INVOTEK

Jurnal Inovasi Vokasional dan Teknologi

http://invotek.ppj.unp.ac.id/index.php/invotek

Vol. 24, No. 2, June 2024

ISSN: 1411 – 3411 (p)
ISSN: 2549 – 9815 (e)

Sales Segmentation Analysis of Tobacco Products Using the K-Means Clustering Method

Yance Sonatha[1] *, Aldo Erianda[1], Redhatul Fitri[1]

[1] Department of Informatics Technology, Politeknik Negeri Padang
Jl. Limau Manis Padang, West Sumatera, Indonesia-25164

*Corresponding author: yance@pnp.ac.id

Doi: https://doi.org/10.24036/invotek.v24i2.1221

Abstract

Technological advancements have encouraged businesses to optimize data utilization, including in sales analysis. This study analyzes sales transaction data of tobacco products at Tobacco Shop Taste using the K-Means Clustering method. By implementing the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, the sales data were categorized into three groups: highly sold, moderately sold, and less sold. These clustering results support stock management, marketing strategies, and data-driven decision-making. A web-based system was developed, providing real-time monitoring of analysis results, which distinguishes this study from existing solutions by enabling store management to promptly respond to sales trends. This study significantly contributes to the application of data mining technology in the tobacco retail sector, despite being limited to a single store and basic variables. Future development opportunities include integrating broader datasets and analyzing external variables to enhance the accuracy and relevance of the findings.

Keywords: K-Means Clustering, Sales Analysis, Product Segmentation, Data Mining, Stock Management.

1. Introduction

The development of technology has driven business owners to enhance the effectiveness and efficiency of their operations. One technological advancement that businesses urgently need is related to the processing of sales data. Many businesses still rely on traditional methods for their operational processes, such as recording transactions manually in sales ledgers [1]–[3]. As a result, the data from these sales are not fully utilized, leading to inefficiencies. This is the case with Tobacco Shop Taste.

Tobacco Shop Taste is a store located in Lasi, Candung Subdistrict, Agam Regency, West Sumatra. This shop sells various types of tobacco in different flavors, as well as rolling accessories such as cigarette papers, tobacco rolling machines, filters, and other products. The store still uses a manual system for recording sales data, expenses, profits, and so on, in an administrative book. Every day, employees must report the store's income and expenses to the owner by sending photos of the daily administrative record pages. This method is highly ineffective and prone to problems such as data entry errors and data loss. The data is disorganized and not utilized to assess which products are most or least popular. Furthermore, the absence of inventory tracking leads to stock shortages for popular items.

Existing methods for clustering sales data, such as hierarchical clustering and density-based algorithms, often face limitations when applied to dynamic and large-scale datasets like retail sales. Hierarchical clustering, for instance, is computationally intensive and struggles with scalability, making it less practical for stores with varying inventory sizes and daily transaction volumes. Similarly, density-based clustering methods, while effective for identifying arbitrary-shaped clusters, are sensitive to parameter selection and require pre-knowledge of data density distributions, which may not be feasible for retail sales data characterized by irregular patterns. These limitations highlight the need for an efficient, scalable, and robust method like K-Means clustering, which can handle large datasets, adapt to dynamic changes, and provide actionable insights for decision-making.

Received date 02/12/2024, Revised date 03/01/2025, Accepted date 02/03/2025

P-ISSN: 1411-3414
E-ISSN: 2549-9815

In the tobacco industry, analyzing sales patterns is essential to understanding market needs, determining marketing strategies, and optimizing resource allocation [4]. However, this analysis process often faces challenges, especially when dealing with large and varied sales data [5]. One effective technique for uncovering hidden patterns in data is K-Means Clustering, a method frequently used in data mining to group data based on specific characteristics [6]–[8]. By applying this method, the analysis of tobacco sales patterns can be carried out more systematically, providing valuable information for business decision-making.

