ISSN 1978-0524 (print) | 2528-6374 (online)
http://journal.uad.ac.id/index.php/JIFO/index

Predictive analytics on product sales at heva inc. using k – means method

Qurrota Nastiti Rizqita Aura Syifa[a,1,*], Murein Miksa Mardhia[a,2]

[a] Departement of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, 55191, Yogyakarta, Indonesia
[1] qurrota18000181240@webmail.uad.ac.id*; [2] murein.miksa@tif.uad.ac.id;
* Corresponding Author

Received 30 January 2022; accepted 10 March 2022; published 10 May 2022

ABSTRACT

Prediction is the process of estimating something that is most likely to happen in the future based on previous and current knowledge that is owned, with the goal of minimizing the error. Prediction allows people to recognize and then solve difficulties that are occurring or are expected to arise. This study began with preparation, literature review, data collection, and knowledge discovery in databases (KDD). One of the processes is data mining using the K – Means method, which is critical for obtaining the research's results and conclusions. This research also uses the RapidMiner application as a comparison of the results with the results obtained by python coding.By using 4 clusters, products were categorized into 4 labels, namely very good products, good products, bad products, and very bad products. The research resulted in 11 products in the bad product category, 12 products in the good product category, 10 products in the very good category, and 18 products in the very good product category. The very good product label was further clarified with visualization to show the best time to restock each recommended product.

KEYWORDS
Prediction
Data Mining
K-Means
Clustering
RapidMiner

1. Introduction

Prediction is hunches of future events based on present circumstances while forecasts involve a probability statement, so forecasting is part of the prediction [1][2]. Prediction does not have to provide a definite answer to events that will occur, but rather tries to find answers as close as possible to what will happen [3]. Predictions have similarities such as classification and estimation, but a prediction will look for a new value in the future by observing past data [4]. The use of prediction analysis allows an experiment's designer to estimate the accuracy that should be obtained from the experiment before completing the experimental setup [5].

The assumption underlying prediction analysis is that once an experiment has been performed and data has been collected, the data will be examined using a method such as K – Means Clustering. Data mining is the process of looking through huge volumes of computer data with specific software to uncover valuable information, or the study of enormous amounts of information stored in a computer to seek patterns, trends, and etc [6], [7]. Data mining and Knowledge Discovery in Databases (KDD) are terms that are frequently used interchangeably to describe the process of extracting information from a very large but linked database, as well as the KDD process's schema [8]. Clustering is one of the techniques in data mining which means widely utilized in a variety of applications, including business intelligence, visual pattern recognition, web search, biology, and security [9]. Clustering is the process of grouping data objects into distinct classes known as clusters [7]. The cluster process is carried out without following the hierarchical process once the number of clusters has been determined. K – Means Clustering is the name given to this procedure [9], [10]. The K – Means clustering algorithm divides data

10.26555/jifo.v116i2.a24754

into k groups using a partitioning clustering method [11]. – Means is a widely used clustering algorithm in data mining. The K – Means clustering algorithm divides data into many clusters based on object similarity, which is often determined by attribute values [12], [13]. To find out the number of k that is suitable for the calculation of k means clustering, the elbow method is calculated. The elbow method itself is a heuristic for counting the number of clusters in a data collection. Plotting the explained variation as a function of the number of clusters, the procedure entails choosing the elbow of the curve as the appropriate number of clusters [14], [15].

Some research on predictions has been carried out, for example in [16] that aims utilize available data on the database in company X to be processed automatically by the application that is specialized for forecasting sales transactions with the Time Series method, [17] that aims predict weak students and help the university management to make strategy and decision making related to student's performance improvement. [18] proposed to classify the income of an area based on the APBD using the K – Means method. [19] proposed to create an application that can classify selling products and unsold products and perform analysis using the K – means method and [20] research aims that the clinicians can use the prediction as a recommendation when making medical decisions in diabetics.

In this study, researchers conducted research on Heva Inc. which is a small-medium enterprise with several branches spread across the Special Region of Yogyakarta which sells fishes, aquariums, and other fishery-related items to be delivered across Java Island. In the midst of Covid-19 pandemic, Heva Inc. experienced a sales drop to 50% from previous year's sales thus causing new problems in the stock section. The problem is that the excess stock of goods causes the death of fish because it has not been sold for a long time. In order to solve this problem, an alternative way to be done is making predictions of new category of product sales and this prediction is made to make it easier for the stock provider part of Heva Inc. The grouping of data using the K - Means Clustering method was carried out using Google Collaboratory and the RapidMiner application as a comparison of results. RapidMiner itself is an open-source software for knowledge discovery and data mining to find data according to the purpose of processing the data, not all existing algorithms can match or process existing data sets [19],[21]. The research wil be continued by make a visualization of the new labels that will be produced.

2. Method

2.1. Data Acquisition

Descriptive analytical approach was used as the research method. Analytical descriptive research is a method that is used to describe or offer an overview of the object under investigation using data or samples that have been acquired in their natural state without analyzing and drawing generalizable conclusions. In other words, analytical descriptive research examines problems or focuses attention on problems as they are at the time of research, and the research findings are then processed and analyzed to form conclusions. The data collecting method was used to get the information needed for the investigation. The following are the data collection procedures used by the researcher in the research:

2.1.1 Interview

Interview method is a way of gathering data through interactions between researchers (assigned personnel) and research subjects, respondents, or data sources. For this research, the interviewees were the owner, the marketing and stock division team. This activity was carried out in order to learn more about product sales at Heva Inc. data from November 2019 to November 2021 were used for data mining calculations.

