Journal of Artificial Intelligence and Digital Economy
Homepage : https://journal.antispublisher.id/index.php/JAIDE
Email : admin@antispublisher.com

e-ISSN : 3032-1077
JAIDE, Vol. 2, No. 12, December 2025
Page 838-844

Application of Data Mining to Predict Distro Clothing Sales Using the K-Means Clustering Method

Shafa Arrizqa Az Zahroh[1], Nuril Lutvi Azizah[2], Novia Ariyanti[3], Irwan Alnarus Kautsar[4]
[1,2,3,4]Muhammadiyah University of Sidoarjo, Indonesia

DOI : https://doi.org/10.61796/jaide.v2i12.1554

Sections Info

Article history:
Submitted: September 30, 2025
Final Revised: October 15, 2025
Accepted: October 25, 2025
Published: November 05, 2025

Keywords:
Data mining
K-Means Clustering
Sales Prediction
Customer Segmentation

ABSTRACT

Objective: *The research aims to classify distro clothing products at Aldi Store according to sales levels to support more effective inventory and marketing strategies.* ***Method:*** *Data processing was conducted using Google Colaboratory, applying the K-Means Clustering algorithm combined with evaluation metrics including the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index to determine the optimal cluster structure.* ***Results:*** *The analysis shows that K-Means successfully groups sales patterns with strong cluster performance, indicated by a Silhouette Coefficient of 0.576, a Calinski-Harabasz Index of 19.125, and a Davies-Bouldin Index of 0.308, reflecting high cohesion and clear separation among clusters.* ***Novelty:*** *This study integrates multiple validity indices within a practical retail context, demonstrating a robust clustering approach that enhances customer segmentation accuracy and provides actionable insights for strategic decision-making in product management.*

INTRODUCTION

The rapid development of technology in the modern era has shown a major transformation for a number of areas of people's lives, including the business and trade sectors [1]. Information technology has become an important component to support business processes, from inventory management to marketing strategies [2]. The use of technology is no longer limited to data collection, but also to data analysis to gain useful insights for decision making. In the increasingly competitive world of trade, data utilization is very important, especially for industries engaged in selling goods, such as distro clothing [3].

The distro clothing industry, known for its creativity and design innovation, faces significant challenges in maintaining its competitiveness [4]. Increasingly fierce competition demands that entrepreneurs continually adapt and seek more effective ways to manage their businesses [5]. One important aspect of this industry is the ability to meet consumer needs in a timely and efficient manner. To achieve this goal, entrepreneurs need to understand sales patterns and market trends so they can take appropriate strategic steps in stock management and product marketing [6].

Based on a literature review, previous studies have shown that sales data analysis can have a significant impact on improving the efficiency and effectiveness of business management [7]. For example, the Apriori algorithm has been used to analyze high-frequency item combination patterns, which is useful for inventory decision-making [8].

However, these studies still have limitations, such as difficulty in obtaining up-to-date data and limited data utilization. These weaknesses indicate an opportunity to develop new approaches that are more suited to the needs of specific industries, such as clothing distribution [9].

Data mining, as a branch of artificial intelligence, offers various techniques for extracting patterns and important information from large amounts of data [10]. Data mining enables entrepreneurs to analyze historical data and identify relevant patterns, such as seasonal trends, consumer preferences, or the most popular product categories. In the context of clothing distribution, data mining can be used to optimize inventory management, forecast demand, and develop more effective marketing strategies [11].

K-Means Clustering is a data mining method with significant potential. This algorithm is used to classify data into several groups based on similar characteristics [12]. Using K-Means Clustering, products in a distro store can be grouped based on similar sales patterns, such as high-selling, moderately selling, and low-selling products. This information can be used as a basis for predicting future product demand and optimizing inventory strategies. Furthermore, this method allows stores to identify seasonal trends or specific sales patterns that are not readily apparent from raw data.

This research is motivated by a weakness in previous research, namely the lack of implementation of a K-Means Clustering-based decision support system specifically applied to the distro clothing industry. This research will focus on sales data from Aldi Stores, which includes variables such as transaction date, payment status, quantity of items, unit price, and total price. By applying the K-Means Clustering method, this study aims to analyze this data to identify product groups based on their sales levels. The results of this analysis are expected to assist Aldi Store in developing more effective inventory and marketing strategies, thereby increasing sales and increasing store competitiveness.

Therefore, this research seeks to address the need for a data mining-based decision support system in the context of distro clothing sales.

