

Implementation of the K-Means Algorithm for Analyzing Patient Visits at the Kupang Community Health Center

Mahendra Adi Kusuma
Informatics Department

Universitas Muhammadiyah Banjarmasin
Banjarmasin, Kalimantan Selatan

mahendra_adi_2255201110010@umbjm.ac.id

Finki Dona Marleny
Informatics Department

Universitas Muhammadiyah Banjarmasin
Banjarmasin, Kalimantan Selatan

finkidona@umbjm.ac.id

Ayu Ahadi Ningrum
Informatics Department

Universitas Muhammadiyah Banjarmasin
Banjarmasin, Kalimantan Selatan

ayuahadi@umbjm.ac.id

Abstract—Patient visit data at community health centers are still managed manually, making it difficult for healthcare workers to identify visit patterns and make informed decisions. This study applies the K-Means Clustering algorithm to group patient visit data at Puskesmas Kupang based on insurance type, visit category, clinic or ward, and patient age group. The dataset consists of thousands of patient visit records, with six key features selected through three feature engineering experiments. Preprocessing involved data cleaning and Label Encoding, while the clustering process was configured using PyCaret with normalization, two-component PCA dimensionality reduction to address inter-feature correlation and enable visual cluster representation and outlier removal. The optimal number of clusters was set to four using the Elbow Method. The best-performing model achieved a Silhouette Score of 0.6558, indicating good clustering quality. The four resulting clusters represent patient segments with distinct profiles: adult BPJS users accessing the general clinic, young productive-age patients utilizing Maternal and Child Health services, patients with unplanned service needs, and elderly patients with routine visit patterns. The findings confirm that K-Means is effective as a patient classification tool at community health centers, supporting data-driven decision-making for healthcare personnel.

Keywords—K-Means Clustering, Data Mining, Community Health Center, Patient Visits, PyCaret, Elbow Method.

I. INTRODUCTION

Community Health Centers (Puskesmas) play a central role in the Indonesian National Health Insurance (JKN) system as primary care facilities that serve as the main gateway for the public to access health services. Because the JKN enrollment rate exceeds 96% of Indonesia's population, community health centers experience a steady increase in patient visits year after year [1]. This situation places Puskesmas in an increasingly critical position while demanding more reliable healthcare service data management capabilities, so that they can effectively handle the ever-growing volume of visits and maintain the quality of service to the public.

Behind this rising volume of services lies an unresolved fundamental issue: the management of patient visit data at Puskesmas still largely relies on non-standardized manual processes. The diversity of patient

profiles, which encompass various visit attributes, presents a unique challenge when these profiles must be sorted and mapped without the support of an automated system [2]. As a result, Puskesmas cannot optimally extract the potential information that visit data contains, making it difficult to implement evidence-based resource planning, health worker deployment, and service program design [3]. This situation underscores the urgency of implementing an analytical approach capable of processing patient visit data into information that can be directly used as a basis for decision-making.

A number of studies have demonstrated the effectiveness of K-Means in analyzing health data: Ziqrullah *et al* [4] successfully clustered patient disease data at community health centers using K-Means in a more systematic manner, while Yang *et al* [5] confirmed the superiority of K-Means over other clustering methods in processing multivariate medical data. However, these studies focus on grouping by disease type or clinical records, and none has specifically examined patient segmentation based on administrative visit attributes, insurance type, visit category, clinic or ward, and age group, in a community health center operational context, which constitutes the gap this study addresses. K-Means was selected because its Euclidean distance-based mechanism suits the structured, encoded categorical attributes used in this study and offers a computationally efficient and easily interpretable solution compared to density-based or medoid-based alternatives [6], [7].

K-Means operates iteratively by assigning each data point to the nearest centroid and updating centroid positions until convergence is reached [8], producing structured patient segments that can be directly interpreted by healthcare administrators [9]. Distinct from prior work, this study contributes an iterative feature engineering protocol across three experimental configurations and incorporates PCA to address inter-feature correlation before clustering, yielding a four-cluster patient segmentation based on administrative visit attributes that translates directly into operational decisions at the facility level. This study therefore aims to produce informative patient groupings from visit data at Puskesmas Kupang as an effective data-driven decision-support instrument for healthcare personnel.

II. RESEARCH METHODOLOGY

A. Data Collection

This study utilizes secondary data obtained directly from Puskesmas Kupang through a formal request

submitted to the facility's management. The data were provided within approximately two business days and delivered in a spreadsheet format (.xlsx). The dataset encompasses patient visit records from February through October 2025, totaling 8,092 records across 15 attributes: Date, Patient Name, Medical Record Number, National ID Number, Gender, Date of Birth, Age, Visit Type, Visit, Clinic/Ward, Insurance, Diagnosis Code, Diagnosis, Case Type, and Age_Numeric. The use of secondary data derived from community health center medical records for data mining-based analysis has been established as a common practice in health informatics research in Indonesia [10].