2.1.2 Observation

Observation as a technique or approach for obtaining primary data by directly witnessing the data object. The technique is taken by directly observing events in Heva Inc.'s product sales

2.1.3 Data Acquisition

Literature Study is a data collection approach that involves completing a review study of books, literature, notes, and reports related to the subject at hand. Data is gathered directly from other sources, such as books, theses, journals, and other papers relevant to this research.

2.2. Research Stage

Flowchart in Fig. 1 depicts the stages of research. Describes the research process that will be used as well as the research in general

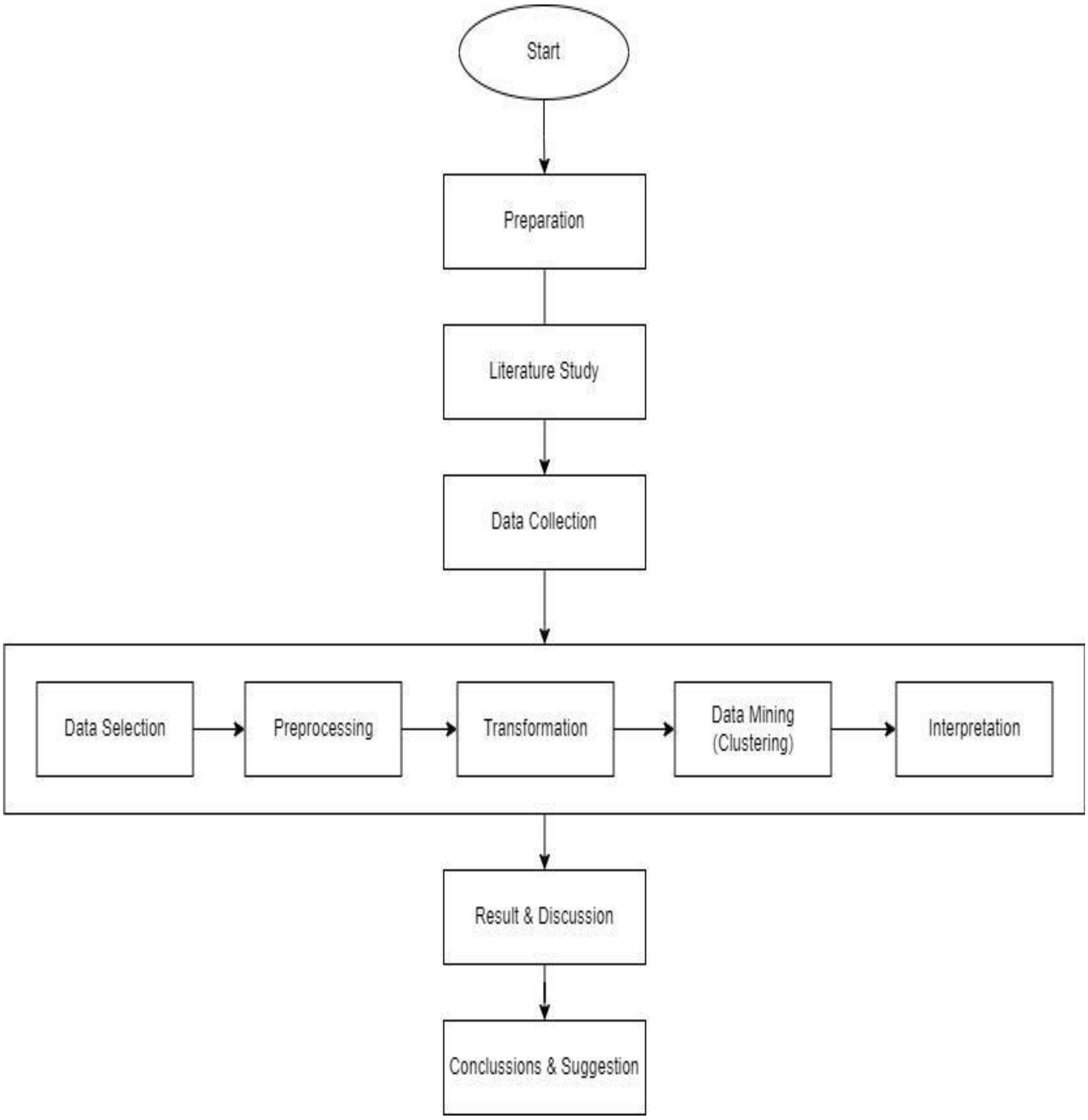


Fig 1. Flowchart of Research Stage

2.2.1 Preparation

The object taken is sales products in Heva Inc., limitations must be identified, and a research strategy must be devised.

2.2.2 Literature Study

The literature review includes literature reviews and studies on sales forecasting and related topics.

2.2.3 Data Collection

Data was gathered through interviews with the accounting, observation, and document divisions. Data collection from a set of operational data that must be completed before the stage of information extraction in KDD begins. The data collection in this study is the result of interviews with the owners of Heva Inc., the marketing division and the stock division as well as observing the data. The data is a weekly sales table from Heva Inc.

2.2.4 Data Selection

The term data selection aims at choosing data that should be stored during data collection or that should be shared/archived after the project is completed. The chosen data will be utilized for data mining and will be saved in a file separate from the operating database. The data used in this research are product sales in Heva Inc. during September 2019 until November 2021.

