GOLDEN RATIO OF DATA IN SUMMARY
VOLUME 5, ISSUE. 4 (2025)
Website: https://goldenratio.id/index.php/grdis

GOLDEN RATIO
ISSN [Online]: 2776-6411

DATA IN SUMMARY | ACCOUNTING, MANAGEMENT, BUSINESS, ECONOMICS

Classification of Reject Patterns Based on Production Stages Using the K-Means Clustering Method

Renita Lestari[1], Elfina Novalia[2], Tukino[3], Fitria Nurapriani[4]

[1,2,3,4] Department of Information Systems, Faculty of Computer Science, Universitas Buana Perjuangan Karawang, Karawang, Indonesia. Email: si22.renitalestari@mhs.ubpkarawang.ac.id[1], elfinanovalia@ubpkarawang.ac.id[2], tukino@ubpkarawang.ac.id[3], fitria.apriani@ubpkarawang.ac.id[4]

ARTICLE HISTORY

Received: May 13, 2025
Revised: October 20, 2025
Accepted: October 20, 2025

DOI
https://doi.org/10.52970/grdis.v5i4.1301

ABSTRACT
This study aims to classify reject patterns in the production process using the K-Means Clustering method. The dataset consists of 870 records collected from the production line, containing information such as product name, reject type, process stage, and production quantity. Through a data mining approach, data preprocessing steps such as cleaning, encoding, and normalization were performed prior to the clustering process. The Elbow Method indicated that the optimal number of clusters is three. Each cluster exhibits distinct characteristics: light rejects with small quantities in early stages, heavy rejects with large quantities, and moderate rejects with random distribution. These findings are expected to assist management in formulating more targeted strategies for process improvement and quality control. By identifying common reject patterns within each cluster, companies can adopt a more proactive approach to minimizing production defects and enhancing overall operational efficiency.

Keywords: Data Mining, K-Means Clustering, Production Rejects, Manufacturing Process.

I. Introduction

Modern manufacturing industries increasingly depend on efficient production processes to ensure product quality and minimize defects at each stage of operation. In the semiconductor sector, production errors commonly referred to as rejects represent a major challenge that not only reduces productivity but also increases operational costs and undermines customer trust. To overcome these challenges, companies need to implement data-driven approaches that can analyze and identify reject patterns across the entire production process. (Segai et al., 2024). With the rapid advancement of information technology, data mining techniques have become powerful tools for managing large and complex industrial datasets. Among the various approaches in data mining, clustering is widely utilized to group data based on similar attributes. The K-Means clustering algorithm, in particular, is favored for its simplicity, efficiency, and capability to uncover hidden patterns in data.(Haekal & Mu'min, 2024)

This study aims to classify reject patterns across different production stages using the K-Means clustering method. The analysis focuses on identifying clusters of rejects with similar characteristics based on attributes such as product name, type of reject, production stage, and the quantity of both products and rejects. By identifying these patterns, manufacturers can gain valuable insights into which stages contribute most to production defects and which types of rejects occur most frequently. (Harani et al., 2020). The research

ISSN [Online]: 2776-6411

process involves several key phases, including data collection, data cleaning, data transformation through encoding and normalization, and determining the optimal number of clusters using the Elbow Method. The K-Means algorithm is then applied to group the data based on similarities in patterns. The resulting clusters are subsequently analyzed and visualized to uncover the distinctive traits of each group.(Barimbing & Hidayat, 2024)

Furthermore, the integration of clustering techniques such as K-Means into quality control systems reflects a broader shift towards Industry 4.0, where data plays a central role in predictive decision-making and operational optimization. By embedding analytical methods into production oversight, companies are better equipped to transition from reactive quality measures to proactive strategies, allowing for continuous improvement and reduced defect rates over time. This research, therefore, not only contributes to the academic exploration of data mining applications in manufacturing but also serves as a practical reference for industries aiming to enhance their competitiveness through intelligent data utilization. (Ellang Putro Priambodo & Arief Jananto, 2023)