This study aims to apply the K-Means Clustering method to determine tobacco product sales patterns to facilitate market segmentation, identify sales trends, and provide strategic recommendations based on the clustering results [9]. By using this method, it is hoped that the segmentation of sales data will provide deeper insights into consumer preferences for various types of tobacco products over a certain period [10].

The research questions addressed in this study include: (1) How can K-Means Clustering be applied to group tobacco product sales data? (2) What sales patterns can be identified through this method? (3) How can the results of this clustering help in making better business decisions? By addressing these questions, this research contributes to the literature by demonstrating the potential of K-Means clustering in optimizing inventory and marketing strategies in the tobacco retail sector, thus bridging the gap between theoretical data mining applications and practical business needs.

The novelty of this study lies in the use of K-Means Clustering to analyze sales patterns in the tobacco sector, a topic that has been rarely explored in this context [11]. This research contributes by adapting the K-Means method to uncover varied consumption patterns and factors affecting tobacco product sales, thus providing new insights for industry players in managing marketing strategies and meeting market demand more effectively.

Sales segmentation analysis using the K-Means clustering method can significantly enhance marketing strategies by identifying distinct customer groups based on purchasing behavior [12], [13]. This approach allows for targeted marketing, optimal resource allocation, and overall improvement in sales performance. For instance, the K-Means algorithm processes customer data to identify clusters based on attributes such as purchase frequency and product preferences [14], [15]. Additionally, there are studies in the tobacco retail sector that categorize customers into nine different segments, demonstrating the effectiveness of clustering in understanding diverse customer bases [16].

Clustering can also help businesses understand customer segments, enabling them to tailor marketing strategies to meet the specific needs of each group, thereby improving customer retention and increasing sales [17]. Moreover, clustering methods allow companies to allocate resources more efficiently by focusing on high-value segments to maximize revenue [18].

2. Methodology

The method used in the development of this final project is a standard data mining method known as the Cross-Industry Standard Process for Data Mining, or CRISP-DM for short [19]–[22]. The process flow of CRISP-DM can be seen in Figure 1 below.

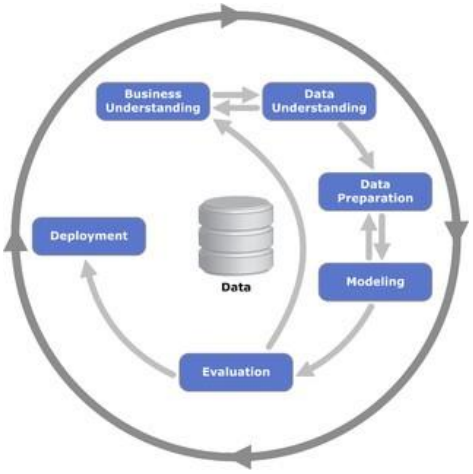


Figure 1. The CRISP-DM Process Flow

P-ISSN: 1411-3414
E-ISSN: 2549-9815

2.1 Business Understanding

In this stage, the business needs and project objectives are identified. The goal is to determine which tobacco products in the Tobacco Shop Taste are the best-selling and least-selling. The main purpose is to support stock management and devise appropriate promotional strategies.

2.2 Data Understanding

This stage aims to collect and understand the data that will be analyzed, specifically the tobacco sales data from the store. The raw data attributes include:

- Product Name: The name of the tobacco product.
- Sales Frequency: The number of times the product was sold during the observed period.
- Total Revenue: The total income generated from the sales of the product.
- Date of Sale: The specific dates on which the products were sold.

These attributes were selected to provide a comprehensive view of sales trends and product performance.

2.3 Data Preparation

This stage involves the processes of data cleaning, selection, and integration to make the data ready for further analysis. The data is organized and processed so that it can be applied to the selected algorithm.

2.3.1 Data Cleaning Process

The sales table is cleaned by removing irrelevant data, eliminating duplicates, correcting erroneous data, and addressing missing values.