2.2.5 Preprocessing

Before the data mining process can be implemented, the preprocessing stage needs to be done, at this stage the data integration process will be done to combine data from different databases, then data cleaning will be done to produce a clean dataset so that it can be used in the next stage, namely mining. This stage is the initial stage of the KDD process. At this stage irrelevant data, missing values, and redundant data must be cleared. This is because relevant data, not missing value, and not redundant is the initial requirement in conducting data mining. A data is said to be missing value if there are attributes in the dataset that do not contain a value or empty, while the data is said to be redundant if in a dataset more than one record that contains the same value, after cleaning the data that is more qualified based on sales data.

2.2.6 Transformation

The Transformation stage is the stage of changing the data that has been selected, so that the data is suitable for the data mining process. The transformation process in KDD is a creative process and is highly dependent on the type or pattern of information to be searched in the database. At this stage, from all operational data, the attribute grouping data used for the data mining transformation process is obtained, namely the moon attribute and classification as the data criteria that are the targets in the mining process

2.2.7 Data Mining Process (Clustering)

This stage is a process of looking for interesting patterns or information in selected data using certain techniques or methods based on the overall KDD process. The method used in this research is the K – Means Clustering This stage includes checking whether the patterns or information found contradict the facts or pre-existing hypotheses. At this stage, the product sales pattern is obtained from the data mining process with the K – Means method, the pattern or information generated from the data mining process is in the form of rules obtained from the K – Means calculation. The ease with which the k-means clustering method works has led to its adoption in a variety of industries. The K – Means clustering algorithm divides data into k groups using a partitioning clustering method [10], [11]. The K – Means clustering algorithm has gained popularity due to its ability to cluster large amounts of data quickly and efficiently. In terms of clustering error, the K – Means algorithm 2nds locally optimal solutions. It's a quick iterative algorithm that's been used in a lot of clustering software. It's a point-based clustering algorithm that starts with the cluster centers at random positions and moves them at each stage to reduce clustering error. The method's fundamental flaw is that it is sensitive to the starting placements of cluster centers. As a result, numerous runs with different initial placements of the cluster centers must be scheduled in order to find near optimal solutions using the K – Means method [22]. K – Means is a widely used clustering algorithm in data mining. The K – Means clustering algorithm divides data into many clusters based on object similarity, which is often determined by attribute values. The goal of this approach is to reduce the objective function used in the clustering process, which aims to reduce variances within a cluster. Also, keep cluster variation to a minimum. K – Means clustering algorithm is divided into many steps, which are as follows:

- Enter the number of clusters.
- Numeric attributes are the only ones that can be handled.

There are two essential stages in the K – Means method: determining the central position of each cluster and searching for members of each cluster. The K – Means approach works as follows [5], [22]:

- Calculate k, which is the number of clusters that will be produced
- Set the initial centroid (cluster's center point) to k at random.
- Using the Euclidean Distance formula, calculate the distance between each data point and its centroid. Equation (1) is used to calculate Euclidean distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{k} \quad (xi - yi)^2} \tag{1}$$

Where d = distance of value of the data, k = number of data, xi = data to i from testing and yi = data to I from training.

- A centroid with the shortest distance will be assigned to each data set.
- To find the location of the new centroid position, calculate the mean value of the data in the same centroid
- If the new centroid's position differs from the previous centroid, repeat step 3.

2.2.8 Interpretation

Data interpretation (or can also be called evaluation) is examining data and arriving at relevant conclusions using various analytical methods. Data interpretation assists researchers in categorizing, manipulating, and summarizing information to answer critical questions. The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties.

2.2.9 Result and Discussion

The discussion at this point explains the outcomes of the data mining procedure, which was performed using the K – Means Clustering approach.

2.2.10 Conclussion and Recomendation

Draw conclusions from the research findings and make recommendations for the firm to improve

3. Results and Discussion

3.1. Data Collection

The data obtained are 2 files. The first file is a hardcopy containing the number of weekly sales per product and the second file is a softcopy in excel form containing weekly sales per product so that both files are still raw. Then proceed with entering the data into excel for easy use by researcher. Based on the results of interviews and requests from Heva Inc. weekly sales results are calculated into monthly sales and produce raw data that is used as a dataset in this study

3.2. Data Acquisition

The data used in this research is product sales data from Heva Inc. for the last three years, 2019, 2020, and 2021. Item Name and Month-Year are the attributes used to determine the group. To properly interpret the data, it is necessary to execute a data comprehension once the data selection is complete. The substantial decline that occurs each year in every product is clearly shown in table 1 below:

Table 1. Dataset for K – Means Clustering

Product name	September 2019	Oktober 2019	November 2019	Desember 2019	Januari 2020	Februari 2020	…	November 2021
Ikan Botia	105	88	100	118	58	69	…	20
Ikan Cupang	119	89	118	100	65	48	…	16
Ikan Koi	105	111	99	96	58	52	…	25
Ikan Black	109	110	87	109	53	48	…	31
Ikan Manfis h	88	85	118	93	49	49	…	37
…	…	…	…	…	…	…	…	…
Undergravel filter	93	93	109	91	70	54	…	19

3.3. Preprocessing

Preprocessing and data cleaning are fundamental operations that include removing noise data. Before beginning the data mining process, it is necessary to clean the data that will be the focus of KDD. The cleaning process includes checking for missing values, NaN, and outliers in the dataset.

