CLUSTERING ANALYSIS WITH K-MEANS FOR ELECTRONIC SALES CLASSIFICATION IN BATAM IT STORE COM

Anggia Dasa Putri*[1], Tukino[2]
[12]Putera Batam University
Corresponding Email: anggia.dasa@puterabatam.ac.id

Vol. 18 No.1 2024

Submit :
10/08/2023

Accept :
23/03/2024

Publish :
02/04/2024

Abstract

This study aims to assist Batam IT Store Com in refining its sales categorization process and identifying patterns in product demand. The objective is to derive valuable insights from the store's sales data for future decision-making. The research employs the K-means Clustering method, which involves grouping data into distinct clusters through a calculation process. This results in the identification of products that are highly popular, best-selling, and those with low demand. The electronic sales data used for this study spans from March 2020 to November 2023. The findings reveal three clusters: Cluster 0 comprising 19 items, Cluster 1 with 2 items, and Cluster 2 with 4 items. In conclusion, the analysis and grouping facilitated by the K-means Clustering method prove beneficial for segmenting sales data at Batam IT Store Com. The outcome is a renewed understanding of product clusters that aids in analyzing and segmenting goods sold at the store.

Keywords: Clusters, Analysis, Data Mining, Electronic, K-Means Algorithm.

ISSN :1979-9292 / 2460-5611
JIT is Sinta 3 Journal (https://sinta3.kemdikbud.go.id/journals/profile/2143) accredited by Ministry of Research & Technology, Republic Indonesia

INTRODUCTION

In the current age of rapid technological advancement, electronic devices play a crucial role in supporting various aspects of daily life, even though they may not be considered primary necessities. The heightened societal reliance on electronic equipment has resulted in robust growth in sales within this sector. Notably, data cables represent a rapidly selling product category, particularly for charging mobile phones, meeting diverse needs in daily life. According to data from the Ministry of Industry of the Republic of Indonesia, a significant portion, 60%, of the estimated imported products in the country belongs to the electronics sector. Furthermore, Indonesia imports raw materials primarily from China, Taiwan, Singapore, South Korea, and Vietnam, contributing to the substantial value of imports in this sector. The increased investment in Indonesia's electronics sector has fueled this growth.

In this context, Batam IT Store Com emerges as a key player in the electronic equipment retail sector. Despite accumulating substantial data from sales, the current practice involves storing data without further processing using supporting algorithmic methods. Typically, data is retained until the storage medium is full, after which it is deleted. Addressing this issue, this research focuses on data analysis, employing data mining methods, specifically clustering techniques. The chosen method is the k-means method, which involves grouping data with similar characteristics and subsequently evaluating the created clusters. This approach aims to extract meaningful insights from the abundant data amassed by Batam IT Store Com, contributing to informed decision-making and potential improvements in sales strategies.

In 1989, the name Knowledge Discovery in Database was first put forward to signify the skills that can be obtained from data mining in a database which aims to mine important data from within the database [1] .

There are 5 stages of KDD [2].

1) *Data Selection*
   At this stage, data is selected from all the data that will be analyzed further.
2) *Pre-processing*
   At this stage, the data will enter a cleaning process to filter the data that will be used in the calculation process.
3) *Transformation*
   Transformation is carried out to change the data format into a format that can be processed in the data mining process.
4) *Data Mining*
   This stage is the data stage where the calculation process is carried out according to a predetermined algorithm.
5) *Evaluation*
   At this stage, the processing data is evaluated using a processing application to ensure the validity of the results obtained.

*Data Mining*is a procedure that uses data on a large scale to explore valuable information contained in a collection of data which can later be used to assist the process of making crucial decisions [3].

*Clustering*is a procedure commonly used to divide data into several clusters with characteristics that are similar and different from other clusters [4].