2.3.2 Data Selection Process

In this phase, the data collected is filtered to retain only the columns relevant to the clustering process. The objective is to select the most pertinent data and exclude unnecessary information. The attributes used for clustering include product name, sales frequency, and total revenue.

2.3.3 Data Integration

This step merges the sales table and product list into a single table that will be processed. This process generates new attributes, namely "Total J," which represents the total sales of a product over a specific period, and "Total Price J," which represents the total revenue from product sales.

2.4 Modeling

The modeling phase of this study utilizes the K-Means algorithm to cluster the sales data based on products that are best-selling, moderately selling, and least-selling. The detailed process is as follows:

2.4.1 Determining the Number of Centroids (k)

The decision to use three clusters (k = 3) was based on the practical needs of the store to categorize products into clear and actionable groups: best-selling, moderately selling, and least-selling. This categorization aligns with the store's operational priorities, such as inventory planning and promotional strategy development. Additionally, initial exploratory data analysis showed that three clusters effectively captured the main variations in product sales patterns without overcomplicating the results.

2.4.2 Initializing Centroid Centers

The initial centroids are selected based on three values: the highest, average, and lowest values from the "total sales" and "total price" attributes. For example, the first centroid is initialized with the highest value, the second with the average value, and the third with the lowest value to ensure stable clustering results in each iteration. The initial centroid values are displayed in Table 1 below:

Table 1. Initial Centroid Values

	Total J	Total Price
C1	1	4000

P-ISSN: 1411-3414
E-ISSN: 2549-9815

C2	9	68875
C3	119	324000

2.4.3 Calculating the Distance to Centroids

The distance of each data point to the centroids is calculated using the Euclidean distance formula, which measures the proximity between the data point and the cluster center. The Euclidean distance formula used is:

$$D_{(ij)} = \sqrt{(x1i - x1j)^2 + (x2i - x2j)^2 + \cdots + (xki - xkj)^2} \quad (1)$$

where $D_{(i,j)}$ is the distance between data point iii and cluster center j, and xki and xkj are the attribute values of data i and centroid j, respectively.

2.4.4 Recalculating Centroids

After the data is classified, the cluster centers are updated based on the average of the data points within each cluster. This process is repeated until the positions of the cluster centers no longer change or converge.

2.4.5 Clustering Results

Once the iterations are complete, the data is divided into three groups: best-selling, moderately selling, and least-selling products. These results provide insights into the sales patterns of products, aiding the store in managing stock and planning product promotions.

2.5 Evaluation

Evaluation is performed to ensure that the clustering results meet the analysis objectives. The evaluation results include the grouping of products, which will help store management make more effective decisions.

2.6 Deployment

The final stage involves the implementation of a web-based system that allows the store owner to monitor sales data and clustering results in real-time, facilitating strategic decision-making.

3. Result and Discussion

3.1 System Interface Display

The system is developed using a web-based platform. Figure 2 below illustrates the initial interface of the developed system, which features a login page.