II. Literature Review and Hypothesis Development

2.1. Data Mining in Manufacturing

Data mining is a process of extracting valuable insights and patterns from large datasets, especially when traditional analytical methods are insufficient due to the volume or complexity of the data. In the manufacturing industry, data mining techniques are widely adopted to support quality control, production forecasting, defect detection, and process optimization. Studies such as have emphasized the relevance of data-driven decision-making to improve product quality and reduce manufacturing costs.(Widodo et al., 2024). Clustering, a subset of unsupervised learning in data mining, is particularly useful when there are no predefined categories and the goal is to discover inherent groupings within the data. This approach enables companies to identify patterns and anomalies in production processes that may not be immediately visible.(Ika Anikah et al., 2022)

2.2. K-Means Clustering Algorithm

K-Means is one of the most commonly used clustering algorithms due to its simplicity and computational efficiency. The algorithm partitions data into k clusters by minimizing the variance within each cluster and maximizing the distance between clusters. It works iteratively by assigning data points to the nearest cluster centroid and then recalculating centroids based on the new assignments until convergence is reached. (Budi et al., 2022). K-Means has been successfully applied in various manufacturing contexts. For example, utilized K-Means to detect process deviations in automated production lines. In the context of quality management, clustering rejected data helps isolate frequently occurring issues and enables targeted corrective actions. (Dila et al., 2025).

2.3. Reject Analysis and Quality Control

Rejects in manufacturing refer to products that do not meet specified quality standards and must be discarded or reworked. Understanding the patterns and causes of rejects is critical for implementing effective quality control measures. Categorizing reject types based on production stages and defect characteristics allows companies to trace back the root causes and prevent future occurrences. (Pooja et al., 2022). In semiconductor manufacturing, where processes are highly complex and sensitive to slight variations, clustering reject data can reveal systematic errors that are otherwise difficult to detect manually. By grouping similar reject patterns, companies can identify high-risk stages and prioritize improvements accordingly. (Syani et al., 2024)

ISSN [Online]: 2776-6411

2.4. Application of Clustering for Quality Control

The application of clustering methods in quality control allows companies to group reject data based on pattern similarities, so that in-depth analysis of the root cause can be carried out. By grouping data based on attribute similarities, management can formulate more targeted corrective actions, allocate resources more efficiently, and design long-term quality improvement strategies. Clustering also opens up opportunities for real-time process monitoring as part of the implementation of intelligent manufacturing systems. (Zeda Al Widad & Malik, 2022). By applying the K-Means method to production reject data, three main groups of reject patterns were successfully identified. This approach not only helps to understand the causes of rejection but also provides direction in improving the efficiency of the production process and quality control. Further research can deepen the analysis by considering time variables, operators, and production machine parameters. (Setiawan, 2022)

2.5. Hypothesis Development

Based on the reviewed literature and the conceptual understanding of data mining applications in manufacturing, particularly in relation to clustering techniques for quality analysis, the following hypotheses are proposed as the foundation for this research:

H1: Reject data in the production process can be effectively grouped into meaningful clusters using the K-Means algorithm.
H2: Each resulting cluster represents a distinct pattern of rejects, characterized by a combination of product type, reject type, process stage, and quantity.
H3: Identifying these clusters can support better decision-making in quality control and defect prevention strategies.

These hypotheses guide the analytical process in this study, aiming to validate whether clustering techniques can contribute actionable insights to the improvement of manufacturing quality systems.

III. Research Method

This research was conducted using a quantitative approach with the unsupervised learning method, especially the K-Means Clustering algorithm, to classify reject patterns in the production process. The purpose of this method is to group reject data into several clusters based on similar characteristics such as products, types of rejects, production stages, and quantities of products and rejects. (Darmawan & Yudistira, 2017)

