Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Implementation of the K-Means Clustering Algorithm Based on CRISP-DM Using Orange for Sales Product Grouping in Shopee.

Septi Giseila Putriana[1]*, Rika Yani[2], Pujianto[3]

[1,2,3]*Department of Informatics, Universitas Baturaja, Indonesia*
giseilaputriana19@gmail.com[1*], rikayani2233@gmail.com[2], pujianto.mail@gmail.com[3]

Abstract

The growth of e-commerce in Indonesia, especially on Shopee, has generated large product data that are not yet fully used by sellers in decision-making. This study aimed to cluster Shopee product sales data based on physical characteristics and product information completeness using the K-Means Clustering algorithm within the CRISP-DM framework through Orange Data Mining. The dataset included 4,997 product records with eight attributes: product category, product name length, product description length, number of product photos, product weight, and product length, height, and width. After data cleaning, 4,895 records were analyzed. The results identified three clusters C1 contained large and heavy products, C2 contained products with general and dominant characteristics, and C3 contained relatively small to medium products. The silhouette coefficient score was 0.238, indicating a weak cluster structure, but the model still provided an overview of product grouping based on physical characteristics.

Keywords: *Data Mining, K-Means Clustering, CRISP-DM, Orange, Shopee*

1. Introduction

The development of information and communication technology has changed public behavior in buying and selling transactions, shifting from conventional methods to online transactions through e-commerce platforms. Shopee is one of the largest marketplaces in Indonesia, connecting millions of sellers and buyers within a single platform. Every activity on the platform generates a large amount of data, including transaction data and product data, which, if properly processed, can serve as a valuable source of information for sellers in business decision-making. In practice, Shopee sellers manage products with highly diverse quantities and characteristics, ranging from small and lightweight items to large and heavy products. This diversity creates practical challenges because manual product inspection requires substantial time and is prone to error. Sellers often find it difficult to identify product groups with similar characteristics, even though such grouping is needed to manage inventory strategies, estimate packaging and shipping costs, and design promotions that match product types. This problem can be addressed through data mining, which refers to the process of discovering patterns or knowledge from large datasets. One of the most widely used data mining techniques for clustering is the K-Means Clustering algorithm, which partitions data into several groups based on similarity in characteristics. Previous studies have shown the effectiveness of this method[1]. clustered smartphone sales at Konter Mutiara Cell using K-Means in Orange and produced three price clusters: low, medium, and high clustered sales data from CV. Widuri using Orange and produced three product clusters, ranging from fast-moving to slow-moving items[2]. Meanwhile, Nawawi et al. (2021) applied K-Means in Orange to identify the best-selling Muslim fashion products.

Unlike those studies, which generally clustered conventional store sales data based on price and sales volume, this study clusters product data on the Shopee e-commerce platform based on physical characteristics and product information completeness, namely weight, dimensions, number of photos, product name length, and product description length. In addition, the data mining process in this study is systematically structured using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which consists of six stages and allows each research step to be accountable and reproducible[3]. The use of CRISP-DM in a marketplace product case study represents the novelty of this research. Based on this description, this study aims to implement the K-Means Clustering algorithm within the CRISP-DM framework using Orange to cluster Shopee sales products, so that product groups with similar characteristics can be identified and utilized by sellers in inventory management, shipping, and promotional strategies.

2. Theoretical Review

2.1. Data Mining

Data mining is the process of discovering useful patterns, relationships, or knowledge from large datasets by applying statistical techniques, artificial intelligence, and machine learning. It addresses the problem of abundant data that does not always translate into useful

information. Through data mining, data that were previously stored only as records can be transformed into knowledge that supports decision-making. In business contexts, data mining is widely used to understand consumer behavior, group products, and formulate marketing strategies.

2.2. CRISP-DM

CRISP-DM, or the Cross-Industry Standard Process for Data Mining, is a standard industry framework for conducting data mining projects. It consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The framework is iterative and independent of specific software, which makes it applicable to a wide range of case studies. The use of CRISP-DM makes the data mining process more structured, well documented, and easier to replicate.