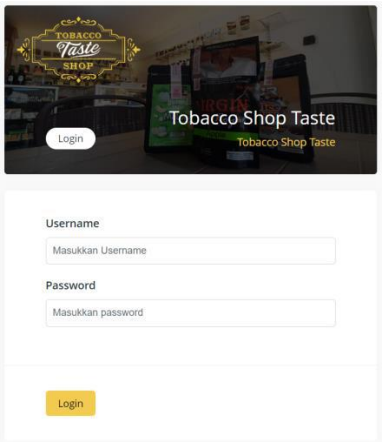


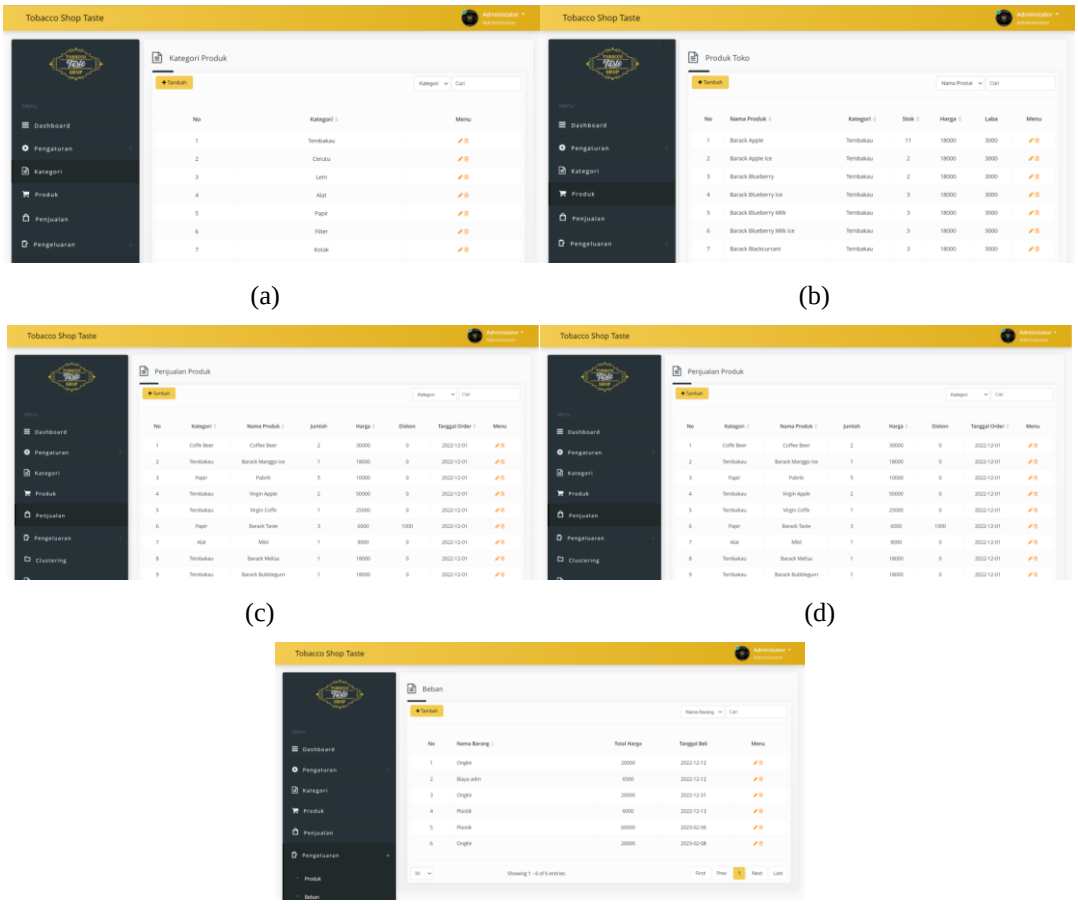
Figure 2. Login Page

After successfully logging in, users are directed to the system's dashboard page. The dashboard displays a monthly sales chart, a clustering chart, and a list of top products, which represent the best-selling items based on clustering data. This is illustrated in Figure 3 below:

P-ISSN: 1411-3414
E-ISSN: 2549-9815

Figure 3. Dashboard Page

The system is designed with two main access roles: store owners and system administrators. Users can access various menus provided by the system, including categories, products, sales, and expenses (which consist of purchases and operational costs), as shown in Figure 4 below:



(a) (b)

(c) (d)

(e)

Figure 4. The Various Menu Options Available on the System Dashboard Include: a) Category Page, b) Product Page, c) Sales Page, d) Purchase Page, e) Expense Page

3.2 Clustering Page Interface

This section focuses on testing the clustering results of sales data using the K-Means algorithm. Figure 5 shows two date fields: the first field specifies the starting date for the sales data to be processed, while the second field defines the end date for the data selection. The system imposes no restrictions on clustering results based on month or year; it even allows clustering to be performed for a single day.