3.3.1. Checking Missing Values

In the process of checking for missing values, listing below is the syntax of checking missing value:

```
df.isnull().sum()
```

After running the syntax, it results that the dataset used is completely filled and there are no missing values.

3.3.2. Checking Not a Number (NaN)

Checking for NaN or Not a Number which is a numeric data type value that represents an undefined or unrepresented value [23]. Listing below is the syntax of checking NaN:

```
df.isna().sum()
```

After running the syntax, it results that the dataset used is completely filled and there are no NaN.

3.3.3. Checking Outlier

The last process in the last preprocessing stage is checking outliers, which are observational data that appear with extreme values, either univariate or multivariate [23]. Fig. 2 below is the syntax of checking missing value:

```
df1=df.select_dtypes(include=['float64', 'int64'])
sns.boxplot(x="variable", y="value", color='green',orient='v',
data=pd.melt(df1))
plt.tight_layout()
```

Fig 2. Syntax Checking missing value

The result of the outlier check can be seen in Fig. 3 , in the dataset there are no outliers so it can be said that the dataset is clean.

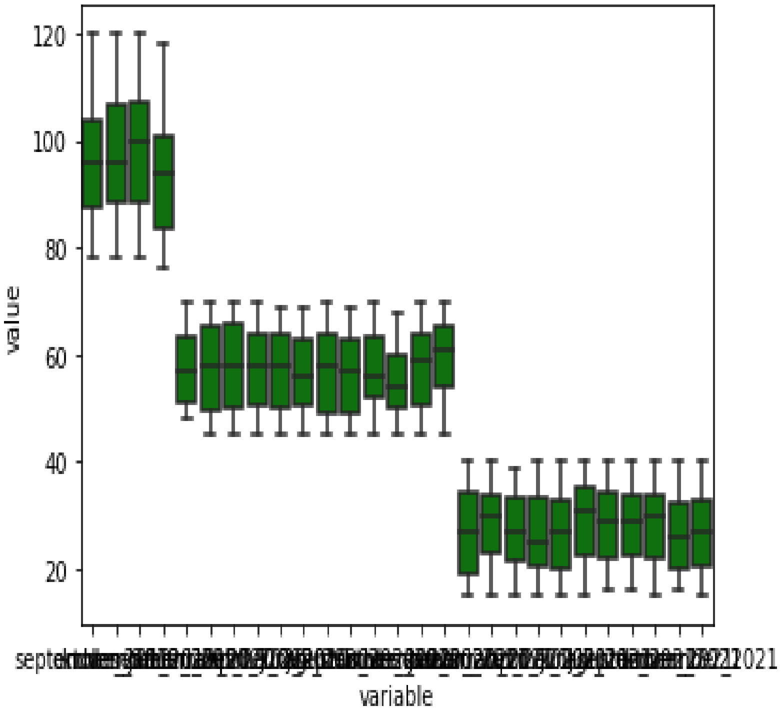


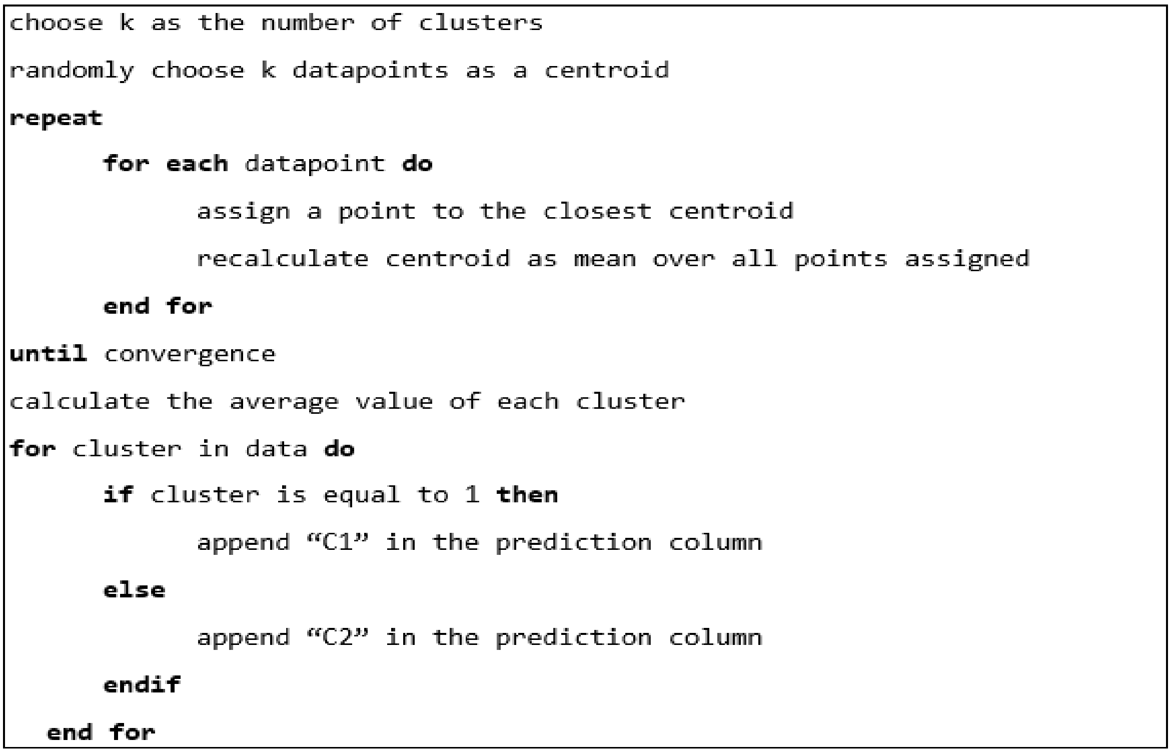
Fig 3. Result of Checking Outlier

3.4. Transformation

The transformation carried out is in the form of smoothing the columns in the data frame table by changing the Month-Year attribute to lowercase. The transformation is continued by changing the space (' ') in the month-year attribute to '_' to simplify the dataset processing. The transformation process aims to simplify the dataset processing process.

