https://doi.org/10.32815/jitika.1249

Volume 20, Issue 2, 2026

Jurnal Ilmiah Teknologi Informasi Asia

Journal Homepage: https://jurnal.asia.ac.id/index.php/jitika

Article

Perfume Product Segmentation on Marketplace Using Web Scraping and K-Means Algorithm

Suryo Atmojo[a,*], Suzana Dewi [a], Isnaini Muhandhis [a], Laily Alfina Wulansari [a], Ruli Utami [b]

[a] *Program Studi Teknik Informatika, Fakultas Teknik,Universitas Wijaya Putra, Surabaya, Indonesia*
[b] *Program Studi Sistem Informasi, ITATS, Surabaya, Indonesia*

***Abstract*—The rapid development of e-commerce in Indonesia, particularly Tokopedia, has significantly shifted consumer consumption patterns towards secondary products such as perfume. However, Micro, Small, and Medium Enterprises (MSMEs) often struggle to understand competitive market characteristics due to limited access to clear, usable analytical data. This study aims to address this issue by developing an automated data collection system and performing perfume product market segmentation using a data-driven approach. The methodology used in this study combines Web Scraping techniques and Machine Learning algorithms. Product data was automatically collected from Tokopedia using Node.js and Playwright scripts, followed by a data preprocessing stage that included price cleaning, sales volume conversion, and attribute normalization. Segmentation analysis was performed using the K-Means Clustering algorithm, with the optimal number of clusters determined using the Elbow Method. The results from 995 successfully collected product data points indicate that the perfume product market on Tokopedia is divided into three main segments (k=3). Cluster 0 is the dominant segment (98% of the population), representing the mass market with affordable prices and high sales volume. Cluster 1 was identified as anomalous data or outliers with extreme prices but no sales, while Cluster 2 represents a niche market (premium) with high prices and limited sales volume. The conclusion of this study suggests that combining web scraping with K-Means is effective for mapping product positions in the digital market. The output of this research is expected to assist MSMEs in formulating more targeted pricing and promotional strategies based on real data rather than mere assumptions.**

***Keywords*—k-means; marketplace; perfume; product segmentation; web scraping.**

1. Introduction

The advancement of information technology has had a significant impact on public consumption patterns, particularly with the rapid growth of Indonesia's e-commerce ecosystem. Secondary and tertiary need products, such as perfume, are now sold not only in physical stores but also on marketplaces like Tokopedia, Shopee, and Lazada. According to recent reports, more than 88% of internet users in Indonesia have engaged in online transactions, including the purchase of beauty products and perfumes (Wardhana et al., 2012; Dewi, 2025).

The figure above illustrates the user interface of the Tokopedia marketplace, where various men’s perfume products are displayed, with key information such as product names, prices, shop locations, and ratings visible to potential buyers.

However, behind these high consumption figures, business actors—especially MSMEs and local brands—often face challenges in positioning their products correctly. One of the

* Corresponding author.
E-mail Address: suryoatm@gmail.com (S. Atmojo)

Author E-mail(s): SA (suryoatm@gmail.com), SD (suzanadewi@uwp.ac.id), IM (isnainimuhandhis@uwp.ac.id), LAW (laily.alfina.wulansari@uwp.ac.id), RU (ruli.utami@itats.ac.id)
Digital Object Identifier 10.32815/jitika.1249
Manuscript submitted 12 Januari 2026; revised 25 April 2026; accepted 5 Mei 2026.
ISSN: 2580-8397(O), 0852-730X(P).

