

Data Mining on Women's Clothing Sales in Market Places with the K-means Clustering Algorithm

¹Rizna Fitriana Dalimunthe, ²Raissa Amanda Putri

¹Sains dan Teknologi, Ilmu Komputer, Universitas Islam Negeri Sumatera Utara

²Sains dan Teknologi, Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Jl. Lap. Golf No. 120, Medan, Indonesia

Email: ¹riznafitriana5@gmail.com, ²raissa.ap@uinsu.ac.id

Article Info

Article history:

Received May 12th, 2019

Revised Jun 20th, 2020

Accepted Jul 26th, 2020

Keyword:

Data Mining

Women's Clothing

Marketplace

K-means Clustering

ABSTRACT

Clothing is a necessity that must be used to cover the body with the main material made of fiber or textile so that the body is completely covered without gaps. Marketplace is an application or website that provides online buying and selling facilities from various sources. On the Shopee marketplace, there are many shops selling women's clothing from various groups and types of clothing. The K-means Clustering algorithm in the research was applied to make it easier for sellers and buyers to find out what kind of women's clothing is currently selling well in the marketplace by grouping it into 3 clusters, namely the best-selling, best-selling, and least-selling. Research data was obtained from the Shopee marketplace with 3 variables, namely product price, number of sales, and buyer assessments of 4 types of women's clothing in the form of tunics, dresses, abayas/gamis, and shirts totaling 1200 data. The results of this research make it easier for buyers to make decisions and sellers to develop shop ideas.

Copyright © 2022 Puzzle Research Data Technology

Corresponding Author: (10 pt)

Third Author,

Departement of Electrical and Computer Engineering,

National Chung Cheng University,

168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan, ROC.

Email: thirdauthor@uin-suska.ac.id

DOI: <http://dx.doi.org/10.24014/ijaidm.v2i2.xxxx> (10 pt)

1. INTRODUCTION

In the current era of technology, people use technology to simplify various tasks, one of which is utilizing the internet for conducting transactions without leaving their homes. A marketplace is an application or website that provides online buying and selling facilities from various sources [1]. It is an electronic product marketing platform that brings together many sellers and buyers to transact with each other [2]. Selling Muslim fashion online allows sellers to market, promote, and transact with customers without meeting face-to-face or being limited by time constraints. It can be done without having a physical shop [3]. Consumers prefer online shopping because it is practical, modern, and can be done at any time without the hassle of leaving the house to get the desired goods [4].

One marketplace that sells clothes is Shopee. The advantage of this marketplace is its easy payment access, which appeals to consumers, especially housewives [5]. Additionally, through its programs, campaigns, and promotions, Shopee has succeeded in ranking among the top five e-commerce sites with the highest number of visitors [6]. As a result, people have started to take advantage of this opportunity by setting up shops selling various types of women's clothing [7]. However, this marketplace lacks an algorithm to determine the best-selling clothes, particularly for women's clothing. By implementing such an algorithm, buyers can receive recommendations from trusted stores and stay updated with the best-selling women's clothing. This feature benefits not only buyers but also sellers who want to open a shop on the platform. Sellers can understand consumer preferences, identify effective marketing strategies, make informed

decisions, and keep up with changing trends for their shop's benefit. Some sellers waste money by continuously buying products that suit trends but do not adapt themselves [8].

The use of data mining supports every businessman in making quick and accurate decisions. Data mining is a process that employs artificial intelligence, statistical techniques, mathematics, and machine learning to identify and extract useful information and related knowledge from large databases [9]. It involves exploring added value in the form of information produced by extracting and recognizing patterns contained in databases [10]. According to the Gartner Group in Larose, data mining is a process of finding meaningful relationships, patterns, and trends by examining large sets of data stored in storage using pattern recognition techniques such as statistical and mathematical techniques [11]. Analyzing consumer purchasing trends to determine sales patterns, and processing them correctly can help identify which products are the best sellers, best sellers, or slowest sellers. This aids in inventory management and provides valuable input for businesses when formulating their marketing strategies [12]. Data mining is suitable for addressing this problem, as it is useful for extracting data about sales and best-selling women's clothing on the Shopee marketplace. This data mining can be assisted by programming languages to determine the results.

Data mining encompasses various techniques, including classical methods such as statistics, nearest neighbors, clustering, and decision trees [13]. For this problem, the K-Means Clustering algorithm is particularly suitable. K-Means clustering is a data mining technique that provides a cluster description of a product. It can also be interpreted as a data segmentation method applied in several fields, including marketing, business problem analysis, market segmentation and predictions, computer vision patterns, regional zoning, object identification, and image processing [14]. The K-Means clustering method aims to group existing data into several clusters, where data within a cluster share similar characteristics and differ from data in other clusters [15]. The goal of this grouping is to minimize the objective function set in the clustering process [16]. The K-Means algorithm is a distance-based, non-hierarchical clustering method that divides data into clusters and works on numeric attributes [17]. Similar objects are placed in adjacent clusters, while dissimilar objects are placed in separate clusters [18]. This K-Means Clustering method can be used to group best-selling women's clothing, analyzed using data obtained from the Shopee marketplace.