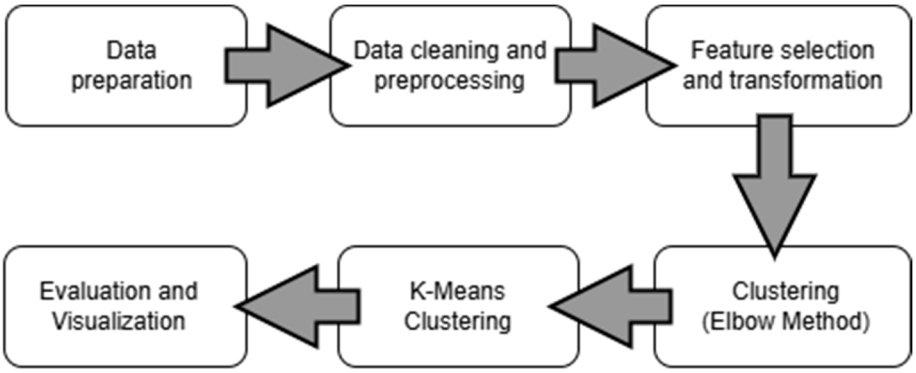

Figure 1. Conceptual Framework

ISSN [Online]: 2776-6411

The research stages broadly include five main stages, namely: (1) Data Preparation, (2) Data Cleaning and Pre-processing, (3) Feature Selection and Transformation, (4) Elbow Method Clustering Process, (5) K-Means Clustering, and (6) Evaluation and Visualization of Results. The following diagram illustrates the overall process flow:

3.1. Data Preparation

The dataset is collected from internal sources of the manufacturing company and stored in .xlsx format. The dataset contains attributes such as Product Name, Stage From, Stage To, Qty From, Qty To, Reject Name, Reject Qty, and Reject Rate (%). Next, the data is loaded and checked to ensure integrity.(Miwan Kurniawan Hidayat & Rina Fitriana, 2022)

3.2. Data Cleaning and Pre-processing

This process includes removing duplicates, fixing inconsistent values, and refining the logic between Qty From and Qty To, to ensure that the reject value is calculated correctly. In addition, encoding is performed on categorical variables, and normalization of numeric data is performed to accelerate convergence during the clustering process.(Watmah et al., 2024)

3.3. Feature Selection and Transformation

Relevant features are retained based on the needs of the analysis. Product name, reject type, and production stage name are simplified to one word for processing efficiency. Label Encoding is performed for categorical variables, while Min-Max normalization is used for numeric variables such as Qty and Rate.(Mentu et al., 2023)

3.4. Elbow Method

Before applying K-Means, the Elbow Method is used to determine the optimal number of clusters. This method calculates the Within-Cluster Sum of Squares (WCSS) for different values of k. The result is plotted in a line graph, and the point at which the curve begins to flatten (forming an "elbow") is selected as the optimal k. In this study, the elbow was identified at k = 3, indicating that three clusters represent the most efficient grouping of the dataset.(Sri Irtwaty, 2022)

3.5. Clustering with K-Means

The clustering process begins with determining the optimal number of clusters using the Elbow method. The results of the Elbow analysis show that 3 clusters are the ideal number. Then, the K-Means algorithm is applied to divide the data into three groups based on pattern similarity. (Kurnia & Muhammad, 2023)

3.6. Evaluation and Visualization

The clustering results are analyzed visually using PCA for dimensionality reduction. Each cluster is then analyzed based on dominant characteristics such as reject volume, production stages that are often problematic, and types of rejects that often appear. This aims to provide an interpretation of each cluster.(Arrosyad et al., 2024)

GOLDEN RATIO
ISSN [Online]: 2776-6411

IV. Results and Discussion

4.1. Dataset

The dataset used in this study contains information related to production and rejection activities in manufacturing companies. Figure 1 below displays the first five rows of the dataset.

	Product Name	Stage From	Stage To	Qty From	Qty To	Reject Name	Reject Qty	Reject Rate (%)
0	Produk A	Final Test	Visual Inspection II	5000	4754	Bin 1	246	2.460000
1	Produk A	Final Test	Visual Inspection II	5000	4754	Bin 3	246	2.460000
2	Produk B	Dejunk	Time Cycle Test	2500	2400	Coplanarity	100	4.000000
3	Produk C	Marking	Singulation	1990	1988	Double Mark	2	0.100503
4	Produk C	Molding	Post Mold Cure	2000	1990	Chipping	10	0.500000

Figure 2. Initial Dataset

Next, the Pre-Processing data stage in Figure 2 shows the data representation after going through the encoding process before the clustering stage with the K-Means algorithm. This process converts all categorical attributes into a numeric format so that they can be analyzed mathematically. This display represents the data structure used in the reject pattern classification process based on production and quantity attributes, as a basis for forming groups that have similar characteristics.