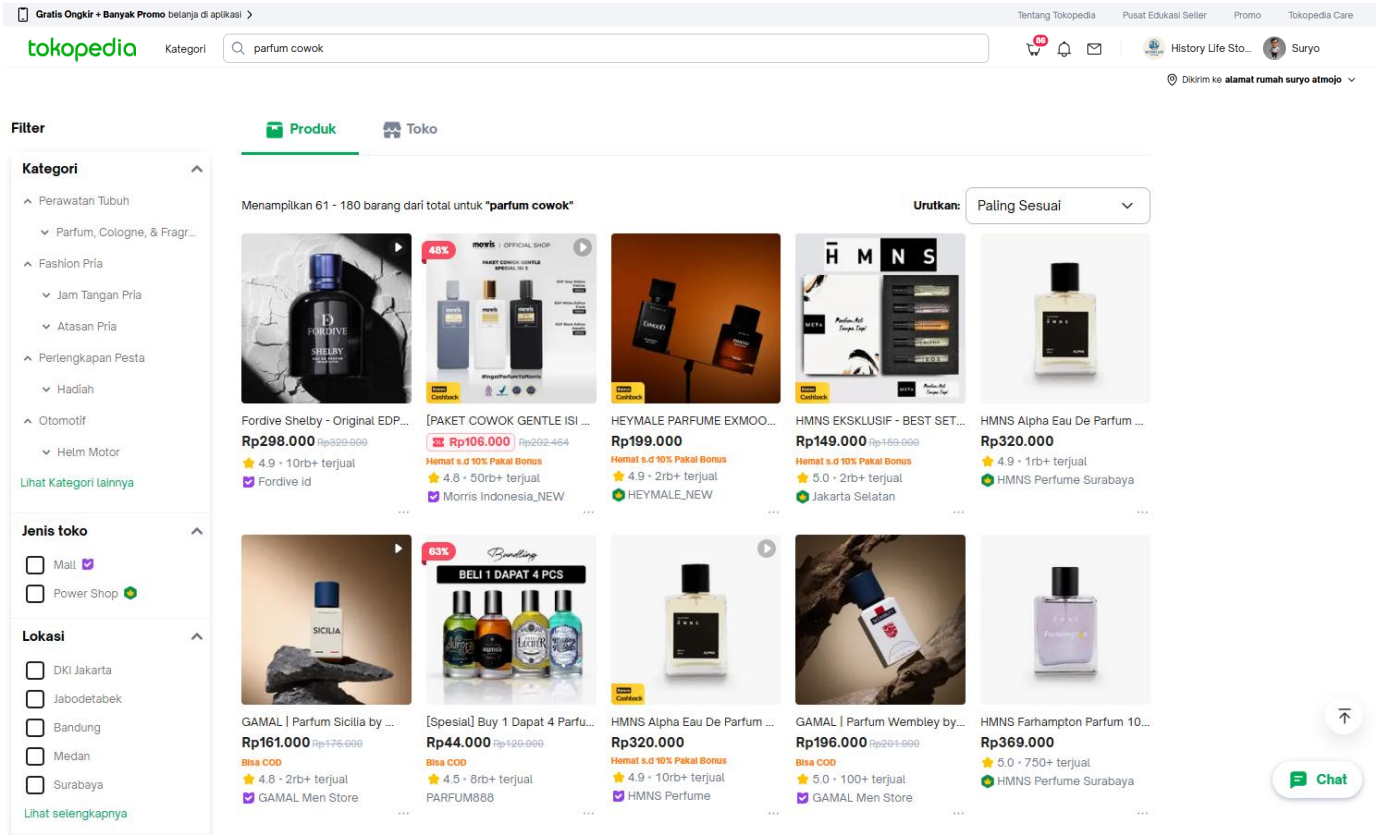

Fig 1. Tokopedia marketplace display showing perfume products

difficulties lies in understanding how competitive product characteristics are shaped by consumer preferences, particularly with respect to gender, price, and popularity. Information such as who the main competitors are, at what price range products sell best, and which market segments remain open, is often not directly available in an easily analyzable format (wulandari et al., 2024; Rahmawati et al., 2025; Fatonah et al., 2024).

The researcher's curiosity stems from the phenomenon that, although transaction and product data are publicly available on marketplaces, their use in market analysis remains minimal. By applying a web scraping approach and the K-Means Clustering algorithm, this study aims to segment perfume products based on attributes of gender, price, and popularity. This segmentation will form product clusters with similar characteristics, thereby assisting business actors or brands in determining product placement strategies, pricing adjustments, and more effective promotional campaigns (Nugroho, 2025; Sabrina & Yasin, 2025; Putri et al., 2025).

Furthermore, this approach supports the advancement of applied informatics, particularly in leveraging data mining and machine learning techniques to address challenges in the digital industry. The use of unsupervised learning methods, such as K-means clustering, enables the discovery of hidden patterns in large datasets. This process can serve as a foundational step toward developing more advanced systems, including recommendation engines and predictive models, which utilize real-world market data to generate insights and improve decision-making accuracy for businesses. (Zhang & Wu, 2024; Legito et al., 2023; Yu, 2024).