2. RESEARCH METHOD

The method used in this research is quantitative, where this research begins with the data collected consisting of variables that measure numbers or numbers to the processed data and its output. In this research, data mining analysis was carried out using a programming language for business strategy and decision-making at the Shopee Market Place in determining the best sellers from women's clothing sales data. The following is the research framework:

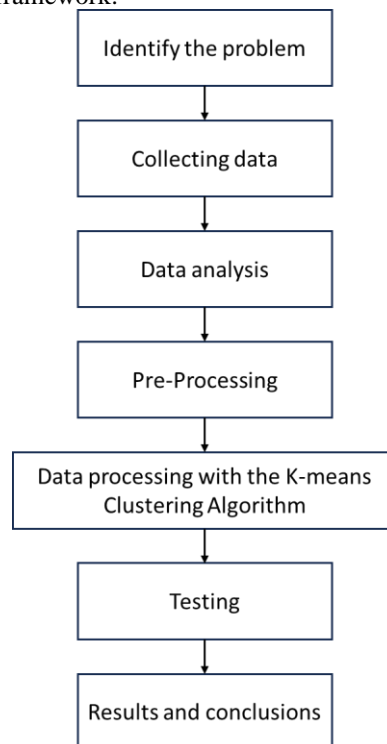


Figure 1. Research Framework

To help buyers and sellers in marketplace shops with data mining using the K-means clustering algorithm, the author wants to create data visualizations so that buyers or sellers can more easily determine the best-selling women's clothing to help make decisions and develop shop ideas. This research collected 1200 shopee women's clothing sales data with 4 categories, namely tunics, dresses, abayas/gamis, and shirts. The design in the form of a flowchart can be seen in the following image:

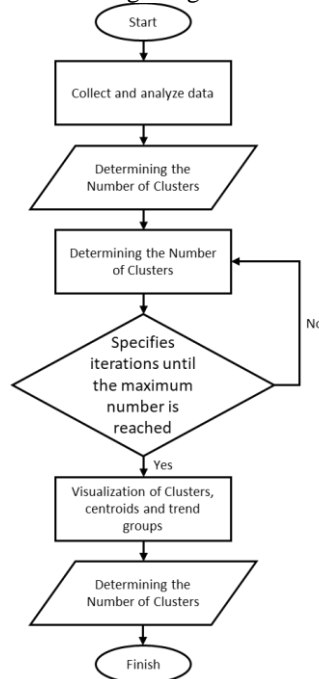


Figure 2. Flowchart of the stages of the K-Means Clustering algorithm

In the flow chart above, it is shown that the stages start with data collection. This data collection was carried out using a web scraper, obtaining data on 1200 sales data for women's clothing on the Shopee marketplace. The collected data is exported in Excel (.xlsx) or CSV (.csv) format. The following is an example of the data collected.

Table 1. Example of collected data

No	Product name	Price	Sale	Rating
1	ATASAN WANITA TUNIK OVAL BAHAN CRINGKLE AIRFLOW PREMIUM//TUNIK OVERSIZE	Rp 36.500	4.800	04.08
2	MIDI DRESS MOTIF BUNGA SOKA BALI	Rp 36.000	10.000	04.08
3	gamis simer silky mdl lengan gantung	Rp 72.000	3.700	04.05
1200	Kemeja Wanita Premium Kerut Depan Oversize UNGGUL FASHION	Rp 44.999	8.800	04.04

After the data is collected, the number of clusters is determined. In this problem, researchers grouped 3 clusters, namely less popular, best selling, and best selling with the attributes of price, sales, and valuation. Then initialize the centroids randomly. In this section the researcher takes some data randomly to determine the initial centroid using the Euclidean Distance formula as follows: [19]

$$d(x_i, \mu_j) = \sqrt{\sum (x_i, \mu_j)^2}$$

Where: x_i = criteria data
 μ_j = centroid of cluster j_s

After determining the initial centroid, the researcher then proceeds with iterations to determine which cluster the data belongs to. For example, the groupings of least-selling, best-selling, and most popular items are marked with C1, C2, and C3, respectively. After the iterations reach the maximum number, the researcher

creates a visualization of the data using the Python programming language with Google Colab. In this research, the necessary libraries include Keras, TensorFlow, NumPy, Pandas, and other supporting libraries, such as Matplotlib for creating graphs. With the guarantee of stable server capabilities, almost all processing runs smoothly on Google Colab as long as the internet connection is stable [20]. The data visualization, in the form of graphs, facilitates comparisons to help make decisions regarding the problem. Additionally, the interpretation of grouping results and conclusions from the clustering modeling will be discussed.

3. RESULTS AND ANALYSIS

3.1. Data Analysis

This data analysis stage includes the results of data collection and exploration. Data collection was carried out using a web scraper by creating keywords representing four types of women's clothing: Tunics, Dresses, Abayas/Gamis, and Shirts in the product search feature on Shopee. For each type of clothing, 300 pieces of data were sequentially collected and then combined into an MS Excel file. For example, data entries 1-300 represent sales of Tunic-type clothing, while entries 301-600 represent sales of Dress-type clothing, continuing this pattern until the total data reached 1200 entries, encompassing women's clothing sales from various stores.

Data exploration was conducted to analyze the collected data, aiming to identify and remove any duplicate or empty entries. The data used in this research was obtained from the Shopee marketplace, consisting of sales data for women's clothing. The dataset included 1200 entries, with a sample of 20 entries manually examined by researchers for the analysis.

3.2. Data Representation

Application of the K-Means Clustering algorithm to group women's clothing sales data on the Shopee Market Place with the data represented in the form of relevant attributes. These attributes include product name, product price, number of sales, and buyer ratings. The following is a data representation of Market Place Shopee women's clothing sales data.