2.3. Clustering and the K-Means Algorithm

Clustering is an unsupervised data mining technique that groups data objects into several clusters so that objects within the same cluster are highly similar, while objects in different clusters are substantially different. K-Means is a partition-based clustering algorithm that divides data into k clusters. The algorithm works by defining the number of clusters, selecting initial centroids, calculating the distance between each data point and each centroid, assigning each point to the nearest centroid, and updating the centroid positions based on the mean of the cluster members. This process is repeated until the centroids no longer change or the maximum number of iterations is reached. The distance between data points and centroids is generally calculated using Euclidean distance, which is the square root of the sum of squared differences between the attributes of the data object and the centroid. K-Means is widely used because it is simple, fast, suitable for numerical data, and easy to interpret[4].

2.4. Silhouette Coefficient

The silhouette coefficient is a metric used to evaluate clustering quality by comparing the closeness of an object to members of its own cluster and to the nearest neighboring cluster. The silhouette value ranges from -1 to 1, and values closer to 1 indicate that the object is more appropriately assigned to its cluster. Kaufman and Rousseeuw classify average silhouette values above 0.70 as indicating a strong structure, 0.51 to 0.70 as a good structure, 0.26 to 0.50 as a weak structure, and below 0.26 as a poor structure.

2.5. Orange Data Mining

Orange is an open-source data mining and machine learning software based on visual components and developed using Python. Analysis in Orange is performed by connecting widgets such as File, Data Table, Preprocess, and k-Means into a workflow, allowing users to analyze data without writing program code. This ease of use has made Orange widely applied in sales data clustering studies[5].

2.6. Previous Studies

Several previous studies serve as references for this research. applied K-Means in Orange to cluster smartphone sales and obtained three price-based clusters. grouped sales data from CV. Widuri into three product clusters based on sales performance. Used K-Means in Orange to identify best-selling Muslim fashion products, and Mutaqin 2023 applied K-Means to online transportation balance sales. These studies show that K-Means is effective for sales data clustering. Based on this foundation, this study assumes that Shopee product sales data can also be grouped into several distinct clusters based on product physical characteristics and information completeness.

3. Methodology

This study employed a descriptive quantitative approach with a data mining framework based on CRISP-DM. The research design consisted of six stages, as shown in Figure 1, namely business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Fig. 1: Research Stage Using the CRISP-DM Framework

The sample comprised 4,997 product records obtained from the Shopee marketplace product dataset. After data cleaning and the removal of incomplete records, 4,895 data records were used in the clustering process. Data were collected through documentation, namely by downloading and recording product data from the store management page on Shopee into a Microsoft Excel file, as well as through literature review to establish the theoretical foundation from relevant journals and books.

The analytical instrument used in this study was Orange Data Mining version 3 with the K-Means Clustering algorithm. The number of clusters was set at three (k = 3), with centroid initialization using the Kmeans+ method, a maximum of 5,000 iterations, and ten repetitions

of the initialization process to obtain stable results. Prior to modeling, all numerical attributes were normalized using the min-max method into the interval from 0 to 1, in which the original value was subtracted by the minimum attribute value and divided by the difference between the maximum and minimum attribute values. This was done to prevent differences in attribute scales from dominating the distance calculation. The distance between data points and centroids was calculated using Euclidean distance, as described in the theoretical review, while the quality of the clustering results was evaluated using the silhouette coefficient, with assessment categories referring[6].

4. Results and Discussion

Data collection was conducted using a Shopee marketplace product dataset obtained from a secondary data source and stored in the file. The initial dataset consisted of 4,997 product records. After data cleaning and the removal of records with missing values, 4,895 product records remained and were used in the analysis. The dataset contained one product identity column and eight attributes describing product characteristics, namely product category, product name length, product description length, number of product photos, product weight, product length, product height, and product width. Data analysis was carried out using Orange Data Mining through stages based on the CRISP-DM framework, namely business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

4.1. Business Understanding

This stage aimed to understand the business problem faced by Shopee sellers, namely the difficulty of grouping products with diverse characteristics for inventory management, packaging and shipping cost estimation, and promotion planning. The defined data mining objective was to form product groups with similar physical characteristics and information completeness using the K-Means algorithm. [7]

4.2. Data Understanding

At this stage, the structure of the dataset was examined. The dataset consisted of 100 product records with attributes as described in Table 1. The product_category_name attribute was categorical, while the other seven attributes were numerical and used as the basis for distance calculation in the clustering process. The product_id column served as a meta attribute or identifier and was therefore excluded from the calculation process.