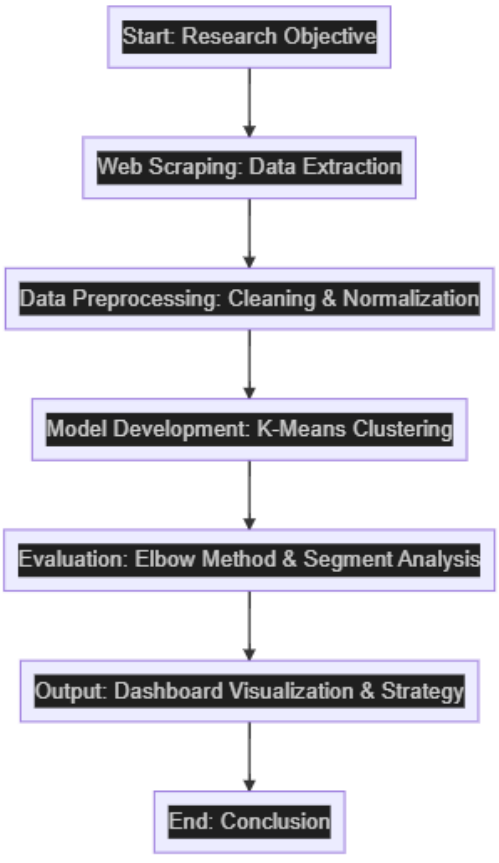

Fig 2. Research Flowchart

Table 1. Research Stages

No	Stages	Main Activities	Output
1	Perfume Product Web Scraping	Scraping perfume products from Tokopedia. Data includes: product name, brand, price, rating, review count, target gender, and sales volume (if available).	Perfume product dataset
2	Data Preprocessing	Price normalization, gender encoding, and missing/duplicate cleaning. • Invalid data will be cleaned. • Categorical attributes like gender will be encoded (0 = male, 1 = female, 2 = unisex). • Numerical attributes will be normalized to be equivalent in the clustering process.	Processed data
3	K-Means Clustering	• The K-Means model will be applied to group products into several clusters. • Determination of cluster count (K) will be done using the Elbow Method and Silhouette Score. • Each cluster will be analyzed based on dominant attributes.	Product segmentation model
4	Interpretation & Strategy Formulation	• Each cluster will be interpreted as a market segment with specific characteristics. • Analysis results are used to formulate recommendations for pricing, positioning, and promotion strategies.	Cluster strategy document
5	Dashboard Visualization	Creating an interactive dashboard. The visualization is designed to help MSMEs better understand their product's market position.	Segmentation dashboard

2. Methods

The research methodology is structured into five systematic stages to ensure a rigorous data-driven analysis. The process begins with automated data acquisition and concludes with interactive visualization, following a standard data mining workflow. This structured approach allows for reproducibility and clear interpretation of the perfume market segments. The research flow is visualized in Fig. 2 below.

2.1. Research design and approach

This study uses a quantitative exploratory approach with data mining methods based on unsupervised learning, specifically using the K-Means Clustering algorithm. This approach was chosen for its ability to uncover hidden patterns in numerical and categorical attributes, such as price, target gender, and product popularity in the marketplace (Rajapandian et al., 2025; Zhao et al., 2024). The main objective is to systematically form product segmentation for perfume products to support data-driven decision-making by MSME actors (Sakinah & Awaliyah, 2025).

2.2. Data sources and collection period

Primary data will be collected through web scraping from the Tokopedia and Shopee marketplaces, spanning April to June 2025. Scraping focuses on “Men’s Perfume” and “Women’s Perfume” categories, with filtering based on MSME seller criteria, including:

- Not official brands or verified stores,
- Limited product count (less than 100 items),
- Total sales not too high (maximum 10,000 sales per product) (Pranata et al., 2023).

2.3. Sampling technique

The study employs purposive sampling, selecting data based on specific criteria aligned with the research objectives. This method is widely applied in digital data-driven studies to ensure that the collected dataset accurately represents the research object. In this context, the data focus on perfume products offered by MSME actors, enabling more relevant and targeted analysis. (Chrisinta & Simarmata, 2024; Chairunnita et al., 2025; Nurhayati et al., 205).