3.5. Data Mining (Clustering)

K – Means clustering process begins with determining K or the number of clusters and the centroid or cluster center and continued by calculating the distance of each object with each centroid using the distance measured on the similarity size of objects in the cluster. The distance measure used in the K – Means method is the Euclidean distance in Fig. 4



```
choose k as the number of clusters
randomly choose k datapoints as a centroid
repeat
    for each datapoint do
        assign a point to the closest centroid
        recalculate centroid as mean over all points assigned
    end for
until convergence
calculate the average value of each cluster
for cluster in data do
    if cluster is equal to 1 then
        append “C1” in the prediction column
    else
        append “C2” in the prediction column
    endif
  end for
```

Fig 4. Syntax distance measure used in the K – Means method is the Euclidean distance

Experiments have been carried out by comparing the number of clusters used, namely 3 clusters (k = 3) and 4 clusters (k = 4). From the comparison of these clusters, the results are as in table 2 and table 3 below:

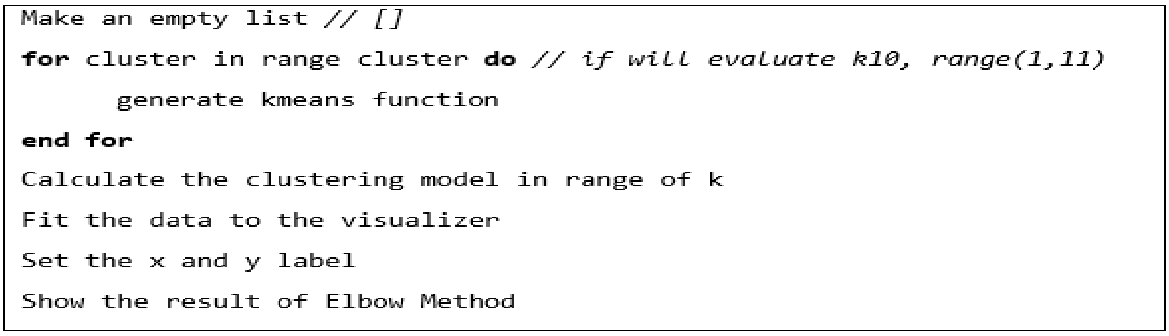
Table 2. Number of Products in Each Cluster for K = 3

Cluster Code	**Trial 1**	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**Trial 8**	**Trial 9**	**Trial 10**	**Trial 11**	**Trial 12**	**Trial 13**	**Trial 14**	**Trial 15**
0	16	27	22	22	16	24	24	15	16	16	18	16	18	16	16
1	11	24	5	5	11	27	18	19	11	12	11	12	17	11	22
2	24	0	24	24	24	0	9	17	24	23	22	23	16	24	13

Table 3. Number of Products in Each Cluster for K = 4

Cluster	**Trial 1**	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**Trial 8**	**Trial 9**	**Trial 10**	**Trial 11**	**Trial 12**	**Trial 13**	**Trial 14**	**Trial 15**
0	16	11	17	10	11	19	11	20	11	11	14	11	13	16	13
1	8	12	9	14	12	6	15	15	12	15	10	12	18	15	18
2	19	10	7	15	10	13	9	10	10	9	11	10	2	20	2
3	8	18	18	12	18	13	16	16	18	16	16	18	18	18	18

Based on the results of k = 3 and k = 4 in the table above, it can be seen that the pattern of data changes at k = 4 is more stable than k = 3, especially in cluster 3 which has relatively the same and stable results.The last step in implementing Google Collaboratory is evaluating the best k value using the Elbow method. The elbow method itself is a method used to generate information in determining the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point [23]. The algorithm of Elbow Method in pseudocode in Fig. 5



```
Make an empty list // []
for cluster in range cluster do // if will evaluate k10, range(1,11)
      generate kmeans function
end for
Calculate the clustering model in range of k
Fit the data to the visualizer
Set the x and y label
Show the result of Elbow Method
```

Fig 5. Syntax algorithm of Elbow Method in pseudocode

Figure 6 explains that produces the best k when the Sum of Square Distance or the sum of the squares of the distances of each data to the centroid point is small, in this case k = 10.