RESEARCH METHOD

The research was conducted at the Aldi Baju Distro Store located in the Mega Asri Larangan Housing Complex, Sidoarjo Regency, and was conducted on June 1, 2024. At this stage, references are collected by reading scientific journals and conducting internet searches related to the application of data mining, especially K-Means Clustering, in sales prediction. The data was obtained after conducting direct interviews with the owner of the Aldi Distro Clothing Store so that the sales transaction data for the clothes sold was known. Data collection for this study was carried out through several stages, including direct interviews with the owner of the Aldi Baju Distro store to obtain information on clothing sales transactions. The dataset was then collected from April to July 2024, comprising 393 items obtained from the store. This research was also supported by a literature review covering various references related to the application of data mining

methods, particularly K-Means Clustering, which is used to analyze and predict sales patterns.

Research stages are the steps taken to complete the research, from initial design to testing. They begin with data collection, then move on to pre-processing, then data processing, and finally validation, as shown in the following sequence in Figure 1:

Figure 1. Research Flow

The literature review stage is an important step taken to collect various relevant references and support research. This process includes reviewing information sources originating from various media, including books, scientific journals, research articles, and audio-visual materials such as videos [13]. This review aims to enrich the theoretical foundation that serves as the basis for designing more effective and innovative programs. In the context of this research, the literature review stage aims to deepen understanding of the basic concepts of data mining and its application that can support data management in Distro Clothing Stores. Therefore, it is intended that the research results can show a significant contribution to the development of programs that are more useful and applicable.

strategically to support this research. The data collected consisted of 393 entries, collected between April and July 2024. This information was obtained through a careful and systematic analysis of sales summaries for the two months, which is expected to provide a comprehensive picture of sales activity at the store.

Data pre-processing is a very important and fundamental step that must be carried out before data can be analyzed using clustering methods such as K-Means [14]. This stage aims to ensure that the data is in the best condition so that the analysis carried out produces a high level of accuracy and validity. Some of the main steps usually carried out in the data pre-processing stage include:

a. Data Cleaning

Eliminating incomplete or missing data, addressing issues related to data duplication, and implementing procedures to ensure consistency in data used in analysis or decision-making.

b. Data Normalization

To improve the accuracy of the data clustering process, data is transformed into a uniform scale through the application of methods such as min-max scaling or standardization. This transformation process aims to reduce imbalances that may arise due to differences in scale between variables, so that variables with a larger range of values, such as price, do not disproportionately affect the clustering results compared to variables with smaller scales, such as sales quantity or sales volume. This allows for a more fair and representative analysis of all variables involved.

c. Feature Selection:

To ensure the accuracy and relevance of the clustering results, a crucial initial step is selecting attributes deemed relevant. These attributes, such as sales quantity and item price, are selected based on theoretical and empirical considerations to represent the main characteristics of the analysis. This selection process aims to reduce noise or unnecessary information, so that the clustering algorithm can work more efficiently and produce more meaningful data groupings.

In the implementation phase, data processing was carried out using the Google Collaboratory platform as the primary tool for running the k-means algorithm. In this process, the k-means algorithm was not only applied independently but also combined with two relevant evaluation techniques, namely the Silhouette Coefficient and the Davies-Bouldin Index. The combination of these two techniques aims to evaluate the performance of the clustering model more comprehensively, thus enabling the determination of the optimal number of clusters and ensuring that the resulting groupings are of high quality and in line with the characteristics of the analyzed data.

RESULTS AND DISCUSSION

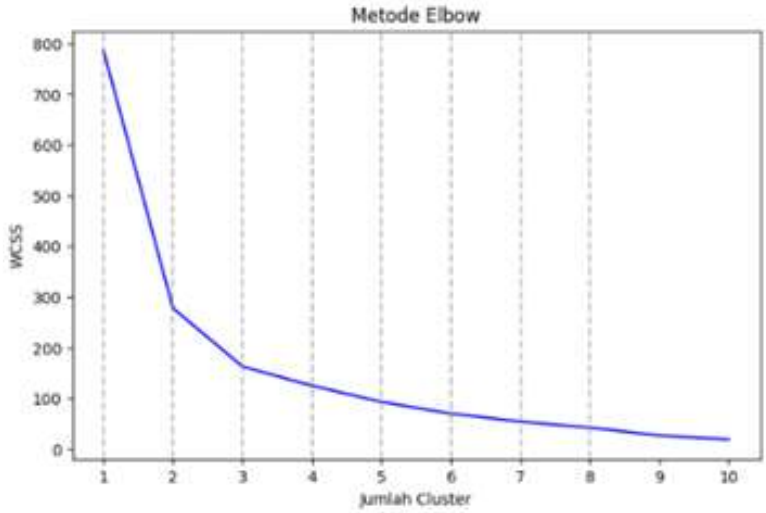
The collected data consists of 393 rows of distro clothing sales data at the Aldi Sidoarjo store in .csv (Microsoft Excel) format. The data contains six main attributes used as variables to support the prediction process: Date, Payment Status, Item Type, Quantity, Price, and Amount. The following is a sample of the data presented in Table 1:

In the next stage, this data collection was processed using Google Collaboratory as the primary tool for running the k-means algorithm. In this process, the data was categorized. Categorization is the process of transforming data to make it easier to process. Because some of the attributes used were non-numeric, it was necessary to convert the data to numeric form. The data converted to categorical form is shown in Table 2.