Table 2. Preliminary Data on Women's Clothing Sales

No	Product Name	Product Price	Number of Sales	Buyer Ratings
1	GFS SHAKILA DAILY TUNIK CICIRA BAHAN SHAKILA PREMIUM	Rp 38.000	10000	04.06
2	Atasan Batik Wanita Modern Tunik Reslestant Depan Terbaru Tunik Batik Najwa Batik Irma	Rp 92.490	2200	04.08
3	ATASAN WANITA TUNIK OVAL BAHAN CRINKLE AIRFLOW PREMIUM//TUNIK OVERSIZE	Rp 36.500	4800	04.08
4	Kemeja Long Tunik Kattun Linen // Kemeja Oversize // Kemeja Linen Panjang	Rp 205.000	641	04.09
5	GG-Atasan Tunik Polos Crinkle Premium Wanita Tunik wanita(Dapat akrilik gold)	Rp 37.900	2100	04.06
6	Tunik Toyobo Wanita LD 100cm & LD 110cm Long Tunik	Rp 62.000	10000	04.09
7	[NEW ARRIVAL] POLO LINEN PREMIUM BAHAN LINEN POLO TERBARU KAIN KATUN BEST SELLER TERBARU DAN TERMURAH UNTUK KEMEJA TUNIK GAMIS OUTER KULOT ROK PER 0.5M BY MACMOHAN KRINKLE CRINKLE	Rp 11.500	10000	04.08
8	HAVA Outer Tunik / Outer Tile dot by DEZHAF	Rp 129.900	10000	04.09
9	Vallina Outfit - Marbel Long Tunic Baju Atasan Wanita Muslim Modern Linen Rami Premium M - XL	Rp 42.900	5100	04.06
10	TUNIK BRIELLE /ATASAN WANITA/BAHAN SHAKILA JUMBO LD 100,LD 110,LD 120,LD126Cm	Rp 62.900	10000	04.07
11	Kemeja Kiana Stripe Tunik Wanita Terbaru	Rp 31.900	9000	04.06
12	OVERSIZE MATT KNIT HORNETS ATASAN	Rp 29.250	6000	04.07

TUNIK LENGAN 7/8 / KAOS POLOS OVESIZE JUMBO				
13	AFI - EC - Kemeja Pocket salur Kemeja Tunik Wanita Sierra Salur Tunik	Rp 24.900	10000	04.06
14	Baju Wanita Asmita Tunik Jumbo Crinkle	Rp 57.000	10000	04.06
15	ALINA tunik by ZALFA OUTFIT / kemeja panjang / tunik polos / tunik rayon	Rp 89.000	10000	04.09
16	RX FASHION - KEMEJA AILEN KATUN TWISTCONE XXL /TUNIK AYUMI / Hanley long shirt / Hanley Tunik Atasan Muslim Fashion Wanita Termurah & Terlaris / azkiya tunik premium - NN	Rp 24.999	10000	04.07
17	DWYNE TUNIK Atasan Kemeja Wanita Linen Pakaian Muslim Wanita Fashion Muslim Wanita Terbaru	Rp 61.999	2300	04.08
18	KIMIA LONG TUNIK HORNET PREMIUM / TANIA TUNIK PANJANG BELAH SAMPING LENGAN KARET (COD)	Rp 49.800	2300	04.08
19	KALINA tunik by ZALFA OUTFIT / tunik rayon twill polos	Rp 88.000	10000	04.08
20	TUNIK POLOS JUMBO CRINKLE AIRFLOW PREMIUM	Rp 44.999	10000	04.07

3.3. Analysis of the K-Means Clustering Algorithm

When using the K-Means algorithm, the initial process that must be carried out by the Clustering method is the formation of clusters by transforming data into numerical form in the form of integers based on the initial data that has been collected. This aims to make it easier for researchers to get more accurate results. After that, the researcher can determine the number of clusters, calculate the centroid, calculate the centroid distance of the moving object or group, and then the calculation to find the iteration value is complete. In this method, the first step in determining a group or cluster of an object is to measure the Euclidean distance between two object points from the transformed attribute. The following are the provisions made for data transformation.

PN = Product Name

PP = Product Price

TS = Total Sales

BR = Buyers Rating

Example:

The product price is IDR 38,000 = 38000

Buyer Valuation is 04.06 = 46

Table 3. Women's Clothing Sales Transformation Data

No	PN	PP	TS	BR
1	P1	38000	10000	46
2	P2	92490	2200	48
3	P3	36500	4800	48
4	P4	78500	1400	49
5	P5	37900	2100	46
6	P6	62000	10000	49
7	P7	11500	10000	48
8	P8	129900	10000	49
9	P9	42900	5100	46
10	P10	62900	10000	47
11	P11	31900	9000	46
12	P12	29250	6000	47
13	P13	24900	10000	46
14	P14	57000	10000	46
15	P15	89000	10000	49
16	P16	24999	10000	47
17	P17	61999	2300	48
18	P18	49800	2300	48
19	P19	88000	10000	48
20	P20	44999	10000	47

After the data has been successfully transformed, clusters are formed into three groups (K=3) and the centroid center point is determined. The following is the clustering calculation process below.

K= 3 centroid

C1 = (38000,10000,46) taken from data 1

C2= (92490, 2200,48) taken from data 2

C3= (36500, 4800,48) taken from data 3

Next, the researcher carries out a calculation process to find iterations as below.

Iteration 1:

1) A(38000,10000,46)

$$C1 = (38000, 10000, 46) = \sqrt{(38000 - 38000)^2 + (10000 - 10000)^2 + (46 - 46)^2} = 0$$

$$C2 = (92490, 2200, 48) = \sqrt{(38000 - 92490)^2 + (10000 - 2200)^2 + (46 - 48)^2} = 55045,44$$

$$C3 = (36500, 4800, 48) = \sqrt{(38000 - 36500)^2 + (10000 - 4800)^2 + (46 - 48)^2} = 5412,02$$

2) B(92490, 2200, 48)

$$C1 = (38000, 10000, 46) = \sqrt{(92490 - 38000)^2 + (2200 - 10000)^2 + (48 - 46)^2} = 55045,44$$

$$C2 = (92490, 2200, 48) = \sqrt{(92490 - 92490)^2 + (2200 - 2200)^2 + (48 - 48)^2} = 0$$

$$C3 = (36500, 4800, 48) = \sqrt{(92490 - 36500)^2 + (2200 - 4800)^2 + (48 - 48)^2} = 56050,34$$

3) C(36500, 4800, 48)

$$C1 = (38000, 10000, 46) = \sqrt{(36500 - 38000)^2 + (4800 - 10000)^2 + (48 - 46)^2} = 5412,02$$

$$C2 = (92490, 2200, 48) = \sqrt{(36500 - 92490)^2 + (4800 - 2200)^2 + (48 - 48)^2} = 5606,33$$

$$C3 = (36500, 4800, 48) = \sqrt{(36500 - 36500)^2 + (4800 - 4800)^2 + (48 - 48)^2} = 0$$

From the calculation above, the results of the calculation for iteration 1 are as shown in the following table.