Table 1. Description of Product Dataset Attributes

No.	Attribute Name	Data Type	Description
1	product_id	Text (meta)	Unique product identifier code
2	product_category_name	Categorical	Category of the sold product
3	product_name_length	Numeric	Number of characters in the product name
4	product_description_length	Numeric	Number of characters in the product description
5	product_photos_qty	Numeric	Number of uploaded product photos
6	product_weight_g	Numeric	Product weight in grams
7	product_length_cm	Numeric	Product length in centimeters
8	product_height_cm	Numeric	Product height in centimeters
9	product_width_cm	Numeric	Product width in centimeters

4.3. Data Preparation

The data preparation stage involved selecting relevant attributes, checking data completeness, and transforming the data. The inspection showed that no missing data were found in any of the variables used. Next, min-max normalization was applied through the Preprocess widget using the Normalize to interval [0, 1] option, so that all numerical attributes were scaled to the same range. This normalization step was important because the range of the weight attribute, measured in grams, was much larger than the range of the dimension attributes, measured in centimeters, and the number of photos[8]. Without normalization, the weight attribute would dominate the distance calculation. An example of the transformed data is presented in Table 2.

Table 2. Example of Min-Max Normalization Results

Attribute	Original Value	Normalized Value
product_weight_g	225 g	0.012
product_length_cm	16 cm	0.146
Product_photos_qty	4 photos	0.200

4.4. Modeling

The modeling stage was carried out by constructing a workflow in the Orange application, as shown in Figure 2. The File widget read the file, after which the data were displayed in the Data Table widget named Shopee Sales Data. The selected data were then sent to the Preprocess widget for normalization, displayed again in the Data Table widget, and subsequently processed by the k-Means widget with the number of clusters set to three[9]. The model output was then forwarded to the Scatter Plot widget for cluster distribution visualization and to the Data Table 1 widget to display the clustering results along with the silhouette value for each data point.

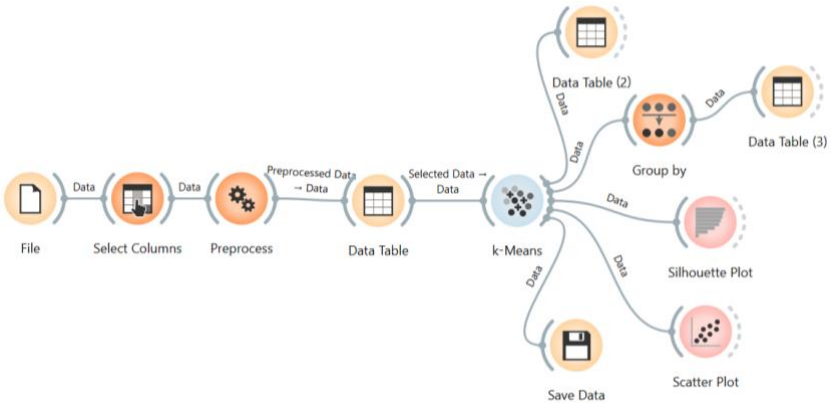

Fig 2. Product Clustering Workflow in Orange

The modeling results added two new columns to the dataset: the Cluster column, which indicates each product's cluster membership (C1, C2, or C3), and the Silhouette column, which indicates the quality of each product's assignment to its cluster.

4.5. Evaluation

The clustering results produced three clusters with different distributions, which were evaluated based on the number of members and the silhouette value of each cluster, as summarized in Table 3. Cluster C2 was the largest cluster, containing 2,764 products (56.47%), followed by Cluster C3 with 1,315 products (26.86%), and Cluster C1 with 816 products (16.67%). The silhouette coefficient evaluation in Orange produced a value of 0.238. This value indicates that the cluster structure is still relatively weak; however, the grouping was still able to separate products based on their physical characteristics and product information. This score is classified as a weak cluster structure, although it still provides a useful basis for product grouping.