The initial data processing process is carried out with the aim of improving data quality and making the data ready for further processing [8]. The stages carried out

include taking the required attributes, namely the attributes "Type of Goods", "Quantity", and "Amount" and removing unused attributes. Next, remove the symbol "Rp" and the punctuation mark "period (.)" in the contents of the rental cost attribute per day from the original "Rp. 35,000" to "35000". Sample data that has gone through the previous data pre-processing process will be shown in table 3

K-Means Clustering is a data mining method that aims to classify data based on certain characteristics. In this study, the K-Means method was utilized to analyze and group distribution clothing sales patterns by considering attributes such as sales volume, product category, and sales period. The optimal number of clusters was determined using the Elbow method, which utilizes the Within-Cluster Sum of Squares (WCSS) graph [15]. In the graph, the optimal number of clusters can be identified at the elbow point, which indicates the number of clusters that provide a balance between variation within the cluster and model complexity. Based on the analysis results, the elbow point was identified at k = 2 or k = 3, indicating the optimal number of clusters is within that range. The following calculation results from the elbow method are shown in Figure 2:



Gambar 2. Grafik Metode Elbow

After calculations using the Elbow method, sampling clusters ranging from 1 to 10, a curvature was observed in clusters 2 and 3. This proves that the optimal clusters are in clusters 2 and 3. However, to test the accuracy of the clusters, a matrix is required for calculation.

Once the optimal number of clusters is determined, the K-Means algorithm is applied through a series of steps. This process begins with the initial initialization of randomly determined cluster centers. Next, each data item is calculated for its Euclidean distance to the predetermined cluster center and then grouped into the closest cluster. The cluster centers are then updated based on the average value within each cluster. This process continues iteratively until there is no significant change in the cluster centers. The clustering results yield three main categories describing sales characteristics: high-selling products, medium-selling products, and low-selling products.

The results of this method indicate that high-selling products can be prioritized in marketing strategies, while medium-selling products require additional promotional

efforts to increase their appeal. Products with low sales require further evaluation to determine their sustainability feasibility or implement innovative strategies to increase consumer interest. The insights generated from this clustering provide significant benefits to the distro industry in developing more effective marketing strategies. With a better understanding of sales patterns, business owners can allocate resources more efficiently, optimize promotional strategies, and increase competitiveness in the market.

Overall, the K-Means Clustering method has proven to be an effective approach for clustering distro clothing sales patterns and providing valuable insights for business decision-making. With this data-driven approach, marketing strategies can be more focused and tailored to market characteristics, thereby increasing the chances of success in the increasingly dynamic distro industry.

Results Evaluation

The clustering results visualized in the scatterplot indicate that the data is segmented into a number of color-coded groups, with each group representing a cluster formed by the K-Means algorithm.

CONCLUSION

Fundamental Finding : The study concludes that K-Means Clustering effectively identifies three distinct sales performance groups—low, medium, and high—supported by strong validation indices, showing that sales patterns are multidimensional and influenced by both quantity sold and revenue contribution. **Implication :** The findings imply that business owners should utilize data-driven segmentation to optimize marketing strategies, improve inventory management, and prioritize high-performing product clusters to enhance operational efficiency and competitiveness. **Limitation :** This study relies solely on quantitative transaction data, without incorporating additional influencing factors such as customer behavior, seasonal variations, product categories, or promotional activities, which may limit comprehensive interpretation. **Future Research :** Further research should integrate more diverse variables and explore advanced analytical methods or predictive models to examine broader determinants of sales performance and provide deeper insights for strategic business decision-making.

REFERENCES

[1] R. G. Solechati and A. Jananto, "Penerapan Algoritma K-Means Clustering Pada Data Brain Stroke Untuk Pengelompokan Profile Pasien," *semanTIK*, vol. 9, no. 1, p. 39, 2023, doi: 10.55679/semantik.v9i1.29446.

[2] R. Maulana, "OPTIMISASI PENGGUNAAN ALGORITMA MACHINE LEARNING," vol. 1, no. 6, pp. 1–16, 2024.