Table 4. Results from Iteration 1 calculations

No	PN	PP	TS	BR	C1	C2	C3	Cluster
1	P1	38000	10000	46	0	55045,44	5412,02	1
2	P2	92490	2200	48	55045,44	0	56050,34	2
3	P3	36500	4800	48	5412,02	56050,33	0	3
4	P4	78500	1400	49	41403,02	14012,85	42137,39	2
5	P5	37900	2100	46	7900,633	54590,09	3041,382	3
6	P6	62000	10000	49	24000	31471,89	26024,8	1
7	P7	11500	10000	48	26500	81364,73	25535,07	3
8	P8	129900	10000	49	91900	38214,5	93544,64	2
9	P9	42900	5100	46	43202,11	49674,72	6407,028	3
10	P10	62900	10000	47	24900	30600,79	26907,25	1
11	P11	31900	9000	46	33145,32	60970,39	6228,965	3
12	P12	29250	6000	47	9620,941	63354,07	7348,639	3
13	P13	24900	10000	46	68038,58	68038,58	12712,2	3
14	P14	57000	10000	46	19000	36337,03	21149,23	1
15	P15	89000	10000	49	51000	8545,18	52756,9	2
16	P16	24999	10000	47	13001	67940,23	12621,93	3
17	P17	61999	2300	48	62041,67	30491,16	25621,26	3
18	P18	49800	2300	48	14090,07	42690,12	13532,92	3
19	P19	88000	10000	48	50000	9000,006	51761,86	2
20	P20	44999	10000	47	0	55045,44	5412,024	1

After carrying out calculations using the cluster formula as in the table above, the cluster grouping based on the minimum distance to the closest centroid is:

Old cluster : (0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0)

New cluster : (1 2 3 2 3 1 3 2 3 1 3 3 3 1 2 3 3 3 2 1)

Judging from the statement above that there has been a change in the cluster, then continue to the next iteration:

Iteration 2 = K = 3

Centroid 1 cluster 1

$$C1 = \left(\frac{38000+62000+62900+57000+44999}{5} \right) = 52979,8$$

$$C2 = \left(\frac{10000+10000+10000+10000+10000}{5}\right) = 10000$$

$$C3 = \left(\frac{46+49+47+46+47}{5}\right) = 47$$

Centroid 2 cluster 2

$$C1 = \left(\frac{92490+78500+129900+89000+88000}{5}\right) = 95578$$

$$C2 = \left(\frac{2200+1400+10000+10000+10000}{5}\right) = 6720$$

$$C3 = \left(\frac{48+49+49+49+48}{5}\right) = 48,6$$

Centroid 3 cluster 3

$$C1 = \left(\frac{36500+37900+11500+42900+31900+29250+24900+24999+61999+49800}{10}\right) = 35164,8$$

$$C2 = \left(\frac{4800+2100+10000+5100+9000+6000+10000+10000+2300+2300}{10}\right) = 6160$$

$$C3 = \left(\frac{48+46+48+46+46+47+46+47+48+48}{10}\right) = 47$$

So we can produce K=3 centroid

C1 = (52979, 8, 10000, 47)

C2 = (95578, 6720, 48,6)

C3 = (35164, 8, 6160, 47)

- 1) A (38000,10000,46)

$$C1 = (52979, 8, 10000, 47) = \sqrt{(38000 - 52979,8)^2 + (10000 - 10000)^2 + (46 - 47)^2} = 14979,8$$

$$C2 = (95578, 6720, 48,6) = \sqrt{(38000 - 95578)^2 + (10000 - 6720)^2 + (46 - 48,6)^2} = 57671,35$$

$$C3 = (35164, 8, 6160, 47) = \sqrt{(38000 - 35164,8)^2 + (10000 - 6160)^2 + (46 - 47)^2} = 4773,25$$

- 2) B (92490, 2200,48)

$$C1 = (52979,8, 10000, 47) = \sqrt{(92490 - 52979,8)^2 + (2200 - 10000)^2 + (48 - 47)^2} = 40272,77$$

$$C2 = (95578, 6720, 48,6) = \sqrt{(92490 - 95578)^2 + (2200 - 6720)^2 + (48 - 48,6)^2} = 5474,134$$

$$C3 = (35164, 8, 6160, 47) = \sqrt{(92490 - 35164,8)^2 + (2200 - 6160)^2 + (48 - 47)^2} = 57461,81$$

- 3) C (36500, 4800,48)

$$C1 = (52979, 8, 10000, 47) = \sqrt{(36500 - 52979,8)^2 + (4800 - 10000)^2 + (47 - 47)^2} = 17280,74$$

$$C2 = (95578, 6720, 48,6) = \sqrt{(36500 - 95578)^2 + (4800 - 6720)^2 + (47 - 48,6)^2} = 59109,19$$

$$C3 = (35164, 8, 6160, 47) = \sqrt{(36500 - 35164,8)^2 + (4800 - 6160)^2 + (47 - 47)^2} = 1905,87$$

From the calculations carried out above, the results from iteration 2 are obtained, namely as in the following table.

