of the first cluster test of the 60 data that were carried out gave no less than 36 objects presented in the following table:

Table I. First Cluster Results

NO	MERK	PRODUCT	TYPE	PRICE	AMOUNT
1	LG	TV	21FU6RLR	14	52
2	LG	TV	14FU7ABF	8	13
3	LG	TV	19LH20RC	25	6
4	LG	DVD	DV-452	4,5	14
5	LG	KULKAS	GN-VL12SL-1K	23	5
6	LG	MESIN CUCI	WF-15CR-1T	30	11
7	LG	MESIN CUCI	WP-60IN-2T	20	3
8	PANASONIC	BLENDER	MX-J1G	3,95	1
9	PANASONIC	KIPAS ANGIN	FEU-409	3,7	8
10	PANASONIC	KULKAS	NR-A17KX-1P	23	10
11	PANASONIC	MESIN CUCI	NAW-70BC1-2T	18,5	19
12	POLITRON	AC	PLE-686F-1	29	9
13	POLITRON	DISPENCER	PWC-107F	15	10
14	POLITRON	DVD	DVD-2183	4,5	18
15	POLITRON	KULKAS	PR-211SM-2P	24,5	1
16	POLITRON	KULKAS	PR-169VSB-1P	17	1
17	POLITRON	MESIN CUCI	PAW-6T205-2T	19	1
18	POLITRON	TV	MIX-1403	10	1
19	POLITRON	TV	PS-21UM	16	13
20	POLITRON	TV	APS-29UM	27	3
21	SHARP	AC	AH-A9KCY-1	27	5
22	SHARP	KULKAS	SJ-G170TZB-1	16,25	24
23	SHARP	KULKAS	SJ-G200UZS-2	29	11
24	SHARP	TV	21DXS888	17	20
25	SHARP	TV	29DXS200	26	5
26	TOSHIBA	MESIN CUCI	VH-8200E-1T	21	1
27	TOSHIBA	KULKAS	GR-H240PD-2P	30	1
28	PHILIPS	BLENDER	HR-2071	4	2
29	PHILIPS	SETRIKA	HI-114	2	2
30	RINAI	KOMPOR GAS	RI-522A	3	1
31	SAMSUNG	TV	CS-29M21ML	21	1
32	SAMSUNG	AC	AS-05RLN-1	28	4
33	SONY	CAMERA DIGITAL	DSC-W180-B	18	9
34	SONY	CAMERA DIGITAL	DSC-W210-C	24	24
35	SONY	PLAYSTATION	SCPH-90006-1	14	8
36	TECHTRON	DVD	P-DVD9500	18	1

The results of the second cluster when testing 60 data points produce 12 items presented in the following table:

Table 2. Second Cluster Results

NO	MERK	PRODUCT	TYPE	PRICE	AMOUNT
1	LG	AC	CT-07LCS-2	40	2
2	LG	AC	ST-05ICE-1.5	35	37
3	LG	KULKAS	GN-B352YLC-1B	40	1
4	LG	KULKAS	GN-M392YPC-2	55	9
5	PANASONIC	AC	CS-C10KKP-2	49,5	3
6	PANASONIC	KULKAS	NR-B203G-2P	37	2
7	PANASONIC	TV	TH-L32C10X	48	13
8	PANASONIC	TV	TH-L37X10	77	2
9	SHARP	AC	AH-AP5HHY-1.5	38	15
10	SHARP	AC	AH-AP9KHL-2	40	2
11	SONY	HANDYCAM	DCR-SX60E	50	2
12	SONY	MINI COMPO	MHC-GN1300D	61	1

The K-Means algorithm is considered an algorithm that helps in grouping data whose properties can be clearly recognized and makes it easier for users to extract information from relevant data.

5. CONCLUSION

From the descriptions and explanations that have been discussed in the study, several conclusions can be drawn, including: With the K-Means algorithm, determining the level of sales of electronic products is easier than using the manual method. The resulting grouping is used to conclude electronic goods sales data. Berkah Electronics uses the K-Means algorithm to determine which products are performing well and which are not. The selection of variables and attributes to be used has a significant impact on the data used. The development of this system from a manual or data system to a computerized system is carried out from the perspective of problems that arise from the old system.

Here are suggestions to consider for the future: In this study, we experimented with one of the existing data mining clustering methods, namely the K-Means algorithm. In the future, it will be necessary to develop more in-depth research to compare results and find out which theory gives good results. The software used is Tanagra 1.4.48. Although this software is easy to use, you first need to understand how to use it.