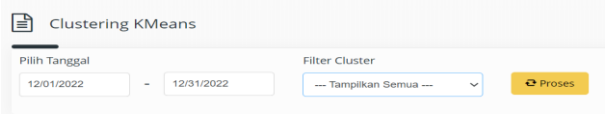
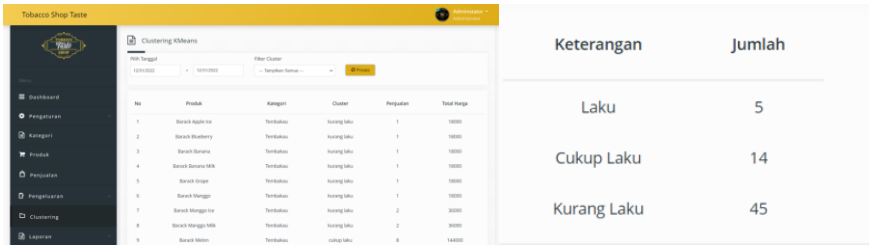


Figure 5. Clustering Page Features

P-ISSN: 1411-3414
E-ISSN: 2549-9815

The clustering results for sales data in December 2022 are shown in Figures 6. These figures indicate that 64 types of products were sold during December 2022. Among them, 5 products were classified as highly sold, 14 products as moderately sold, and 45 products as less sold.



(a) (b)

Figure 6. Display of Clustering Results and Summary

The data displayed in the clustering results table can be clicked to open a new page, referred to as the clustering detail page. As shown in Figure 7, this page provides information about the products, their monthly grouping, and a monthly sales chart for each product.

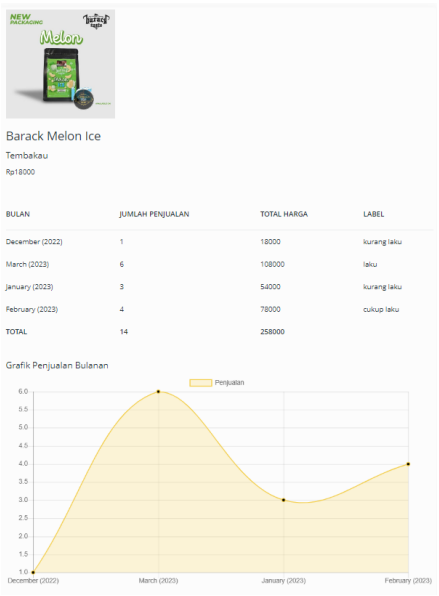


Figure 7. Product Clustering Details

In Figure 7, a comparison of the monthly product clustering results is shown. Each month, the centroid value varies depending on the total sales and total price attributes within a particular data group. As a result, it is possible for a product to have higher total sales than the previous month but still fall into a lower group.

The clustering results from the system's processing are displayed on the dashboard. The visualization of these results, derived from the clustering process available in the system's main menu, is shown in Figure 8 below:

Figure 8. Clustering Diagram in the System

The system's clustering results provide significant insights into optimizing marketing strategies for the store. By categorizing products into three groups—highly sold, moderately sold, and less sold—the store can focus its resources effectively. Best-selling products can be prioritized for promotional campaigns to sustain or further boost their sales. Marketing efforts such as loyalty programs or discounts can target moderately sold products to increase their attractiveness. For less popular items, the store could either re-evaluate their stocking decisions or explore niche marketing techniques to reach specific customer segments.