	Product	StageFrom	StageTo	QtyFrom	QtyTo	Reject	RejectQty	RejectRate
0	0	5	2	202	199	3	3	1.49
1	2	5	0	279	276	0	3	1.08
2	0	0	4	192	188	1	4	2.08
3	2	0	5	114	106	3	8	7.02
4	0	0	2	206	200	9	6	2.91

Figure 3. Dataset After Encoding Process

4.2. Elbow Method

To determine the optimal number of clusters, the Elbow method is used, which visualizes the Within-Cluster Sum of Squares (WCSS) value against the number of clusters (k). In Figure 2, the Elbow graph shows a clear elbow point at a value of k = 3, indicating that three clusters are the optimal number for grouping data based on the similarity of reject patterns.

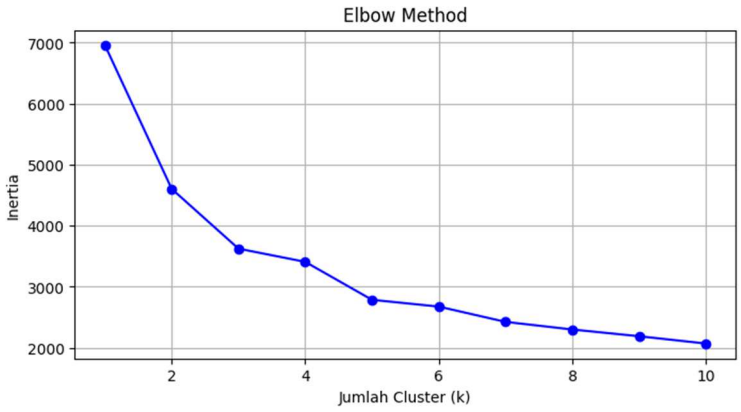

Figure 4. Elbow Method

ISSN [Online]: 2776-6411

4.3. Cluster Distribution

After preprocessing 870 cleaned and simplified production data, the K-Means algorithm was applied to group the data based on the similarity pattern between attributes. Based on the analysis using the Elbow method, it was found that the optimal number of clusters is 3. The selection of the number of clusters is based on the elbow point, which shows the last significant decrease in the WCSS value. With this division, each cluster represents a different reject pattern in the production process, which is then analyzed to obtain relevant insights into potential causes and solutions to improve production quality.

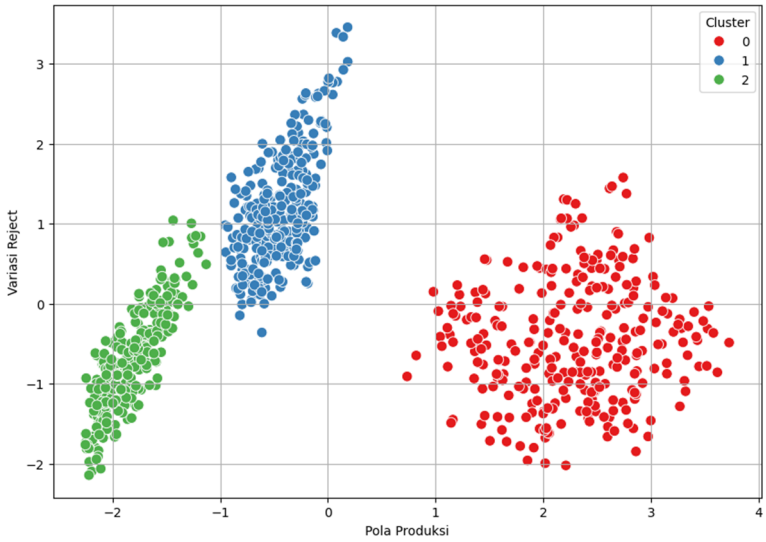

Figure 5. K-Means Cluster Count

After the number of clusters is set at three, the K-Means algorithm is applied to the data based on the similarity of the reject patterns. For modeling purposes, categorical features are encoded using the Label Encoding technique, while numeric features are normalized using MinMaxScaler.