Table 5. Results from Iteration 2 calculations

No	PN	PP	TS	BR	C1	C2	C3	Cluster
1	P1	38000	10000	46	14979,8	57671,35	4773,25	3
2	P2	92490	2200	48	40272,77	5474,134	57461,81	2
3	P3	36500	4800	48	17280,74	59109,19	1905,87	3
4	P4	78500	1400	49	26930,29	17887,44	43595,84	2
5	P5	37900	2100	46	17023,82	57862,73	4895,398	3
6	P6	62000	10000	49	9020,2	33737,82	27108,55	1
7	P7	11500	10000	48	41479,8	84141,95	23974,33	3
8	P8	129900	10000	49	76920,2	34478,37	94812,99	2
9	P9	42900	5100	46	11207,69	52702,9	7807,491	3
10	P10	62900	10000	47	9920,2	32842,2	27999,77	1
11	P11	31900	9000	46	21103,51	63718,8	4327,184	3
12	P12	29250	6000	47	24064,57	66331,91	5916,964	3
13	P13	24900	10000	46	28079,8	70754,07	10959,55	3
14	P14	57000	10000	46	4020,2	38717,19	22170,29	1
15	P15	89000	10000	49	36020,2	7350,407	53971,98	2
16	P16	24999	10000	47	27980,8	70655,17	10866,88	3
17	P17	61999	2300	48	11859	33868,65	27110,4	1
18	P18	49800	2300	48	8330,734	45990,89	15135,68	1

19	P19	88000	10000	48	35020,2	8257,39	52974,56	2
20	P20	44999	10000	47	7980,8	50685,24	10557,32	1

After carrying out calculations using the cluster formula as in the table above, the cluster grouping in iteration 2 based on the minimum distance to the closest centroid is:

Old cluster : (1 2 3 2 3 1 3 2 3 1 3 3 3 1 2 3 3 3 2 1)

New cluster : (3 2 3 2 3 1 3 2 3 1 3 3 3 1 2 3 1 1 2 1)

If a Cluster change occurs, it will continue with the next iteration

Iteration 3 = K=3

Centroid 1 cluster 1

$$C1 = \left(\frac{62000+129900+62900+57000+44999}{5} \right) = 56449,6$$

$$C2 = \left(\frac{10000+10000+10000+10000+10000}{5} \right) = 7433,3$$

$$C3 = \left(\frac{49+49+47+46+47}{5} \right) = 47,5$$

Centroid 2 cluster 2

$$C1 = \left(\frac{78500+89000+88000}{3} \right) = 95578$$

$$C2 = \left(\frac{1400+10000+10000}{3} \right) = 6720$$

$$C3 = \left(\frac{49+49+48}{3} \right) = 48,6$$

Centroid 3 cluster 3

$$C1 = \left(\frac{38000+92490+36500+37900+11500+42900+31900+29250+24900+24999+61999+49800}{12} \right) = 30872,1$$

$$C2 = \left(\frac{10000+2200+4800+2100+10000+5100+9000+6000+10000+10000+2300+2300}{12} \right) = 7444,4$$

$$C3 = \left(\frac{46+48+48+46+48+46+46+47+46+47+48+48}{12} \right) = 46,66$$

So we can produce K=3 centroids as:

C1 = (56449,6, 7433,3, 47,5)

C2 = (95578, 6720, 48,6)

C3 = (30872,1, 7444,4, 46,66)

1. A (38000,10000,46)

$$C1 = (56449,6, 7433,3, 47,5) = \sqrt{(38000 - 56449,6)^2 + (10000 - 7433,3)^2 + (46 - 47,5)^2} = 18627,28$$

$$C2 = (95578, 6720, 48,6) = \sqrt{(38000 - 95578)^2 + (10000 - 6720)^2 + (46 - 48,6)^2} = 57671,35$$

$$C3 = (30872,1, 7444,4, 46,66) = \sqrt{(38000 - 30872,1)^2 + (10000 - 7444,4)^2 + (46 - 46,66)^2} = 7572,189$$

2. B (92490, 2200,48)

$$C1 = (56449,6, 7433,3, 47,5) = \sqrt{(92490 - 56449,6)^2 + (2200 - 7433,3)^2 + (48 - 47,5)^2} = 36418,37$$

$$C2 = (95578, 6720, 48,6) = \sqrt{(92490 - 95578)^2 + (2200 - 6720)^2 + (48 - 48,6)^2} = 5474,134$$

$$C3 = (30872,1, 7444,4, 46,66) = \sqrt{(92490 - 30872,1)^2 + (2200 - 7444,4)^2 + (48 - 46,66)^2} = 61840,68$$

3. C (36500, 4800,48)

$$C1 = (56449,6, 7433,3, 47,5) = \sqrt{(36500 - 56449,6)^2 + (4800 - 7433,3)^2 + (47 - 47,5)^2} = 20122,64$$

$$C2 = (95578, 6720, 48,6) = \sqrt{(36500 - 95578)^2 + (4800 - 6720)^2 + (47 - 48,6)^2} = 59109,19$$

$$C3 = (30872,1, 7444,4, 46,66) = \sqrt{(36500 - 30872,1)^2 + (4800 - 7444,4)^2 + (47 - 46,66)^2} = 6218,208$$

From the calculations carried out as above, the results from iteration 3 are obtained, namely as in the following table.