Additionally, the real-time monitoring system enables quick adaptation to emerging sales trends. The dashboard visualizations, such as monthly sales and clustering charts, allow store owners to make informed, data-driven decisions. However, potential biases in the clustering results must be considered. Seasonal variations could affect the demand for certain products, leading to fluctuating sales patterns that might not reflect long-term trends. For example, holiday-specific items might temporarily fall into the highly sold category. Furthermore, incomplete or inconsistent historical data could impact the clustering's accuracy. To mitigate these issues, future iterations of the system could incorporate external data, such as local events or seasonal demand patterns, and validate the clustering results using broader datasets. By addressing these limitations and leveraging the clustering results, the system offers a robust tool for enhancing inventory management and tailoring marketing strategies to customer needs.

4. Conclusion

This study successfully demonstrates that the K-Means Clustering method can be effectively used to analyze tobacco product sales patterns. The developed system facilitates product segmentation into three categories: best-sellers, moderately sold, and slow-sellers. The analysis results, such as those from December 2022, provide valuable insights for inventory management and store promotion strategies. The web-based system also allows for real-time sales data monitoring, enhancing store operational efficiency. However, this study has limitations, such as the scope of data being restricted to a single store and the use of basic variables without considering external factors, such as market trends or seasonal sales. Future research can expand by incorporating data from multiple stores and regions to uncover more comprehensive patterns. Additionally, the inclusion of external variables and the integration of other algorithms could improve the accuracy of the analysis. With further development, this system has the potential to make a significant contribution to the data-driven business sector.

References

[1] S. D. Golden *et al.*, "Trends in the Number and Type of Tobacco Product Retailers, United States, 2000–2017," *Nicotine Tob. Res.*, vol. 24, no. 1, pp. 77–84, Jan. 2022, doi: 10.1093/ntr/ntab150.

[2] Y. Bi *et al.*, "Tobacco chemical value quantifying method based on near-infrared spectrum wave number K-means clustering." 2015.

[3] W. S. Utami, N. Pratiwi, and F. Muhammad, "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Clustering Perokok Usia Lebih dari 15 Tahun," *Bull. Inf. Technol.*, vol. 4, no. 4, pp. 501–507, 2023, doi: https://doi.org/10.47065/bit.v4i4.1067.

[4] Z. Dzalilov and A. Bagirov, "Cluster analysis of a tobacco control data set," *Int. J. Lean Think.*, vol. 1, no. 2, pp. 40–50, 2010.

[5] H. Fan, C. Lu, and H. Chen, "Exploring the spatial agglomeration characteristics of cigarette brand sales in Guizhou province, China," in *2017 6th International Conference on Agro-Geoinformatics*, 2017, pp. 1–4. doi: 10.1109/Agro-Geoinformatics.2017.8047048.

[6] S. N. Wahyuni, N. N. Khanom, and Y. Astuti, "K-Means Algorithm Analysis for Election Cluster Prediction," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 1–6, 2023, doi: http://dx.doi.org/10.30630/joiv.7.1.1107.

[7] R. F. Ramadhan, S. H. Wijoyo, and M. C. Saputra, "Penerapan Metode K-Means Clustering pada Ulasan Perumahan PT XYZ di Google Maps untuk Formulasi Strategi Bisnis dengan Analisis SWOT," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, pp. 2879–2888, 2023.

[8] M. Ridzki, I. Hadijah, M. Mukidin, A. Azzahra, and A. Nurjanah, "K-Means algorithm method

for clustering best-selling product data at XYZ grocery stores," *Int. J. Soc. Serv. Res.*, vol. 3, no. 12, pp. 3354–3367, 2023, doi: https://doi.org/10.46799/ijssr.v3i12.652.

[9] Z. Potton and B. Baharuddin, "MARKET SEGMENTATION ANALYSIS TO INCREASE THE EFFECTIVENESS OF MARKETING STRATEGIES," *IJMA (Indonesian J. Manag. Accounting)*, vol. 5, no. 1, pp. 242–248, 2024, doi: https://doi.org/10.14421/EkBis.2022.6.1.1555.