The study targets a minimum dataset of 1,000 perfume products collected from various MSME shops, ensuring a balanced distribution across men’s, women’s, and unisex categories. Primary data is obtained through a web scraping process from marketplace platforms such as Tokopedia and Shopee. The collected data is then cleaned, organized, and processed through a series of systematic analytical stages. These steps aim to generate meaningful product segmentation by grouping similar items into several distinct clusters based on selected attributes. (Mubarok et al., 2025; Aldi & Fatah, 2025).

3. Results

3.1. Data collection

The research data was gathered through web scraping on the Tokopedia e-commerce platform using Node.js and the Playwright framework. The scraping process was executed programmatically to handle infinite scrolling and dynamically

Table 2. SSE Values for Elbow Method

Number of Clusters (k)	Sum of Squared Errors (SSE)
1	45.70
2	30.52
3	13.59
4	11.17
5	8.48
6	7.46

Table 3. Statistical Analysis of Clusters

Cluster	Market Segment	Total Products	Average Price	Average Rating
0	Budget / Mass Market	42	Rp 69,464	4.1
1	Mid-Range Market	508	Rp 191,969	3.5
2	Premium / Niche Market	443	Rp 226,890	4.5

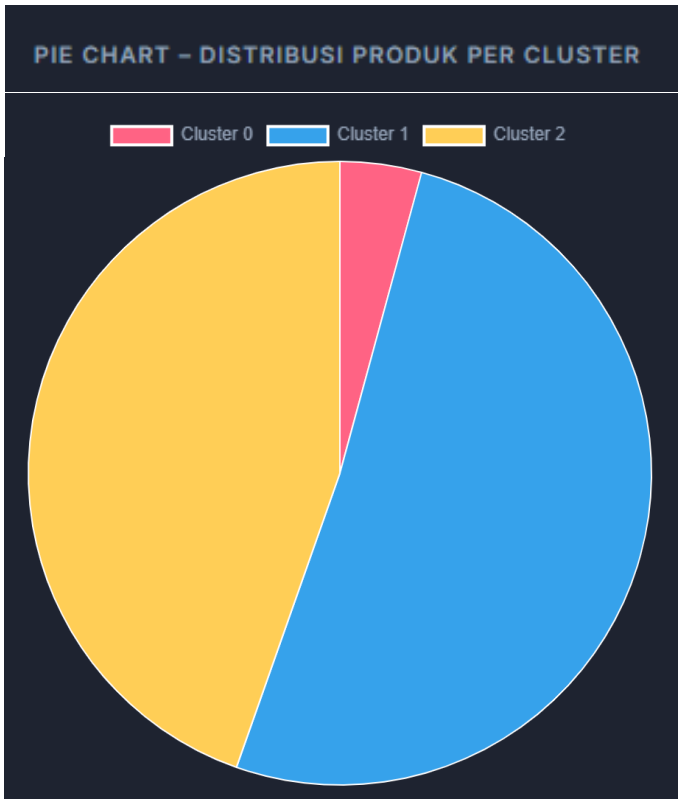

Fig 3. K-Means Clustering Result Visualization

load items based on various search queries (such as "parfum pria", "parfum wanita", "parfum unisex", "parfum branded", and others). The raw dataset successfully collected consisted of 1,000 unique perfume products. The data attributes obtained include Product Name, Price, Sold Quantity, Rating, Shop Location, and Shop Name.

3.2. Data preprocessing

Before the clustering process, the raw data underwent strict pre-processing to ensure the mathematical validity of the K-Means algorithm and to address issues of outlier dominance. This stage involved:

1. Price and Sales Cleaning: Removing non-numeric characters and converting text formats (e.g., "10rb+ terjual" into 10000).
2. Outlier Filtering: To prevent the algorithm from merely separating valid data from noise, products with zero sales (dead listings) and extreme pricing anomalies (price > Rp 1,500,000 with sales < 5) were entirely removed from the dataset. From 1,000 raw products, 7 outliers were discarded, leaving 993 valid products.
3. Missing Values Handling: Products without a specific rating were assigned an interpolated rating to ensure the dimension remains active.
4. Min-Max Normalization: Features (Price, Sold Quantity, and Rating) were normalized using the Min-Max scaling method to ensure all three dimensions contribute equally to the distance calculation.

3.3. Determination of optimal cluster count (elbow method)

To determine the most optimal number of clusters (k), the Elbow Method was used by calculating the Sum of Squared Errors (SSE) for values of k=1 to k=10 on the 3-dimensional normalized dataset.