REFERENCES

[1]. Kusrini & Emha Taufiq Luthfi (2019), "Algoritma Data Mining", Andi Yogyakarta, Yogyakarta

[2]. Santosa,, Budi (2017), "Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis", Graha Ilmu, Yogyakarta, Yogyakarta

[3]. Hermawati, Fajar Astuti (2013), "Data Mining", Andi Yogyakarta, Yogyakarta

[4]. Wijaya, Arim (2019), "Jurnal Analisi Algoritma Means Untuk Sistem Pendukung Keputusan Penjurusan Siswa Di MAN BINONG SUBANG"

[5]. Dana Sulistiyo Kusumo, Moch. Arief Bijaksana, Dhinta Darmantoro "Jurnal Data Mining Dengan Algoritma Apriori Pada RBMS Oracle"

[6]. Irdiansyah, Enur (2013), "Jurnal Penerapan Data Mining Pada Penjualan Produk Minuman di PT.Pepsi Cola Indobeverages Menggunakan Metode Clustering "

[7]. Mewati Ayub (2017), "Jurnal Proses Data Mining Dalam Sistem Pembelajaran Berbantuan Komputer"

[8]. Wakidah, Nur (2015), "Jurnal Clustering Menggunakan K-Means Algorithm (K-Means Alorithm Clustering)"

[9]. Pramudiono, Iko (2013), "Pengantar Data Mining : Menambang Permata Pengetahuan di Guning Data"

[10]. Andayani, Sri (2017), "Pembentukan Cluster ke Dalam Knowledge Discovery in Database Dengan Algoritma K-Means"

[11]. Sunjana (2014), "Jurnal Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi Decision Tree"

[12]. Agusta, Yudi (2015), "Jurnal K-Means Penerapan, permasalahan dan Metode Terkait "

[13]. M. Abinaya and R. U. Devi, "Intelligent Vehicle Control Using Wireless Embedded System in Transportation System Based On GSM and GPS Technology," vol. 3, no. 9, pp. 244–258, 2014.

[14]. D. Hartanti, R. N. Aziza, and P. C. Siswipraptini, "Optimization of smart traffic lights to prevent traffic congestion using fuzzy logic," Telkomnika (Telecommunication Comput. Electron. Control., vol. 17, no. 1, pp. 320–327, 2019, doi: 10.12928/TELKOMNIKA.v17i1.10129.

[15]. P. C. Siswipraptini, W. H. Martono, and D. Hartanti, "Reducing a congestion with introduce the greedy algorithm on traffic light control," J. Phys. Conf. Ser., vol. 974, no. 1, 2018, doi: 10.1088/1742-6596/974/1/012013.

[16]. J. Y. Kim, H.-J. Lee, J.-Y. Son, and J.-H. Park, "Smart Home Web of Objects-based IoT Management Model and Methods for Home data mining," 2015.

[17]. S. Ghosh, "Smart homes: Architectural and engineering design imperatives for smart city building codes," in International Conference on Technologies for Smart City Energy Security and Power: Smart Solutions for Smart Cities, ICSESP 2018 - Proceedings, 2018, vol. 2018-Janua, pp. 1–4, doi: 10.1109/ICSESP.2018.8376676.

[18]. D. E. Paul, "Smart Energy Meter Using Android Application And Gsm Network," Int. J. Eng. Comput. Sci., vol. 05, no. 16058, pp. 16058–16063, 2016, doi: 10.18535/ijecs/v5i3.36.

[19]. B. Lashkari, Y. Chen, and P. Musilek, "Energy management for smart homes-state of the art," Applied Sciences (Switzerland), vol. 9, no. 17. 2019, doi: 10.3390/app9173459.

[20]. M. Samarth Pandit, M. Snehamandhre, and M. Meghananichal, "Smart Energy Meter using internet of Things (IoT)," J. Eng. Res. www.vjer.in, vol. 1, no. 2, pp. 222–229, 2017.

[21]. Y. S. Tak, J. Kim, and E. Hwang, "Hierarchical querying scheme of human motions for smart home environment," Eng. Appl. Artif. Intell., vol. 25, no. 7, pp. 1301–1312, Oct. 2012, doi: 10.1016/j.engappai.2012.03.020.

[22]. S. Z. Reyhani and M. Mahdavi, "User authentication using neural network in smart home networks," Int. J. Smart Home, vol. 1, no. 2, pp. 147–154, 2007.