[10] J. Laurenso, D. Jiustian, F. Fernando, V. Suhandi, and T. H. Rochadiani, "Implementation of K-Means, Hierarchical, and BIRCH Clustering Algorithms to Determine Marketing Targets for Vape Sales in Indonesia," *J. Appl. Informatics Comput.*, vol. 8, no. 1, pp. 62–70, 2024, doi: https://doi.org/10.30871/jaic.v8i1.4871.

[11] A. Y. N. Harahap, R. E. Sari, H. Gunawan, and A. Buyung, "Evaluation of Product Sales Data Using Clustering Method and Hierarchical Divisive Clustering at PT. AYN," 2024, doi: https://doi.org/10.55927/marcopolo.v2i7.10442.

[12] G. Nathiya, S. C. Punitha, and M. Punithavalli, "An analytical study on behavior of clusters using k means, em and k* means algorithm," *arXiv Prepr. arXiv1004.1743*, 2010.

[13] W. A. Barbakh, Y. Wu, C. Fyfe, W. A. Barbakh, Y. Wu, and C. Fyfe, "Review of clustering algorithms," *Non-standard Param. Adapt. Explor. data Anal.*, pp. 7–28, 2009, doi: https://doi.org/10.1007/978-3-642-04005-4_2.

[14] B. I. Nugroho, A. Rafhina, P. S. Ananda, and G. Gunawan, "Customer segmentation in sales transaction data using k-means clustering algorithm," *J. Intell. Decis. Support Syst.*, vol. 7, no. 2, pp. 130–136, 2024, doi: https://doi.org/10.35335/idss.v7i2.236.

[15] Yosia and B. Siregar, "Comparative Analysis of K-Means and K-Medoids Algorithms for Product Sales Clustering and Customer," *J. Math. Comput. Stat.*, vol. 7, no. 2, pp. 360–370, 2024, doi: https://doi.org/10.35580/jmathcos.v7i2.4053.

[16] H. Han and J. Zhang, "Application of AHP and clustering, discriminant analysis in categorization of cigarette retailers," 2014.

[17] V. Prasad and T. Srikanth, "A survey on clustering algorithms and their constraints," *Int. J. Intell. Syst. Appl. Eng.*, no. 11, pp. 165–179, 2023.

[18] Mrs. J. Sirisha, V. Lakshmi Prathyusha, P. Naga Anupriya, M. Suma Sri, and P. Naga Hema, "Customer Segmentation using K-Means Clustering," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 2, no. 3, pp. 170–175, 2022, doi: 10.48175/ijarsct-7618.

[19] D. B. Saputra, V. Atina, and F. E. Nastiti, "PENERAPAN MODEL CRISP-DM PADA PREDIKSI NASABAH KREDIT MENGGUNAKAN ALGORITMA RANDOM FOREST," *Idealis Indones. J. Inf. Syst.*, vol. 7, no. 2, pp. 240–247, 2024, doi: https://doi.org/10.36080/idealis.v7i2.3244.

[20] N. Doede, P. Merkel, M. Kriwall, M. Stonis, and B.-A. Behrens, "Implementation of an intelligent process monitoring system for screw presses using the CRISP-DM standard," *Prod. Eng.*, pp. 1–12, 2024, doi: https://doi.org/10.1007/s11740-024-01298-8.

[21] R. H. Bemthuis, R. R. Govers, and A. Asadi, "A CRISP-DM-based Methodology for Assessing Agent-based Simulation Models using Process Mining," *arXiv Prepr. arXiv2404.01114*, 2024, doi: https://doi.org/10.48550/arXiv.2404.01114.

[22] Y. A. Singgalen, "Penerapan CRISP-DM dalam Klasifikasi Sentimen dan Analisis Perilaku Pembelian Layanan Akomodasi Hotel Berbasis Algoritma Decision Tree (DT)," *J. Sist. Komput. dan Inform. Hal*, vol. 237, p. 248, 2023, doi: 10.30865/json.v5i2.7081.