Based on the SSE values in Table 2, a drastic decline occurs from k=1 to k=2, and then again from k=2 to k=3. After k=3, the decline in SSE begins to level off (plateau), indicating that adding further clusters does not provide a significant reduction in intra-cluster variance. Therefore, it was mathematically determined that the optimal number of clusters is 3.

3.4. Clustering results (K-means algorithm)

The K-Means algorithm was executed with k=3 exclusively on the 993 valid data points. To ensure meaningful and actionable business tiers, the resulting clusters were sorted and labeled based on their actual average prices. The statistical breakdown of the product grouping results is presented in Table 3.

3.5. Product segmentation (clustering result)

Using k=3, the K-Means algorithm successfully grouped product data into three distinct, actionable market characteristic segments:

- Cluster 0 (Budget / Mass Market): This segment represents the lowest price tier (~Rp 69,464). Interestingly, consumer satisfaction is relatively high (4.1). This cluster represents highly affordable daily perfumes for which consumers have

realistic expectations given the low price.

- Cluster 1 (Mid-Range Market): This is the most crowded and competitive segment (508 products) with an average price of ~Rp 191,969. Consumers in this segment appear highly critical, as reflected in the lowest average satisfaction rating (3.5).
- Cluster 2 (Premium / Niche Market): This segment contains higher-priced items (~Rp 226,890). Despite the higher prices, this cluster achieves the highest consumer satisfaction (4.5), indicating a loyal customer base that prioritizes quality, scent longevity, and brand authenticity over price.

4. Discussion

The segmentation analysis reveals that the perfume market on Tokopedia is highly competitive, particularly in the Mid-Range and Premium tiers. Unlike previous uncleaned models that merely separated noise from valid data, rigorous filtering for zero-sales products and pricing anomalies has enabled the algorithm to accurately segment the market into three actionable business tiers for MSMEs.

The Mid-Range segment (Cluster 1) is the most densely populated, indicating fierce competition. The low average rating (3.5) suggests that consumers in this price bracket often experience a gap between expectations and product quality. For MSMEs entering this segment, improving product quality, packaging, and customer service is crucial to standing out.

Conversely, the Premium segment (Cluster 2) demonstrates that higher pricing can be justified by superior quality, as evidenced by its high rating (4.5). Consumers in this segment are less price-sensitive and more value-driven. Meanwhile, the Budget segment (Cluster 0) requires a high-volume, turnover-based strategy, where MSMEs must rely on mass sales to offset lower per-unit profit margins.

The successful integration of the "Rating" dimension alongside Price and Sales Volume provides a holistic view of the market, underscoring the need for pricing strategies to align with consumer satisfaction levels. Furthermore, the decision to drop the "Target Gender" variable was justified, as preliminary observations indicated that modern perfume marketing often employs unisex branding, making gender less distinct as a primary clustering feature than price and performance metrics.

5. Conclusion and future research

Based on the results of the research and discussion above, several conclusions can be drawn, First, the application of Web Scraping techniques is effective for collecting product data in real-time from e-commerce platforms, successfully extracting 1,000 unique data points for robust analysis. Second, the complete removal of zero-sales outliers and extreme price anomalies is a mandatory pre-processing step. This prevents mathematical errors caused by outlier dominance and ensures the K-Means algorithm segments the actual market rather than just isolating statistical noise. Lastly, the K-Means Clustering algorithm, incorporating Price, Sales, and Rating dimensions, successfully polarized the market into three actionable tiers: Budget, Mid-Range, and Premium. Each tier exhibits distinct consumer satisfaction trends, providing clear strategic guidance for MSMEs on pricing and product positioning.

Data availability

All data produced or examined during this study are presented in this paper.

Declaration of competing interest

The authors affirm that they have no known financial conflicts of interest or personal connections that could be perceived as affecting the research presented in this article.

Authors' contributions

All authors have reviewed and approved the final manuscript. All authors participated in the study design, writing, and manuscript revision. SA drafted the initial manuscript and performed the analysis, SD and IM revised the methodology, LAW collected the data, and RU supervised the research. All authors have reviewed and approved the final manuscript.