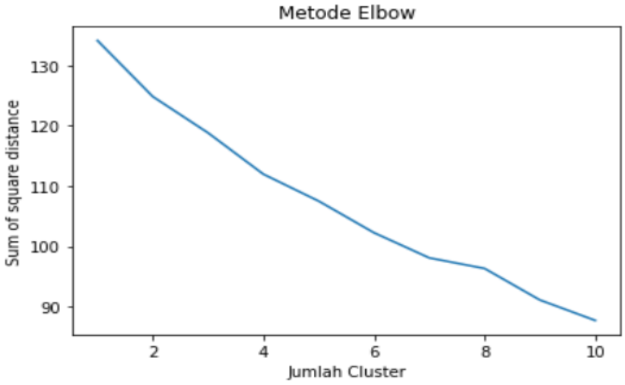


Figure 6. Result of Elbow Method shows that the k used in this study is not the best k because it does not indicate the angle

3.5.1. RapidMiner Implementation

Analysis was continued with the implementation using the RapidMiner application and experimented 5 times. RapidMiner is used as a comparison of results with Google Collaboratory and as an analytical amplifier. Analysis using RapidMiner begins with entering the dataset into the application and continues with making designs as shown in Fig. 7 .The operator used is Multiply which is placed between the dataset and the K – Means Clustering operator.

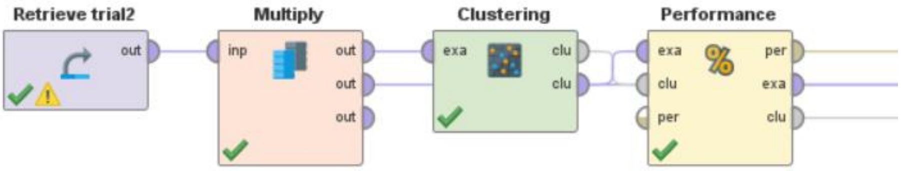
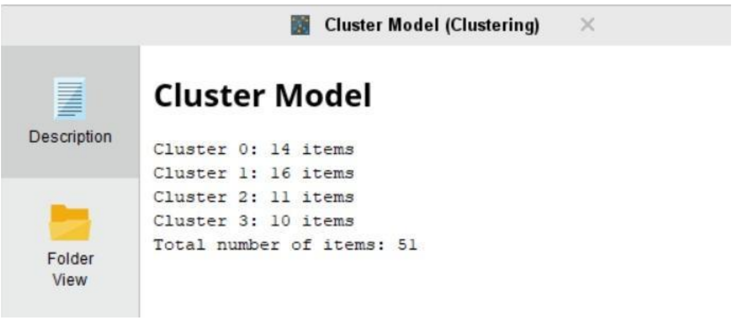


Fig 7. Design for K – Means Clustering on Rapid Miner

After making the design and running the clustering function, the clustering results will appear as shown in Fig. 8 . It can be seen in figure 6 that in cluster 0 there are 14items, cluster 1 are 16 items, cluster 2 are 11items, cluster 3 are 10 items.



Cluster Model

```
Cluster 0: 14 items
Cluster 1: 16 items
Cluster 2: 11 items
Cluster 3: 10 items
Total number of items: 51
```

Fig 8. Result of Cluster Model

Fig. 9 below explains showing which products belong to cluster 0, cluster 1, cluster 2 and cluster 3 in each experiment.

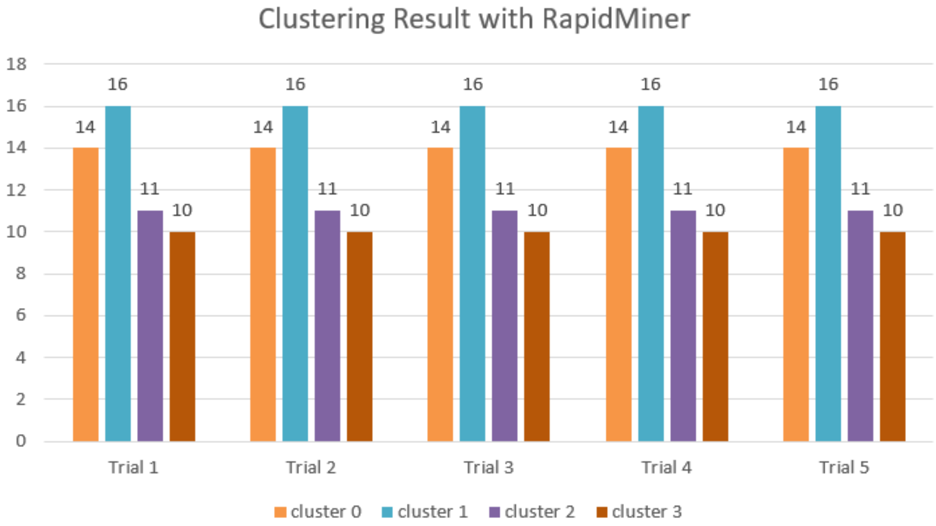


Fig 9. Result of Clustering with RapidMiner

3.6. Interpretation

Based on the results of the comparison between k = 3 and k = 4, then in this study used 4 clusters or k = 4. The labels that represent each cluster formed are cluster 0 is the Bad Product, cluster 1 is the God Product, cluster 2 is the Very Bad Product and cluster 3 is Very Good Product. The label is obtained from the average calculation on K – Means Clustering, if the average cluster has the best value, it means that the cluster belongs to cluster 3 or is labeled Very Good Product and so on.

From 15 experiments using Google Collaboratory, obtained 4 experiments with same results, namely the 2nd, 5th, 8th, and 11th experiments (for the results of each experiment can be seen in the attachment). Meanwhile, from 5 experiments using the RapidMiner application, the results were the same. So that the midpoint can be taken, that the results of the analysis using Google Collaboratory and RapidMiner researcher took data with dominant results, namely products that entered the Bad Product category are 11 products, Good Products are 12 products, Very Bad Products are 10 products and Very Good Products are 18 products.