*K-Means*works by collecting data by repeatedly finding the closest distance to the centroid of each cluster from several existing clusters. Calculation of centroid points can be done using the Euclidean Distance formula [5].

$$d\,(x,y) = \sqrt{\sum_{i=0}^{n} (xi - yi)^2}$$

Formula 1. Euclidean Distance Formula

This distance measure is commonly used in various fields, such as mathematics, physics, and machine learning, to quantify the spatial separation between points.

To find the distance between the object point in the form of x and the centroid point in the form of y, the attribute number i will be calculated from the square of the results of subtracting point x and point y [6].

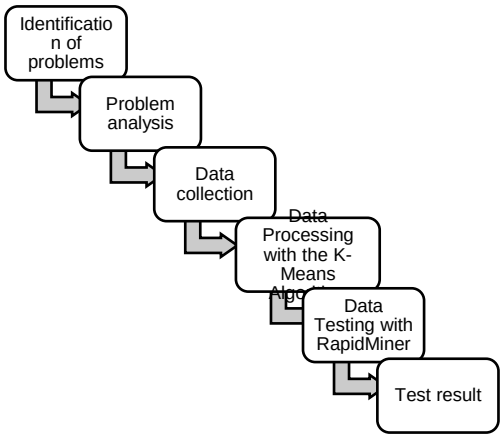
The RapidMiner application is an open-source application designed to have its own framework for the data analysis process and can be integrated into other applications to carry out data mining activities [7].

Sales is the activity of exchanging goods between the seller and the buyer with the aim of meeting daily needs. Sales activities are considered valid if there is an agreement between both parties [8].

Electronics are tools that can do their work with high accuracy by themselves, thereby reducing manual work [9].

RESEARCH METHODS

The research was conducted using quantitative research methods. An overview of the research methods carried out by the author can be seen in the picture below.



(Source: Research Data, 2023)

Picture 1. Research design

1. Identification of Problems:
   Determine the issues or challenges that you aim to address or understand through data analysis. This step involves defining the problem statement.
2. Problem Analysis:
   Analyze the identified problems in-depth. Understand the context, underlying causes, and potential impacts of these problems.
3. Data Collection:
   Gather relevant data that is necessary to address or analyze the identified problems. This may involve acquiring datasets from various sources, ensuring the data's accuracy and completeness.
4. Data Processing with the K-Means Algorithm:
   Utilize the K-Means algorithm for data processing. K-Means is a clustering algorithm that partitions data into distinct groups based on similarities. It is commonly used for segmentation and pattern recognition.
5. Data Testing with RapidMiner:
   Use RapidMiner, a data science platform, to perform testing on the

processed data. RapidMiner provides tools for various data analysis tasks, including visualization, modeling, and evaluation.

6. Test Result:
   Evaluate the results obtained from the data testing process. This involves interpreting the findings, assessing the effectiveness of the K-Means algorithm in addressing the identified problems, and drawing conclusions.

It's important to note that the success of the process depends on the quality of data, the appropriateness of the chosen algorithm (in this case, K-Means), and the effectiveness of the testing and evaluation methods. Additionally, clear communication and documentation of the findings are crucial for sharing insights and facilitating decision-making based on the analysis.

RESULTS AND DISCUSSION

The results of calculations using the k-means clustering algorithm were obtained in the form of sales segmentation at Batam IT Store Com which was clustered into 3 clusters. These clusters have been previously determined by researchers before calculations are carried out using the k-means clustering algorithm.

The clusters formed are cluster 0 which contains items that are not selling well, cluster 1 is containing items that are selling well, and cluster 2 is containing items that are selling very well. From the results of manual calculations, it was found that cluster 0 consists of 17 segments, cluster 1 consists of 4 segments and cluster 2 consists of 4 segments.

The first stage of the research was carried out by pre-processing the data by drawing research samples from which calculations would be carried out.

The stage after pre-processing the data is to clean the data that has been obtained at the pre-process data stage. Cleaning is done by filtering unused data, leaving data that will later be used in the calculation process.

After the data is cleaned, the data is then selected. The result of data selection is data that will be calculated using the k-means clustering algorithm.