Declaration of generative AI in scientific writing

During the preparation of this work, the author(s) used Google Gemini in order to assist in structuring the manuscript and formatting the references according to APA 7th guidelines. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Aldi, M. A., & Fatah, Z. (2025). Implementasi K-means clustering dalam pengelompokan data kunjungan wisatawan asing di Indonesia. Jurnal Ilmiah Multidisiplin Ilmu, 2(1), 13–19. https://doi.org/10.69714/3hhfj353

Alfi Rahmawati, Tasya Kamila Hamdani, & Wahyu Budi Priatna. (2025). Ecommerce Adoption and MSME Business Performance in Indonesia: Systematic Literature Review. Journal Scientific of Mandalika (JSM) E-ISSN 2745-5955 | P-ISSN 2809-0543, 6(9), 3512-3520. https://doi.org/10.36312/10.36312/vol6iss9pp3512-3520

Analisis performa algoritma K-means dan DBSCAN dalam segmentasi pelanggan dengan pendekatan model RFM. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 8(7). https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/13962

Chairunnita, C., Handayanto, A., & Dewanto, F. (2025). Penerapan Algoritma K-Means Clustering dalam Analisis Pengelompokan Produk Toko Oleh-Oleh Berdasarkan Data Penjualan. Journal of Information System Research (JOSH), 6(4), 1816-1825. https://doi.org/10.47065/josh.v6i4.7832

Chrisinta, D., & Simarmata, J. E. (2024). Eksplorasi teknik web scraping pada data mining: Pendekatan pencarian data berbasis Python. Faktor Exacta, 17(1). http://dx.doi.org/10.30998/faktorexacta.v17i1.22393

Dewi, R. K. (2025). The influence of e-commerce on consumer buying behavior in rural Indonesia. Aegaeum Journal, 13(2), 45–56.

Fatonah, S. ., Hidayat, A. ., & Mirnayani. (2024). Exploring the impact of consumer behavior and innovation orientation on business strategy and firm performance: Insights from MSMEs in Indonesia . International Journal of Management and Sustainability, 13(4), 818–831. https://doi.org/10.18488/11.v13i4.3900

Laila Ali Putri, Mazayah Tsaqofah, Dea Syahfira Hasibuan, Hasti Fadillah, Maria Ulfa, & Mhd.Furqan. (2025). Application of K-Means Clustering Algorithm for E-Commerce Data Analysis. Journal of Artificial Intelligence and Engineering Applications

(JAIEA), 4(3), 2364–2367. https://doi.org/10.59934/jaiea.v4i3.1170
Legito, L., Wattimena, F. Y., Rofi'i, Y. U., & Munawir, M. (2023). E-commerce product recommendation system using case-based reasoning (CBR) and K-means clustering. International Journal Software Engineering and Computer Science, 3(2), 162–173.
Mubarok, R., Syahputra, A. A., Permana, A. T., Sholiah, L., & Tarwoto. (2025). Implementasi Data Mining untuk Clustering Lowongan Pekerjaan Menggunakan Metode Algoritma K-Means. Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi), 9(2), 703-712. https://doi.org/10.35870/jtik.v9i2.3438
Nugroho, A. Y. (2025). Application of the KMeans Clustering Algorithm in E-Commerce Transaction Pattern Analysis: Application of the KMeans Clustering Algorithm in E-Commerce Transaction Pattern Analysis. JURNAL ILMIAH MATEMATIKA DAN TERAPAN, 22(1), 21 - 30. https://doi.org/10.22487/2540766X.2025.v22.i1.17884
Nurhayati, I. P., Helmayana, Tiani, A., Maruenci, K., Harahap, Y. S., & Fansyuri, M. (2025). Utilization of Data Mining for Credit Card Customer Segmentation Using the K-Means Method. Journal of Information Technology and Informatics Engineering, 1(1), 62-66. https://journal.jci.co.id/jitie/article/view/162
P. Rajapandian, A. Karunamurthy, V. Vasanth, & M. Meganathan. (2025). E-Commerce Customer Segmentation: A Clustering Approach in A Web-Based Platform. Journal of Engineering Technology and Applied Physics, 7(1), 71–79. https://doi.org/10.33093/jetap.2025.7.1.12
Pranata, F. M., Wijoyo, S. H., & Setiawan, N. Y. (2024).
Sabrina, A. E., & Yasin, M. (2025). Clustering data konsumen e-commerce menggunakan algoritma K-Means dan dataset Kaggle. Journal of Computer Science and Informatics Engineering, 4(2), 96–109. https://doi.org/10.55537/cosie.v4i2.1123
Sakinah, A., & Awaliyah, D. S. (2025). Optimization of e-commerce consumer segmentation based on K-means clustering and machine learning. Journal of Mathematics, Computations and Statistics, 8(2). https://doi.org/10.35580/jmathcos.v8i2.9548
Wardhana, A., Pradana, M., Shabira, H., Buana, D. M. A., Nugraha, D. W., & Sandi, K. (2021). The influence of consumer behavior on purchasing decision process of Tokopedia e-commerce customers in Indonesia. In Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management (pp.—). IEOM Society International. https://doi.org/10.46254/AN11.2021099
Wulandari, A., Purnomo, & Rustandy, A. . (2024). The Role of Market Segmentation and Competitive Positioning in Increasing Brand Awareness in the Indonesian Retail Industry. Ilomata International Journal of Management, 5(4), 1390–1412. https://doi.org/10.61194/ijjm.v5i4.1293
Yu, L. (2024). The application of K-means clustering algorithm in the evaluation of e-commerce websites. Journal of Electrical Systems, 20(6s). https://doi.org/10.52783/jes.2738
Zhang, W., & Wu, Z. (2024). E-commerce recommender system based on improved K-means commodity information management model. Heliyon, 10(9), e29045. https://doi.org/10.1016/j.heliyon.2024.e29045
Zhao, Y., et al. (2024). Research on e-commerce user segmentation and customized marketing strategy based on cluster analysis. Applied Mathematics and Nonlinear Sciences, 9(1). https://doi.org/10.2478/amns-2024-2668