To examine how each product's sales changed over the course of a year, there was formerly a product visualization in the Very Good Product category (September 2019 - 2020). The ideal month to buy shares can be suggested using the information in this visualization. The visualization in Fig 10 below shows that sales decreased in October 2019, but then started to increase in October and November. This was followed by a steep reduction in sales and a slight gain in July 2020. Sales of the products were very consistent after July. This demonstrates that the best months for product replenishment are November and July.

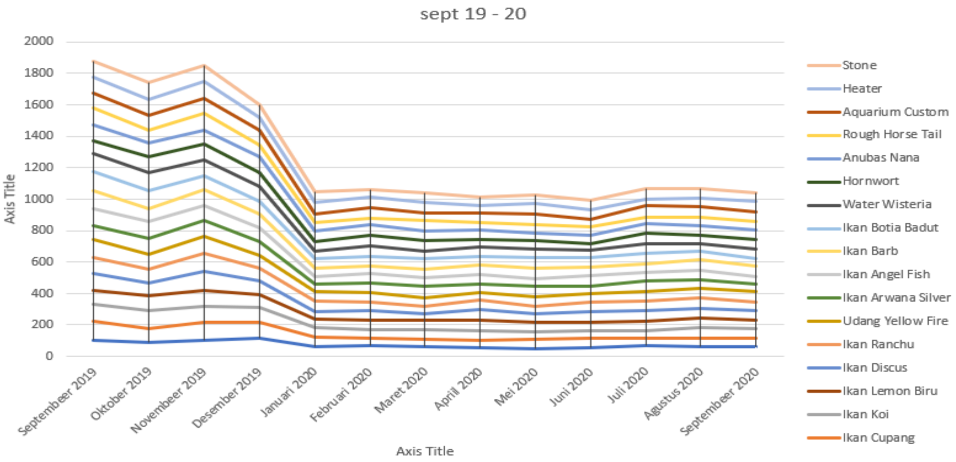


Fig 10. Visualization in Very Good Products Category

3.7. Result

After analyzing the design with data mining stages to produce product category predictions using the K – Means algorithm, the research team carried out the actual data mining process to find out category predictions based on product sales data in November 2019, 2020, November 2021 at Heva Inc.. The results of the analysis namely cluster 0 is the Bad Product, cluster 1 is the God Product, cluster 2 is the Very Bad Product and cluster 3 is Very Good Product. Table 4 is the name of the product that has been included in each cluster according to the results of the clustering that has been done.

Table 4. Results Details with RapidMiner

Product Name	Cluster	Label
Ikan Botia	3	Very Good Product
Ikan Cupang	3	Very Good Product
Ikan Koi	3	Very Good Product
Ikan Black Ghost	2	Very Bad Product
Ikan Manfish	0	Bad Product
Ikan Swordtail	1	Good Product
Ikan Lemon Kuning	2	Very Bad Product
Ikan Lemon Biru	3	Very Good Product
Ikan Guppy	2	Very Bad Product
Ikan Neon Tetra	2	Very Bad Product
Ikan Molly	1	Good Product

Product Name	Cluster	Label
Ikan Platy	0	Bad Product
Ikan Discus	3	Very Good Product
Ikan Ranchu	3	Very Good Product
Ikan Maskoki	2	Very Bad Product
Ikan Oranda	0	Bad Product
Ikan Komet	0	Bad Product
Ikan Zebra	0	Bad Product
Ikan Louhan	1	Good Product
Ikan Sapu sapu	2	Very Bad Product
Udang Blue Pearl	0	Bad Product
Udang Yellow Fire	3	Very Good Product
Udang Bamboo	0	Bad Product
Ikan Arwana Silver	3	Very Good Product
Ikan Arwana Super Red	1	Good Product
Ikan Oscar	2	Very Bad Product
Ikan Angel Fish	3	Very Good Product
Ikan Barb	3	Very Good Product
Ikan Botia Badut	3	Very Good Product
Ikan Pedang	0	Bad Product
Ikan Peacock Bass	2	Very Bad Product
Ikan Corydoras	1	Good Product
Water Wisteria	3	Very Good Product
Java Moss	1	Good Product
Java Fern	0	Bad Product
Hornwort	3	Very Good Product
Anubas Nana	3	Very Good Product
Amazon Frogbit	1	Good Product

Product Name	Cluster	Label
Amazon Sword	0	Bad Product
Rough Horse Tail	3	Very Good Product
Aquarium Kecil	2	Very Bad Product
Aquarium Sedang	1	Good Product
Aquarium Besar	1	Good Product
Aquarium Custom	3	Very Good Product
Filter Aquarium	1	Good Product
Lampu Aquarium	1	Good Product
Aerator	2	Very Bad Product
Power Head	1	Good Product
Heater	3	Very Good Product
Stone	3	Very Good Product

4. Conclusion

Knowledge Discovery in Database (KDD) stage. Based on the results of data mining calculations using the clustering technique with the K – Means algorithm and analysis using the RapidMiner application, the results of the new group are as many as 11 products fall into the bad product category, 12 products fall into good product category, 10 products fall into the very good category and 18 products fall into the very good product category. Based on the results of clustering, the very good product label is taken as a support for recommendations in the form of visualization of the best month to restock products. The visualization that has been done shows that the middle of the year can be the best time to restock product

Acknowledgements

The author would like to thank family, friends, colleagues, work partners for their support and prayers during this time. The author is grateful to all lecturers at UAD who have provided a lot of valuable knowledge, especially Mrs. Murein Miksa Mardhia, S.T., M.T. as a supervisor. And all parties who can't be mentioned one by one. Last but not the least, author thanks herself for her hard work, dedication, and for never quitting.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper

References

[1] G. D. Rudebusch and J. C. Williams, “Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve,” *http://dx.doi.org/10.1198/jbes.2009.07213*, vol. 27, no. 4, pp. 492–503, 2012, doi: 10.1198/JBES.2009.07213.