Table 6. Results from Iteration 3 calculations

No	PN	PP	TS	BR	C1	C2	C3	Cluster
1	P1	38000	10000	46	18627,28	57671,35	7572,189	3
2	P2	92490	2200	48	36418,37	5474,134	61840,68	2
3	P3	36500	4800	48	20122,64	59109,19	6218,208	3
4	P4	78500	1400	49	22860,9	17887,44	48009,91	2
5	P5	37900	2100	46	19301,08	57862,73	8829,156	3
6	P6	62000	10000	49	6115,136	33737,82	31232,63	1
7	P7	11500	10000	48	45022,82	84141,95	19539,94	3
8	P8	129900	10000	49	73495,23	34478,37	99060,87	2
9	P9	42900	5100	46	13749,03	52702,9	12254,25	3
10	P10	62900	10000	47	6942,306	32842,2	32129,7	1
11	P11	31900	9000	46	24599,54	63718,8	1864,529	3
12	P12	29250	6000	47	27237,34	66331,91	2171,981	3
13	P13	24900	10000	46	31653,83	70754,07	6495,927	3
14	P14	57000	10000	46	2625,051	38717,19	26252,59	1
15	P15	89000	10000	49	32651,44	7350,407	58184,05	2
16	P16	24999	10000	47	31555,16	70655,17	6405,029	3
17	P17	61999	2300	48	7559,538	33868,65	31549,15	1
18	P18	49800	2300	48	8400,473	45990,89	19614,54	1
19	P19	88000	10000	48	31654,63	8257,39	57185,03	2
20	P20	44999	10000	47	11734,74	50685,24	14356,2	1

After carrying out calculations using the cluster formula as in the table above, the cluster grouping in iteration 3 based on the minimum distance to the closest centroid is:

Old cluster : (3 2 3 2 3 1 3 2 3 1 3 3 3 1 2 3 1 1 2 1)

New cluster : (3 2 3 2 3 1 3 2 3 1 3 3 3 1 2 3 1 1 2 1)

As can be seen in the centroid results obtained from iteration 3 there was no change, so the search for the next iteration was stopped.

3.4. K-Means Clustering Visualization

Data import was carried out using the Python programming language using Google Colab. This initial step is important in the K-Means Clustering visualization process. To import data, you can upload data to the Google Colab directory. After that, make sure the file has been uploaded to the Google Colab directory. Then make sure Pandas is installed so we can display the data on the Google Colab worksheet. In this research, researchers used the format (.xlsx) as a dataset.

This transformation stage is a process that is also important to carry out when visualizing K-Means Clustering, where this stage is related to data analysis to change, adjust, or analyze so that it can meet the requirements for K-Means Clustering. Data is identified by deleting empty rows or columns and then taking the form of integer numbers. The results of this transformation can be seen in the image below.

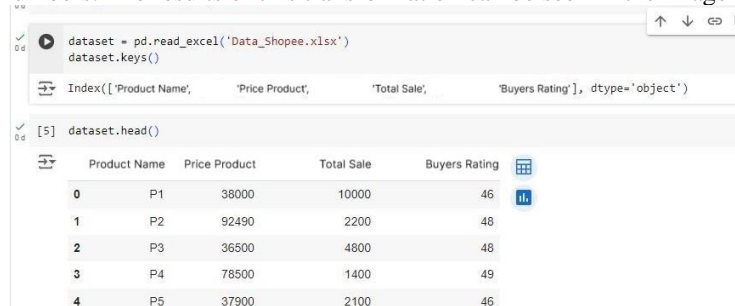


Figure 3. Displays the transformed dataset

The data selection stage involves analyzing the data to identify and select a subset that is relevant and suitable for the intended analysis. Researchers must choose which attributes are directly related to the grouping objectives. In this study, the researchers deleted the Product Name column because this attribute was not needed. The resulting attributes after selection can be seen in the following image.

```
dataset=dataset.dropna(axis=1)
X=dataset.drop(['Product Name'], axis=1)
X.head()
```

	Price Product	Total Sale	Buyers Rating
0	38000	10000	46
1	92490	2200	48
2	36500	4800	48
3	78500	1400	49
4	37900	2100	46

Figure 4. Display of data selection results

At this stage, it was carried out on Google Colab with the Python programming language. This clustering process is carried out by starting with determining the number of cluster as shown in the following image.

```
kmeans = KMeans(n_clusters=3)
kmeans.fit(X)
```

/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: KMeans

KMeans(n_clusters=3)

Figure 5. Determine the number of Clusters

Next, we can see on the label that the grouping has been carried out marked with numbers (0, 1, 2) with the initial centroid point randomly determined as follows.

```
0] kmeans.labels_
array([1, 2, 1, ..., 2, 1, 0], dtype=int32)
```

```
ypredict=kmeans.labels_
dataset['cluster']=ypredict
dataset.head()
```

	Price Product	Total Sale	Buyers Rating	cluster
0	38000	10000	46	1
1	92490	2200	48	2
2	36500	4800	48	1
3	78500	1400	49	2
4	37900	2100	46	1

Figure 6. The results of the data have been grouped by cluster

The dataset that has been successfully downloaded from Google Colab is the result of data that has been grouped based on clusters. The data set consists of 1200 rows of data. The data format can be seen in Table 7.

Table 7. Cluster results dataset on the downloaded program

Product Name	Price Product	Total Sale	Buyers Rating	Cluster
P1	38000	10000	46	0
P2	92490	2200	48	1
P3	36500	4800	48	0
P4	78500	1400	49	1
P5	37900	2100	46	0
P1200	259000	652	49	2

After that, the researcher visualized the data that had been determined by searching for clusters and centroids into a graph marked with a round 'rainbow' color and black centroids with a star shape. The following are the results of Clustering where the centroid center point has been determined.