Before entering the data mining stage, the data will go through a transformation stage. In this study, the data obtained was in the form of numbers so the data did not change. The results of the previous stages can be seen in the table 1.

Table 1. Research data

No	Name Of Goods	2019	2020	2021	2022
1	Light Bulb	12	4	6	28
2	5050 5v Usb Light	20	17	15	0
3	Light Clip/4 Pcs	12	20	6	1
4	3aaa Battery	84	14	11	2
5	3.5 M/ 6.5f	18	16	18	10
6	3.5 M/M Cable	144	23	38	49
7	Cat6e Cable	83	13	170	90
8	Dc 5.5 Cable	4	32	12	10
9	Hdmi F/F	27	25	13	5
10	Hdmi To Mini + Micro Hdmi	7	33	4	5
11	Iphone Cable	202	57	47	58
12	Led Cable Type C	14	7	0	9
13	Metal Cable Type-C	19	32	0	0
14	Round Cable 20cm	8	18	3	8
15	Round Cable 30cm	101	9	11	7
16	Usb 2.0 M/F	27	35	22	42
17	Universal Battery Charger	66	13	13	28
18	Usb 3.4a Charger Mj-A05	15	62	0	11
19	16gb Usb	30	0	8	2
20	32gb Usb	32	53	2	7
21	Card Readers	52	132	90	8
22	Headset	149	0	84	62

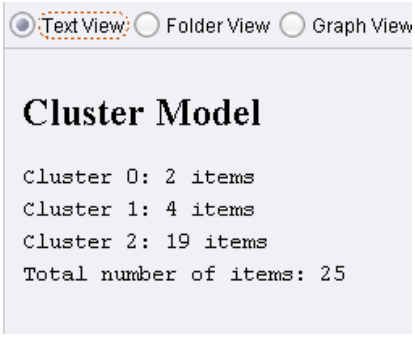
23	Otg Type C 16gb	6	10	12	3
24	Sandisk Ultra Microsdxc 64 Gb	20	9	20	23
25	Adapter On/Off	18	93	13	14

After the data transformation stage, the centroid point will be determined. The centroid point is determined by taking data with the lowest (C0), middle (C1), and highest (C2) sales levels, so that centroid points are obtained for data numbers 12 (C0), 18 (C1), and 11 (C2). By dividing these three clusters, we get cluster 0 with 19 segments, cluster 1 with 2 segments, and cluster 2 with 4 segments. Calculations are carried out using the Euclidean Distance formula and calculated repeatedly until the same results are obtained. In this research, calculations were carried out 6 times to get the same calculation results, the centroid point values which were the same can be seen in the table below.

Table 2. Point *Sixth iteration centroid*

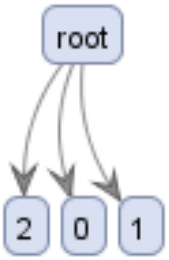
c0	27.47	21.53	9.26	10.58
c1	35	112.5	51.5	11
c2	144.5	23.25	84.75	64.75

To ensure whether or not the calculation results calculated manually are valid, auxiliary software in the form of RapidMiner is used to carry out the calculations. The results obtained from the RapidMiner software can be seen below.



Picture 2. *TextView*

There is a difference in the order of the clusters from manual calculations to the application because in manual calculations, the centroid points taken do not match the data sequence. In the application, cluster 0 is a cluster of goods that sell well, cluster 1 is a cluster of goods that sell very well, and cluster 2 is a cluster of goods that sell poorly.



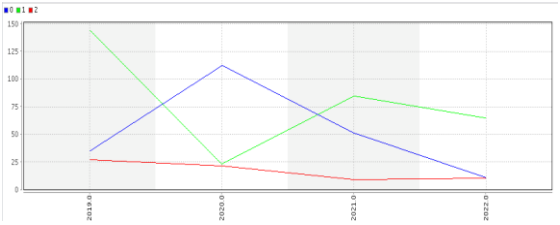
Picture 3 Graph View

In the graph view, the number of clusters calculated as a result of 3 clusters is displayed.