Table 3. Distribution of Members and Silhouette Values for Each Cluster

Cluster	Number of Members	Percentage	Dominant Characteristics
C1	816	16,67%	Large and heavy products
C2	2.764	56,47%	Products with relatively low dimensions and weight
C3	1.315	26,86%	Small to medium-sized products
Total	4.895	100%	

The average silhouette score from the clustering evaluation was 0.238. This value indicates that the resulting cluster structure is still relatively weak, but it was still able to distinguish product characteristics based on the selected attributes. The characteristics of each cluster can be observed from the centroid values in Table 4. Cluster C1 had low centroid values across all physical attributes, indicating that it contained small and lightweight products, such as accessories and beauty products. Cluster C2 had the highest centroid values for weight and dimensions, indicating that it contained large and heavy products, such as household equipment and furniture. Cluster C3 was positioned between the two, representing medium-sized products, such as fashion items and toys.

Table 4. Centroid Values of Each Cluster (Normalized Data)

Attribute	C1	C2	C3
poduct_name_length	0,694	0,754	0,438
product_description_length	0,229	0,198	0,161
product_photos_qty	0,102	0,081	0,051
product_weight_g	0,285	0,032	0,031
product_length_cm	0,516	0,181	0,178
product_height_cm	0,262	0,114	0,124
product_width_cm	0,325	0,143	0,139

The cluster distribution visualization is shown in Figure 3, which maps the normalized product weight attribute (product_weight_g) against product width (product_width_cm). The figure shows three fairly distinct groups: C1 concentrated in the low-value area, C3 in the middle-value area, and C2 in the high-value area. This pattern is consistent with the centroid values presented in Table 4.

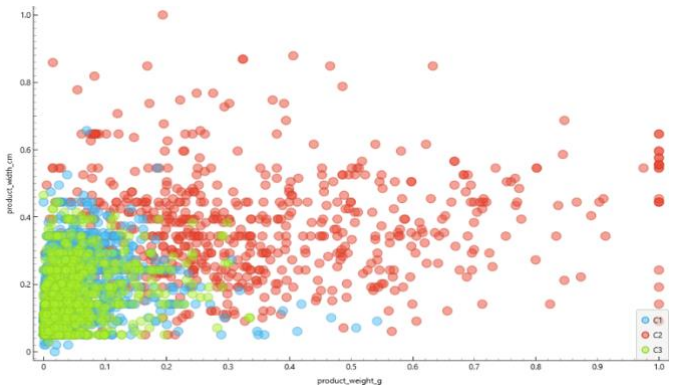

Fig. 3: Scatter Plot of Clustering Results for Product Weight and Width Attributes

4.6. Deployment

The clustering results were applied as a strategic recommendation for sellers. Products in Cluster C1, which are small and lightweight, can be used as complementary items in bundling packages and for free-shipping promotions because their shipping costs are low. Products in Cluster C2, which are large and heavy, require special attention in packaging, the selection of cargo delivery services, and shipping cost calculations to avoid reducing profit margins. Products in Cluster C3 can be managed using standard inventory strategies. This grouping also helps sellers organize warehouses and determine stock priorities based on product categories.

4.7. Discussion

The findings of this study are consistent, who also obtained three clusters using the K-Means algorithm in the Orange application[10]. These studies reinforce the view that K-Means is effective and easy to interpret for numerical sales data[11]. The difference is that previous studies grouped data based on price or sales performance in conventional stores, whereas this study grouped marketplace products based on physical characteristics and product information completeness using a CRISP-DM-based process.
From a theoretical perspective, these findings support the initial assumption that e-commerce products can be meaningfully clustered based on physical attributes. However, the silhouette score of 0.238 indicates that the boundaries between clusters are not yet strongly formed. This condition may result from the high similarity among product characteristics, which causes some products to be close to more than one cluster[12]. Even so, the clustering results still provide initial insight into product grouping patterns based on weight, dimensions, and product information completeness[13].