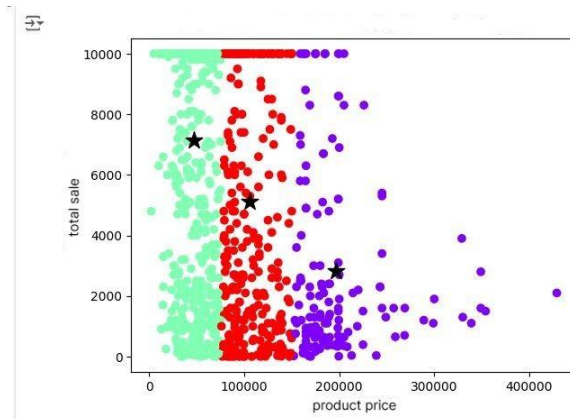


Figure 7. K-Means Clustering Visualization Results

To find out the determination of the final centroid point and the number of each cluster member and the cluster diagram can be seen in the image below.

```
print(kmeans.cluster_centers_)
[[4.65315216e+04 7.13912517e+03 4.64478442e+01]
 [1.05360479e+05 5.11664634e+03 4.68750000e+01]
 [1.96913830e+05 2.81147059e+03 4.84183007e+01]]
```

Figure 8. The final result of placing the centroid point

It can be seen in the picture above, the results of determining the final centroid point are in cluster 0 (4.6531, 7.1391, 4.6447), in cluster 1 (1.0536, 5.1166, 4.6875) and cluster 2 (1.9691, 2.8114, 4.8418) from the three variables, namely product price, number of sales and buyer ratings.

```
dataset['cluster'].value_counts()
cluster
0    719
1    328
2    153
Name: count, dtype: int64
```

Figure 9. Number of each Cluster member

In terms of the number of members of each cluster, it can be seen in the picture above that cluster 0 has 719 data, cluster 1 has 328 data, and cluster 2 has 153 data from 4 types of women's clothing, namely tunics, dresses, abaya/gamis, and shirts. The following is an explanation in the form of a bar chart.

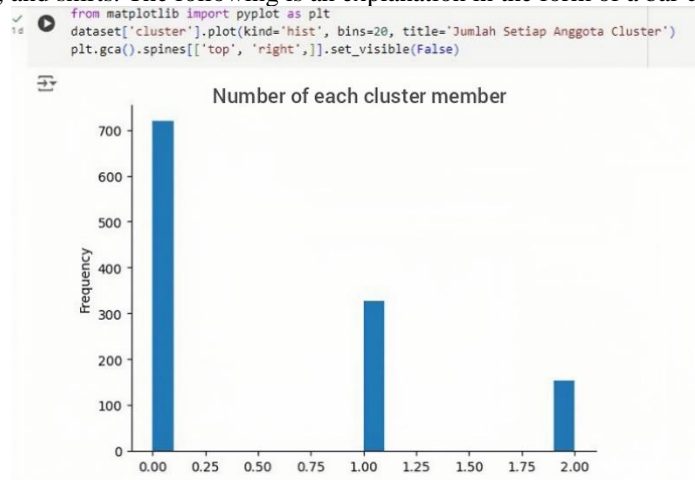


Figure 10. Cluster diagram

The diagrams were created using MS Excel with data obtained from downloading previous data for grouping by type of women's clothing according to the cluster as follows.

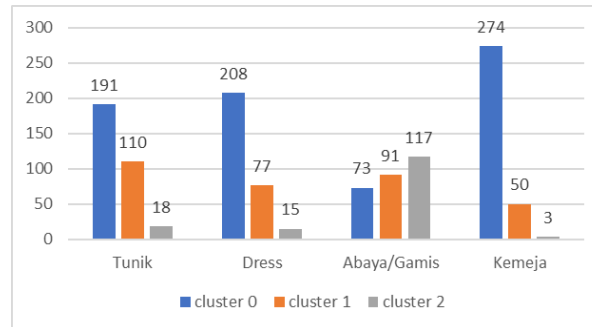


Figure 11. Cluster diagram of women's clothing types

After carrying out calculations using Google Collab with the Python programming language, the calculation results are as follows:

- 1) Cluster 0 with centroid 4.64315, 7.13912, 4.64478 (there are 719 data, consisting of 191 Tunic data, 208 Dress data, 73 Abaya/Gamis data, and 274 Shirt data) it can be decided that Cluster 1 is the best-selling group with the number of sales in an average of 7000 products with a price of around > Rp 80.000.
- 2) Cluster 1 with centroid 1.05360, 5.11664, 4.6875 (there are 328 data, consisting of 110 Tunic data, 77 Dress data, 91 Abaya/Gamis data, and 50 Shirt data) it can be decided that Cluster 2 is the best-selling group with average sales. -an average of 50,000 products with a price of around Rp 80.000 - Rp 150.000.
- 3) Cluster 2 with centroid 1.96913, 2.81147, 4.84183 (there are 153 data, consisting of 18 Tunic data, 15 Dress data, 117 Abaya/Gamis data, and 3 Shirt data) it can be decided that Cluster 0 is the least popular group with the number of sales in an average of 3000 products with a price of around > Rp 150.000.

4. CONCLUSION

The conclusions that can be drawn from the application of the K-Means Clustering algorithm for women's clothing sales data in the marketplace on Shopee can be described as follows. Based on calculations carried out by researchers using the K-Means Clustering algorithm, the characteristics of each group of women's clothing sales data can be identified. Cluster 0 is the best-selling group with average sales data of 7000 products with a price of around > Rp 80.000 and the most common type of women's clothing is shirts. Cluster 1 is the best-selling group with average sales data of 5000 products at a price of around Rp 80.000 - Rp 150.000 and the most common type of women's clothing is tunics. Cluster 2 is the least popular group with average sales data of 3000 products with a price of around > Rp 150.000 with the most common type of women's clothing being Abaya/Gamis.

The K-Means Clustering algorithm is a centroid-based algorithm or distance-based algorithm, where we calculate the distance to assign a point to a cluster. This research aims to group data on sales of the best-selling women's clothing in the marketplace based on clusters with 3 attributes, namely product price, number of sales, and buyer assessment. The clustering results of several attributes provide information about the characteristics of each group in the women's clothing sales data in that marketplace.