[3] C. Purnama, W. Witanti, and P. Nurul Sabrina, "Klasterisasi Penjualan Pakaian untuk Meningkatkan Strategi Penjualan Barang Menggunakan K-Means," *J. Inf. Technol.*, vol. 4, no. 1, pp. 35–38, 2022, doi: 10.47292/joint.v4i1.79.

[4] A. T. Suseno, A. R. Naufal, and M. Al Amin, "Market Based Analysis Sebagai Peningkatan Penjualan Produk Menggunakan Algoritma K-Medoids Dan Fp-Growth," *J. Tek. Inf. dan Komput.*, vol. 5, no. 2, p. 301, 2022, doi: 10.37600/tekinkom.v5i2.646.

[5] P. Putra, "PENGEMBANGAN MODEL PREDIKSI RISIKO KREDIT," vol. 1, no. 6, pp. 1–18, 2024.

[6] O. B. Ginting, A. Anita, and E. Y. Tumanggor, "Penerapan Metode Trend Moment Untuk Memprediksi Jumlah Penjualan Dan Stok Kopi Pada Omilen Coffee," *J. Tekinkom (Teknik Inf. dan Komputer)*, vol. 7, no. 1, pp. 395–401, 2024, doi: 10.37600/tekinkom.v7i1.1332.

[7] I. B. Perkasa and I. Komputer, "STRATEGI DATA MINING UNTUK IDENTIFIKASI POLA," vol. 1, no. 6, pp. 1–16, 2024.

[8] R. Dwi Putra, "Klasifikasi Penjualan Produk Customer Relationship Management dengan Algoritma K-Nearest Neighbors," *J. Comput. Scine Inf. Technol.*, vol. 8, pp. 48–55, 2022, doi: 10.35134/jcsitech.v8i2.34.

[9] Ismai, "Metode Klasifikasi Menentukan Kenaikan Level UKM Bandung Timur Dengan Algoritma Naive Bayes Pada Sistem JURAGAN Berbasis Komunitas," vol. 03, no. 01, pp. 24–31, 2020.

[10] N. Farida, M. T. Chulkamdi, and Z. Wulansari, "Application of Data Mining By Using a Priori Algorithm To Improve Customer Purchasing Decisions At Mikamart Blitar Store," *Int. J. Multidiscip. Res. Lit.*, vol. 1, no. 5, pp. 526–534, 2022, doi: 10.53067/ijomral.v1i5.58.

[11] R. Komansilan, V. Tarigan, and A. Yusupa, "Analisis Perbandingan Metode Trend Moment dan Regresi Linear Untuk Meramal Harga Saham Bank BRI," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 7, no. 1, p. 24, 2024, doi: 10.53513/jsk.v7i1.9474.

[12] L. P. Dalova, Nurmawanti, N. E. Faizah, and S. B. Syahputro, "Efektifitas Penerapan Customer Relationship Management Pada Usaha Jasa Desain Iklan Citra Karya Setia (Advertising & Digital Printing) Melalui Pemasaran Electronic Word of Mouth (E-Wom)," *Neraca Manajemen, Akunt. Ekon.*, vol. 1, no. 3, pp. 1–17, 2023.

[13] A. Nugraha, O. Nurdiawan, and G. Dwilestari, "PENERAPAN DATA MINING METODE K-MEANS CLUSTERING UNTUK ANALISA PENJUALAN PADA TOKO YANA SPORT," 2022.

[14] N. Luh, P. P. Dewi, I. Nyoman Purnama, and N. W. Utami, "Penerapan Data Mining Untuk Clustering Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: STMIK Primakara)," Jurnal Ilmiah Teknologi Informasi Asia, vol. 16, no. 2, 2022.

[15] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, I. Setiawan Mangku Negara, I. Ahmad Ashari, P. Studi Teknologi Informasi, F. Sains dan Teknologi, and U. Harapan Bangsa, "Terakreditasi SINTA Peringkat 4 Analisa Cluster Data Transaksi Penjualan Minimarket Selama Pandemi Covid-19 dengan Algoritma K-means," 2021.

Shafa Arrizqa Az Zahro
Muhammadiyah University of Sidoarjo, Indonesia

***Nuril Lutvi Azizah (Corresponding Author)**
Muhammadiyah University of Sidoarjo, Indonesia
Email: nurillutviazizah@umsida.ac.id

Novia Ariyanti
Muhammadiyah University of Sidoarjo, Indonesia

Irwan Alnarus Kautsar
Muhammadiyah University of Sidoarjo, Indonesia