5. Conclusion

Based on the results of the study, the following conclusions can be drawn:

a. The K-Means Clustering algorithm based on the CRISP-DM framework was successfully implemented using Orange to cluster 4,895 Shopee sales product records through six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This method is appropriate because it is suitable for numerical data, easy to apply without programming, and produces results that are easy for sellers to interpret.
b. Clustering with K-Means using three clusters produced Cluster C1 with 816 products (16.67%), Cluster C2 with 2,764 products (56.47%), and Cluster C3 with 1,315 products (26.86%). Evaluation using the silhouette coefficient produced a value of 0.238. The clustering results can be used as a basis for inventory management, packaging, shipping cost calculation, and product promotion strategies on the Shopee marketplace.

Acknowledgement

The authors would like to express their sincere gratitude to Mr. Pujianto, S.Kom., M.CS., as the supervising lecturer, for his guidance, insight, and constructive suggestions throughout the preparation of this study. The authors also extend their appreciation to Rika Yani for her valuable collaboration and contribution as a research partner. In addition, the authors acknowledge all individuals and parties who provided support, encouragement, and assistance during the research process. Their contributions were instrumental in the successful completion of this study.

References

[1] M. Cell, “Implementasi Algoritma K-Means Clustering Menggunakan Orange Untuk Mengelompokan Penjualan Smartphone (Studi Kasus Konter,” vol. 4, pp. 1371–1379, 2024.
[2] P. Cv and W. Menggunakan, “Implementasi Algoritma K-Means Dalam Pengelompokan Data,” vol. 2, no. 1, pp. 188–196, 2023.
[3] P. M. A. K-means, “Data mining clustering dalam pengelompokan buku perpustakaan mengunakan algoritma k-means,” vol. 8, no. 3, pp. 802–814, 2023.
[4] N. N. K. Kaylista, N. Khoirunnisaa, G. V. E. N. F, and A. Y. Pratama, “Implementasi Algoritma K-Means Clustering Menggunakan Aplikasi Orange Untuk Mengetahui Pola Indeks Pembangunan Manusia Tahun 2022,” vol. 4, no. 1, pp. 65–76, 2023.
[5] M. S. Nawawi, F. Sembiring, and A. Erfina, “Implementasi Algoritma K-Means Clustering Menggunakan Orange Untuk Penentuan Produk Busana Muslim Terlaris,” pp. 789–797, 2021.
[6] F. T. Informasi, U. Kristen, and S. Wacana, “Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means,” vol. 3, no. 3, pp. 424–439, 2021.
[7] A. Maulana and K. N. Akbar, “Penerapan Clustering Menggunakan Algoritma K-Means Sebagai Analisis Produksi Komoditas Perikanan Provinsi di Indonesia,” vol. 01, no. 01, pp. 34–39, 2021.
[8] “inotek,+buku+2+artikel+10+(fakhry).pdf.”
[9] S. Nasional, I. Teknologi, F. A. Bramasta, R. Helilintar, T. Informatika, and F. Teknik, “Penerapan Data Mining Untuk Menentukan Strategi Penjualan Toko Sepatu,” pp. 236–241, 2021.
[10] D. Muriyatmoko, D. Fikrianti, and F. R. Ifalus, “Analisis Penjualan Produk Terlaris di Toko Bangunan Pekanbaru Jaya Menggunakan Metode Clustering K-Means,” vol. 4, pp. 88–93, 2025.
[11] F. Randawan and D. W. Widodo, “Penjualan Aksesoris Motor dan Mobil dengan Metode Clustering,” 2021.
[12] M. A. K-means, “Analisa Clustering Pada Penerima Pupuk Subsidi,” vol. 7, pp. 212–219, 2023.
[13] F. Putri, A. Hasibuan, S. Sumarno, and I. Parlina, “Penerapan K-Means pada Pengelompokan Penjualan Produk Smartphone,” vol. 1, no. 1, pp. 15–20, 2021, doi: 10.54259/satesi.v1i1.3.