Data visualization is the process of using visual elements such as diagrams, graphs, or maps to represent data. In this research, visualization of clustering data on sales of the best-selling women's clothing in this marketplace uses the K-Means algorithm by utilizing Google Collab, carrying out several steps, namely, data preparation, data transformation, data selection, determining centroid and clustering, creating graphs of clustering results and displaying points. Final centroid and number of members of each cluster.

REFERENCES

- [1] Hidayat MA. Desain Iklan pada Marketplace untuk Menarik Minat Konsumen (Studi Iklan Online Shop Shopee). Tugas Akhir. Universitas Islam Kalimantan Muhammad Arsyad Albanjari, 2020.
- [2] Rahmawati K. Pelatihan Penjualan Online Menggunakan Marketplace pada UKM di Bantul. *DHARMA: Jurnal Pengabdian Masyarakat*. 2021; 2(1): 79–85.
- [3] Hartono S, Hendrawan T, Pratama AI. "Malik" (Aplikasi Marketplace Busana Muslim). *Infotech: Journal of Technology Information*. 2023; 9(1): 31–6.
- [4] Artaya IP, Purworusmiardi T. Efektifitas Marketplace dalam Meningkatkan Konsentrasi Pemasaran dan Penjualan Produk Bagi UMKM di Jawa Timur. Surabaya: Universitas Narotama Surabaya; 2019.
- [5] Safitri LA, Dewa CB. Analisa Pengaruh Masa New Normal pada Penjualan Online Melalui E-Commerce Shopee. *Jurnal Manajemen DayaSaing*. 2020; 22(2): 117–25.
- [6] Fauziah F. Strategi Komunikasi Bisnis Online Shop "Shopee" dalam Meningkatkan Penjualan. *Abiwarra: Jurnal Vokasi Administrasi Bisnis*. 2020; 1(2): 45–53.

- [7] Sari EP, Pudjiarti E, Susanti H. Sistem Informasi Penjualan Pakaian Wanita Berbasis Web (E-Commerce) pada PT Bunitop Indonesia. *Jurnal Teknologi Informasi Mura*. 2020; 12(1): 1–13.
- [8] Muktar MR, Syam MRF, Kunda A, Natsir MS. Analisis Penerapan Data Mining dalam Klasifikasi Penjualan Pakaian pada Toko Online Shopee Menggunakan Algoritma C4.5. *Jurnal Dipanegara Komputer Sistem Informasi*. 2023; 17(1): 1–8.
- [9] Kurniawan RA, Hasibuan MS, Piramida, Ramadhan RS. Penerapan Algoritma K-Means untuk Clustering Tempat Makan di Batubara. *Cosie (Journal of Computer Science and Informatics Engineering)*. 2022; 1(1): 10–8.
- [10] Nasir J. Penerapan Data Mining Clustering Dalam Mengelompokkan Buku Dengan Metode K-Means. *Jurnal SIMETRIS*. 2020; 11(2): 1–13.
- [11] Dewi AOP. Big Data di Perpustakaan dengan Memanfaatkan Data Mining. *ANUVA*. 2020; 4(2): 223–30.
- [12] Riadi R, Mesran. Penerapan Data Mining Menggunakan Algoritma K-Means untuk Analisa Penjualan Parfume. *Journal of Informatics, Electrical and Electronics Engineering*. 2023; 2(4): 138–45.
- [13] Zai C. Implementasi Data Mining Sebagai Pengolahan Data. *Jurnal Portal Data*. 2022; 2(3): 1–12.
- [14] Aditya A, Jovian I, Sari BN. Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. *Jurnal Media Informatika Budidarma*. 2020; 4(1): 51–8. <https://doi.org/https://doi.org/10.30865/mib.v4i1.1784>.
- [15] Turnip HN, Fahmi H. Penerapan Data Mining pada Penjualan Kartu Paket Internet yang Banyak Diminati Konsumen dengan Metode K-Means. 2021; 4(2): 36–41.
- [16] Gustientiedina G, Adiya MH, Desnelita Y. Penerapan Algoritma K-Means untuk Clustering Data Obat-Obatan. *Jurnal Nasional Teknologi Dan Sistem Informasi*. 2019; 5(1): 17–24.
- [17] Tinendung IS, Zufria I. Pengelompokan Status Stunting pada Anak Menggunakan Metode K-Means Clustering. *Jurnal Media Informatika Budidarma*. 2023; 7(4): 2014–23.
- [18] Nuryani I, Darwis D. Analisis Clustering pada Pengguna Brand HP Menggunakan Metode K-Means. *Prosiding Seminar Nasional Ilmu Komputer*. 2021; 1(1): 190–211.
- [19] Dinata RK, Safwandi, Hasdyna N, Azizah N. Analisis K-Means Clustering pada Data Sepeda Motor. *Informatics Journal*. 2020; 5(1): 10–7.
- [20] Guntara RG. Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7. *Jurnal Teknologi Dan Sistem Informasi Bisnis*. 2023; 5(1): 55–60. <https://doi.org/10.47233/jteksis.v5i1.750>.

BIBLIOGRAPHY OF AUTHORS

	<p>Rizna Fitriana Dalimunthe is currently a computer science undergraduate student at the State Islamic University of North Sumatra, Indonesia. She is interested in deep learning about data mining, namely data processing, methods such as K-means Clustering and visualizing data with the Python programming language. She can be contacted at the email address: riznafitriana5@gmail.com</p>
	<p>Raisa Amanda Putri received her S.Kom degree from STMIK Mikroskil University, Medan, North Sumatra in 2011. Then she continued her education by earning an M.T.I degree from Bina Nusantara University in 2015. Currently she serves as a lecturer in information systems at the State Islamic University of Sumatra. north. One of his research interests is data mining. She can be contacted at the email address: raissa.ap@uinsu.ac.id</p>