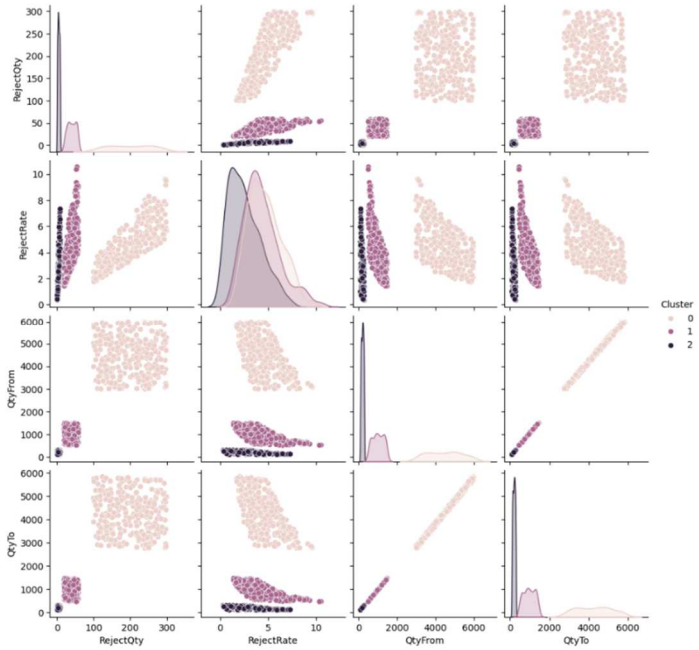

Figure 6. Cluster Distribution Results

ISSN [Online]: 2776-6411

Based on the clustering results using the K-Means algorithm, the rejected data in the production process is divided into three main clusters, each with the following characteristics:

a. Cluster 0 – Light Rejects at the Beginning of Production

This cluster is characterized by low RejectQty and RejectRate values. Reject cases generally occur in the early stages of production (low Stage From) and are dominated by light reject types, such as small stains or dust. Products that experience rejection in this cluster tend to vary and do not show a dominant pattern. Interpretation of this pattern indicates that the problem is likely caused by light human error or less-than-optimal machine settings at the beginning of the process. Therefore, improvements to the initial SOP and increasing the cleanliness of the production area are important.

b. Cluster 1 – Heavy Rejects at the Middle to Late Stages

This cluster has the highest RejectQty and RejectRate values compared to other clusters. Rejects often occur in the middle to late stages of production, with dominant reject types such as misalignment and physical damage. This cluster also shows repeated occurrences in certain products. This indicates a serious problem, possibly from tool wear, fixture problems, or a suboptimal final inspection process. Regular tool audits and maintenance rescheduling are needed to reduce rejects in this cluster.

c. Cluster 2 – Mixed Pattern

This cluster shows moderate RejectQty and RejectRate values. The production stages in this cluster vary, from beginning to end, without any dominant type of reject or product. The interpretation of this pattern is that cluster 2 is a mixture of mild and severe cases that can still be considered a normal condition of the production process. However, regular monitoring is still needed to anticipate a shift in the pattern towards a more problematic cluster.

4.4. Distribution of Cluster Reject Rate

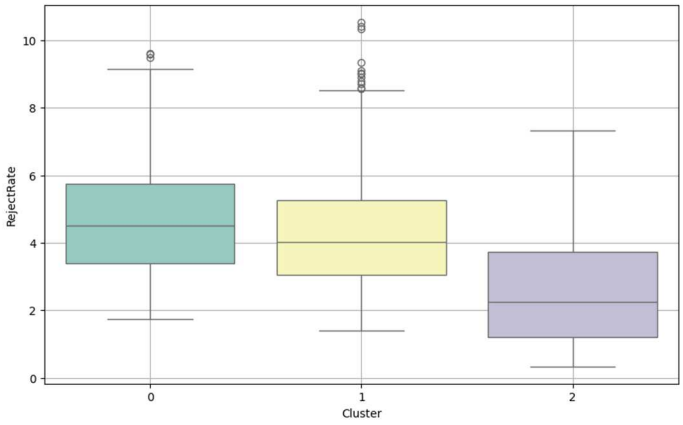

Figure 7. Visualization of Reject Rate for Each Cluster

Figure 5 shows the distribution of Reject Rate values in each cluster resulting from the clustering process using the K-Means method. It can be seen that:

a. Cluster 0 has a median Reject Rate of around 4.5 with a fairly even distribution of data, although there are several outliers at higher values. This indicates mild rejects that tend to be stable but still require attention to the potential for increasing reject values.
b. Cluster 1 has a similar distribution to Cluster 0, but with a greater number of outliers and a higher maximum Reject Rate value. This supports the previous interpretation that this cluster contains severe rejects, especially in the middle to late stages of production.

GOLDEN RATIO
ISSN [Online]: 2776-6411

c. Cluster 2 shows the lowest median Reject Rate among the three, with a narrower distribution and fewer outliers. This confirms that this cluster consists of rejected cases with moderate to low intensity and non-prominent characteristics.

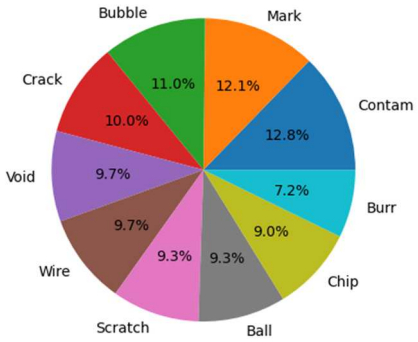

Figure 8. Dominant Reject Type in Cluster 1 (k=0)

Figure 6 shows the percentage distribution of dominant reject types included in Cluster 1, which is the cluster with the highest Reject Qty and Reject Rate characteristics. Based on the pie chart, it is known that the most frequent type of rejection is Contam (contamination) at 12.8%, followed by Mark at 12.1%, and Bubble and Crack at 11.0% and 10.0%, respectively. Other types of rejects, such as Void, Wire, Scratch, Ball, Chip, and Burr, have relatively even proportions, ranging from 7.2% to 9.7%. The dominance of contamination and surface damage type rejects indicates that quality problems in Cluster 1 are most likely caused by production environment factors, suboptimal final quality control, or potential damage due to manual handling.

V. Conclusion

The analysis shows that cluster 0 is a product with light rejects, small quantities, dominated by early stages 1-2. Cluster 1 is a product with large rejects, large quantities, and often in certain production processes. Cluster 2 is a total mixed quantity, moderate rejects, and variations in stages. The results of the analysis show that in the early stages, it is necessary to improve process supervision to prevent light rejects from spreading. The final stage should focus on final validation and strict quality control. And indicates the need for a cross-check process between stages. By applying the K-Means method to production reject data, three main groups of reject patterns were successfully identified. This approach not only helps to understand the causes of rejection but also provides direction in improving the efficiency of the production process and quality control. Further research can deepen the analysis by considering time variables, operators, and production machine parameters.

References

Arrosyad, A. A., Purnamasari, A. I., & Ali, I. (2024). Implementasi algoritma K-Means clustering untuk analisis persebaran UMKM di Jawa Barat. *Jurnal Mahasiswa Teknik Informatika, 8*(3).

Barimbing, A. O., & Hidayat, R. R. (2024). Rancang bangun customer relationship management dengan K-Means untuk reward customer berbasis web. *Jurnal Teknik Informatika dan Sistem Informasi, 10*(2). https://doi.org/10.28932/jutisi.v10i2.7138

Darmawan, H., & Yudistira, N. (2017). Penggunaan model Tabular Retrieval Model (TabR) dengan K-Means clustering untuk efisiensi klasifikasi dan regresi data tabular, *1*(1). http://j-ptiik.ub.ac.id

Dila, R., Saputra, R., & Ramadhanu, A. (2025). Klasifikasi timun segar dan busuk menggunakan K-Means clustering dengan peningkatan noise reduction dan median filter, *6*(1).