Attribute	cluster_0	cluster_1	cluster_2
2019.0	35	144.500	27.474
2020.0	112.500	23.250	21.526
2021.0	51.500	84.750	9.263
2022.0	11	64.750	10.579

Picture 4. *Centroid Table*

In the Centroid Table, the number displayed is the same centroid point from the calculation.



Picture 5. *Centroid Plot View*

The graph displayed in the Centroid Plot View is the location of the values displayed in the centroid table.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -1841.762
Avg. within centroid distance_cluster_0: -2160.500
Avg. within centroid distance_cluster_1: -5177.812
Avg. within centroid distance_cluster_2: -1105.884
Davies Bouldin: -0.742
```

Picture 6. *Performance Vector*

In the Performance Vector section, the average value between centroids is displayed. Apart from the average value, the optimum value from the calculation is also displayed, which is 0.742.

CONCLUSION

After the research is carried out, the author can draw conclusions in the form of:

1) Grouping and analyzing sales using data mining methods using the k-means clustering algorithm can be used to provide segmentation of goods sold at Batam IT Store Com. The results obtained from the test are clusters of goods that are not selling well, are selling well and are selling very well.
2) The clusters obtained from calculations from the data formed using the k-means clustering algorithm are in the form of cluster 0 which contains 19 items that were not selling well, cluster 1 which contains 2 items that were selling well, and cluster 2 contains items that were selling very well. as many as 4 data.

BIBLIOGRAPHY

[1] Alam, S., Official, MG, & Masripah, N. (2022). Classification of Covid-19 vaccine data screening with Naive Bayes algorithm using Knowledge Discovery in database method. Journal of Computer Networks, Architecture and High Performance Computing, 4(2), 177–185. https://doi.org/10.47709/cnahpc.v4i2.1584

[2] Dinata, RK, Safwandi, S., Hasdyna, N., & Azizah, N. (2020). K-Means Clustering Analysis on Motorcycle Data. INFORMAL: Informatics Journal, 5(1), 10. https://doi.org/10.19184/isj.v5i1.17071

[3] Kovács, L., & Ghous, H. (2020). Efficiency comparison of Python and RapidMiner. Multidiszciplináris Tudományok, 10(3), 212–220. https://doi.org/10.35925/j.multi.2020.3.26

[4] Mahalisa, G., & Arminarahmah, N. (2022). Diabetes Classification Analysis Using the Euclidean Distance Method Based on the K-Nearest Neighbors Algorithm. Journal of Computer Technology and Information Systems, 5(3), 178–182.

[5] Nugraha, AR, & Hasan, A. (2019). Control Electronic Devices Using Web-Based Applications Using Arduino. Jumantaka, 03(1), 11–21. Taken from http://jurnal.stmik-dci.ac.id/index.php/jumantaka/article/view/364

[6] Pradana, C., Kusumawardani, SS, & Permanasari, AE (2020). Comparison Clustering Performance Based on Moodle Log Mining. Pradana, C Kusumawardani, SS Permanasari, AE, 722(1), 1–11. https://doi.org/10.1088/1757-

899X/722/1/012012
[7] Robani, AM, Hadi, S., Nurdiawan, O., Dwilestari, G., & Suarna, N. (2021). Android-Based Used Motorcycle Sales Information System to Increase Sales in Mokascirebon. Com. JURIKOM (Journal of Computer Research), 8(6), 205–212. https://doi.org/10.30865/jurikom.v8i6.3629
[8] Shirazi, S., Baziyad, H., & Karimi, H. (2019). An Application-Based Review of Recent Advances of Data Mining in Healthcare. J Biostat Epidemiol., 5(4), 268–278.
[9] Sitinjak, DK, Pangestu, BA, & Sari, BN (2022). Clustering Health Workers Based on Districts in Karawang Regency Using the K-Means Algorithm. Journal of Applied Informatics and Computing, 6(1), 47–54. https://doi.org/10.30871/jaic.v6i1.3855