Suryo Atmojo was born in Falabisahaya Dofa, North Maluku, Indonesia, on January 9, 1989. He received a Diploma degree in management of information technology from STIKI Malang, Malang, Indonesia, in 2009. He then earned his Bachelor's degree in informatics engineering from STIKI Malang, Malang, Indonesia, in 2011, and his Master's degree in informatics engineering from ITS Surabaya, Surabaya, Indonesia, in 2015. His major field of study was informatics engineering.

Suzana Dewi was born in Banjarbaru, South Kalimantan, on April 17, 1983. She completed her Diploma (D3) in Information Systems and Technology at STIKOM Surabaya and later earned her Bachelor's degree in 2007. During her studies, she gained work experience at several companies in Surabaya, including PT. Bumi Jasa and PT. Jatim Tiga Manunggal. In 2011, she began her career as a lecturer. She pursued her Master's degree at Dian Nuswantoro University, Semarang, and completed it in 2015.

She has taught at several institutions, including STMIK Pranata, Universitas Islam Kalimantan (UNISKA), STMIK Indonesia, and the Health Polytechnic in Banjarbaru. From 2015 to 2025, she served as a permanent lecturer at Wijaya Putra University, Surabaya. Since 2026, she has been teaching at Cahaya Bangsa University in Banjar Regency. In addition to her academic career, she has managed a souvenir shop in Banjarbaru since 2025.

Isnaini Muhandhis was born in Malang on October 16, 1988, Her completed undergraduate studies in Informatics Engineering at STMIK Pradnya Paramita Malang in 2012. She received her Master's degree in Informatics Engineering from the Institut Teknologi Sepuluh Nopember Surabaya in 2015, with a specialization in Information Systems.

Her primary research interest lies in Artificial Intelligence, and she has consistently published scholarly works since 2016. She currently serves as a lecturer in the Informatics Engineering Department at Universitas Wijaya Putra, a position she has held since 2015.

Laily Alfina Wulansari, commonly known as Wulan, is a graduate in Informatics Engineering with a strong enthusiasm for continuous learning and professional growth. She enjoys engaging in meaningful work and actively seeks opportunities to build new connections and expand her professional network. Wulan has a particular interest in data analysis, especially using Python to extract insights and solve complex problems. She is passionate about transforming data into valuable information and continuously developing her analytical and technical skills to support data-driven decision-making.

Ruli Utami was born in Lamongan, Indonesia, on March 24, 1984. She received her Bachelor's degree in information systems from Institut Teknologi Adhi Tama Surabaya (ITATS), Surabaya, Indonesia. She then earned her Master's degree in informatics engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2015. Her major fields of study were information systems and informatics engineering.

She is currently working as a Lecturer at Institut Teknologi Adhi Tama Surabaya (ITATS), Surabaya, Indonesia. She is actively involved in research in the field of information systems, particularly in the development and implementation of information technology to support organizational and business needs.