[2] V. Kumar and M. L., “Predictive Analytics: A Review of Trends and Techniques,” *Int. J. Comput. Appl.*, vol. 182, no. 1, pp. 31–37, 2018, doi: 10.5120/ijca2018917434.

[3] “Prediksi Kerusakan Motor Induksi Menggunakan Metode Jaringan Saraf Tiruan Backpropagation.” https://repositori.usu.ac.id/handle/123456789/38400 (accessed Dec. 19, 2020).

[4] G. Atluri, A. Karpatne, and V. Kumar, “Spatio-temporal data mining: A survey of problems and methods,” *ACM Comput. Surv.*, vol. 51, no. 4, 2018, doi: 10.1145/3161602.

[5] E. J. Wolberg, “Prediction Analysis,” *Des. Quant. Exp.*, pp. 90–127, 2010, doi: 10.1007/978-3-642-11589-9_4.

[6] P. Giudici and S. Figini, “Applied Data Mining for Business and Industry,” *Appl. Data Min. Bus. Ind.*, pp. 1–249, Apr. 2009, doi: 10.1002/9780470745830.

[7] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*, vol. 9780470908. 2014.

[8] C. Zhang and J. Han, “Data Mining and Knowledge Discovery,” *Urban B. Ser.*, pp. 797–814, 2021, doi: 10.1007/978-981-15-8983-6_42/FIGURES/6.

[9] A. V. Novikov, “PyClustering: Data Mining Library,” *J. Open Source Softw.*, vol. 4, no. 36, p. 1230, Apr. 2019, doi: 10.21105/JOSS.01230.

[10] A. Nur Khormarudin, “Teknik Data Mining: Algoritma K-Means Clustering,” *J. Ilmu Komput.*, pp. 1–12, 2016, [Online]. Available: https://ilmukomputer.org/category/datamining/.

[11] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, Feb. 2003, doi: 10.1016/S0031-3203(02)00060-2.

[12] N. Shi, X. Liu, and Y. Guan, “Research on k-means clustering algorithm: An improved k-means clustering algorithm,” *3rd Int. Symp. Intell. Inf. Technol. Secur. Informatics, IITSI 2010*, pp. 63–67, 2010, doi: 10.1109/IITSI.2010.74.

[13] Y. Liu, H. P. Yin, and Y. Chai, “An improved kernel k-means clustering algorithm,” *Lect. Notes Electr. Eng.*, vol. 404, pp. 275–280, 2016, doi: 10.1007/978-981-10-2338-5_27.

[14] R. L. Thorndike, “Who belongs in the family?,” *Psychom. 1953 184*, vol. 18, no. 4, pp. 267–276, Dec. 1953, doi: 10.1007/BF02289263.

[15] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.

[16] A. Triayudi, Sumiati, T. Nurhadiyan H, and V. Rosalina, “Data mining implementation to predict sales using time series method,” *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 7, no. October, pp. 1–6, 2020, doi: 10.11591/eecsi.v7.2028.

[17] P. Bachhal, S. Ahuja, S. Gargrish -, and A. Uswatun Khasanah, “A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 215, no. 1, p. 012036, Jun. 2017, doi: 10.1088/1757-899X/215/1/012036.

[18] R. Sari, V. T.-J. I. Terapan, and undefined 2020, “Prediksi Jumlah APBD Kota Payakumbuh dengan metode K-Means,” *ejournal.lldikti10.id*, Accessed: Dec. 19, 2020. [Online]. Available: http://ejournal.lldikti10.id/index.php/jit/article/view/5323.

[19] Y. Darmi, A. Setiawan, J. Bali, K. Kampung Bali, K. Teluk Segara, and K. Bengkulu, “PENERAPAN METODE CLUSTERING K-MEANS DALAM PENGELOMPOKAN PENJUALAN PRODUK,” *J. MEDIA INFOTAMA*, vol. 12, no. 2, Dec. 2016, doi: 10.37676/JMI.V12I2.418.

[20] A. Prasatya, R. R. A. Siregar, and R. Arianto, “Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes,” *PETIR*, vol. 13, no. 1, pp. 86–100, Mar. 2020, doi: 10.33322/PETIR.V13I1.925.

[21] M. F. Akhtar, *RapidMiner Use Cases and Business Analytics Applications - Chapter 5 Naïve Bayes Classification I*, 1st ed. 2014.

[22] “[PDF] Enhancing K-means Clustering Algorithm with Improved Initial Center | Semantic Scholar.” https://www.semanticscholar.org/paper/Enhancing-K-means-Clustering-Algorithm-with-Initial-Yedla-Rao/5b4d09f41f8fa28c8f4ac3e7b8a474ae9f84b197 (accessed Dec. 19, 2020).

[23] F. Corea, “An Introduction to Data,” vol. 50, 2019, doi: 10.1007/978-3-030-04468-8.