ISSN [Online]: 2776-6411

Priambodo, E. P., & Jananto, A. (2023). Perbandingan analisis cluster algoritma K-Means dan …

Haekal, J., & Mu'min, R. (2024). Structured defect data-based K-Means clustering analysis and framework for quality control (QC) prioritization in manufacturing, *18*(3), 405–416.

Harani, N. H., Prianto, C., & Nugraha, F. A. (2020). Segmentasi pelanggan produk digital service Indihome menggunakan algoritma K-Means berbasis Python. *Jurnal Manajemen Informatika (JAMIKA).* https://doi.org/10.34010/jamika.v10i2

Anikah, I., Surip, A., Rahayu, N. P., Al-Musa, M. H., & Tohidi, E. (2022). Pengelompokan data barang dengan menggunakan metode K-Means untuk menentukan stok persediaan barang. *KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer, 4*(2), 58–64. https://doi.org/10.32485/kopertip.v4i2.120

Industri, K., Budi, S., Sakur, H., Silangen, M., & Tuwohingide, D. (2022). Penerapan algoritme K-Means cluster dan metode TOPSIS pada pemilihan mahasiswa kunjungan industri.

Kurnia, N., & Muhammad, A. (2023). IT management dengan menggunakan metode K-Means clustering untuk pengelompokan stok barang. *Jurnal Sains Informatika Terapan (JSIT), 2*(1).

Zeda Al Widad, P. U., & Malik, K. (2022). Implementasi algoritma K-Means clustering untuk loyalitas pelanggan berbasis web di Majutoto Malang, *x*(x), 1–5.

Zahrial, O. T., Siddik, M., & Kom, M. (2023). Penerapan metode K-Means clustering dalam perancangan sistem penjualan. *Jurnal Mahasiswa Aplikasi Teknologi Komputer dan Informasi, 5*(2), 138–143.

Hidayat, M. K., & Fitriana, R. (2022). Penerapan sistem intelijensia bisnis dan K-Means clustering untuk memantau produksi tanaman obat. *Jurnal Teknologi Industri Pertanian, 32*(2), 204–219. https://doi.org/10.24961/j.tek.ind.pert.2022.32.2.204

Pooja, N., Saputra, M., Aisyah, S., & Juanta, P. (2022). Implementasi data mining clustering data valuasi ekspor kertas Indonesia menggunakan algoritma K-Means. *Jurnal Sistem Informasi dan Ilmu Komputer Prima, 5*(2). https://www.bps.go.id

Sebagai, D., Satu, S., Untuk, S., Gelar, M., & Strata, S. (2024). Penerapan metode K-Means clustering untuk analisis profil lulusan teknik industri di dunia kerja (Tugas akhir).

Setiawan, S. (2022). Implementasi data mining clustering dengan metode K-Means untuk mengelola persediaan stok, *3*(2), 146–164. https://katalog.data.go.id/dataset/banyaknya-persediaan-dan-pemakaian-kabupaten

Syani, M., Wahyudi, T., & Studi Sistem Informasi, P. (2024). Klasterisasi penggunaan ban dengan cost per kilometer terendah pada PT. PL menggunakan metode K-Means. *Jurnal Indonesia: Manajemen Informatika dan Komunikasi (JIMIK), 5*(3). https://journal.stmiki.ac.id

Ulfah, M., & Sri Irtwaty, A. (2022). Penerapan data mining clustering menggunakan metode K-Means dalam pengelompokan buku perpustakaan Politeknik Negeri Balikpapan.

Watmah, S., Riana, D., & Astuti, R. D. (2024). Penerapan K-Means dan K-Medoids berbasis RFM pada segmentasi pelanggan di masa pandemi COVID-19. *INTI Nusa Mandiri, 18*(2), 192–200. https://doi.org/10.33480/inti.v18i2.4963

Widodo, A., Widyastuti, R., & Hendrawan, S. A. (2024). Prediksi biaya logistik menggunakan metode K-Means. *Innovation and Technology, 1*(1).