B. Research Tools

All analytical stages in this study were carried out within a Python-based ecosystem accessed through Jupyter Notebook on the Anaconda Navigator platform. Jupyter Notebook was selected as the development environment because it supports interactive, cell-by-cell code execution, enabling each preprocessing, encoding, and modeling step to be verified independently before proceeding to the next stage. Data management and transformation were handled using the Pandas library, which provides a robust DataFrame structure and reliable value-mapping methods for the Label Encoding process. The clustering workflow was executed entirely using PyCaret's clustering module, a machine learning automation library that consolidates normalization, PCA dimensionality reduction, outlier removal, K-Means model creation, and metric evaluation into a single, reproducible pipeline. This pipeline-based approach was adopted because the mixed-type dataset, comprising both continuous numeric and encoded categorical features, requires a strict sequence of preconditions in which normalization must precede PCA and PCA must precede cluster formation, ensuring that results remain valid and consistent across all experimental runs. A summary of the software tools and their respective roles in this study is presented in Table 1.

Table 1. Software Tools Used in the Study

Software	Version / Platform
Python	3.9
Jupyter Notebook	Anaconda Navigator
Pandas	Python library
PyCaret (Clustering Module)	Python library
Microsoft Excel (.xlsx)	Source data format

C. Data Preprocessing

Prior to the modeling stage, data preprocessing was performed using Jupyter Notebook accessed through Anaconda Navigator. Ziqrullah *et al* [4] state that the data cleaning process includes handling missing values, removing duplicate records, and standardizing attribute value formats to ensure data consistency and readiness for further processing. Thorough preprocessing is an essential prerequisite for producing a valid clustering model, given that data quality directly determines the integrity of the resulting groupings [11].

D. Feature Engineering and Experimental Design

This study was conducted through three experimental iterations, each employing a different combination of features, to identify the configuration that yields the highest clustering quality. Feature selection, or feature engineering, represents a critical stage in data mining, as the chosen features directly shape the data representation and ultimately determine the quality of the resulting groupings. The specific features used in each experiment are detailed in Table 2.

Table 2. Feature Engineering Design for Each Experiment

Experiment	Features Used	Number of Features
I	Date, Gender, Date of Birth, Type of Visit, Clinic/Room, Insurance, Case Type	7 features
II	Gender, Date of Birth, Visit Type, Clinic/Room, Insurance, Case Type, Age_Number	7 features
III	Gender, Date of Birth, Visit Type, Clinic/Room, Insurance, Case Type	6 features

Experiment I incorporated the Date attribute as an input variable. Experiment II replaced the Date feature with Age_Numeric, a field representing patient age as a continuous numeric value. Experiment III employed six core features, eliminating both the Date column and the explicit Age_Numeric variable, given that age information was already captured implicitly through the Date of Birth attribute, which is processed internally by PyCaret's pipeline [12].

E. Data Encoding

Since the K-Means algorithm operates in a Euclidean distance-based numeric space, categorical features must be converted into numeric representations prior to modeling [13]. Encoding was carried out using the Label Encoding technique through manual value mapping implemented with the Pandas library, whereby each category within a given feature was consistently mapped to a unique integer value [14]. The complete encoding mapping is presented in Table 3.

Table 3. Value Mapping for Categorical Feature Encoding

Feature	Original Value (Categorical)	Encoded Value (Numeric)
Gender	MALE / FEMALE	0 / 1
Visit Type	RETURN / NEW	0 / 1
Clinic/Room	GENERAL / PEDIATRICS / EMERGENCY / OBSTETRICS / DENTAL / GENERAL INPATIENT	0 / 1 / 2 / 3 / 4 / 5
Insurance	BPJS HEALTH / GENERAL	0 / 1
Case Type	OLD / NEW	0 / 1

F. Clustering Configuration with PyCaret

The clustering process for all three sets of experiments was performed using the PyCaret library in a Jupyter Notebook environment, with the same configuration applied in each case. This configuration included: (1) data normalization to standardize the scale of values across all features, ensuring that no single feature disproportionately dominates the distance calculations; (2) dimension reduction via Principal Component Analysis (PCA) with `'pca_components=2'`, which projects the data onto the two principal components that carry the most informative variance while enabling two-dimensional cluster visualization [15]. PCA is necessary because the dataset contains six mutually correlated features, and analysis shows that “Poli/Ruangan” and “Umur_Angka” contribute dominantly to the first principal component (PC1), while “Jenis Kelamin” and “Asuransi” dominate the second principal component (PC2). This indicates the presence of a covariance structure among the features which, if left unaddressed, could lead to information redundancy in the original feature space and affect the stability of K-Means centroids; (3) automatic outlier removal to eliminate data points that deviate significantly from the general distribution; and (4) the `'verbose=False'` parameter to disable detailed log output during processing.

G. Determining the Optimal Number of Clusters

The optimal number of clusters (k) was determined using the Elbow Method, which plots the Distortion Score (Within-Cluster Sum of Squares/WCSS) against values of k ranging from 2 to 10. The elbow point, the location at which the rate of WCSS decrease begins to decelerate markedly, was adopted as the criterion for selecting the optimal k value [16]. Based on the Elbow Method visualization, Experiment I exhibited an elbow at k=4 yet was run with k=3; Experiment II showed an elbow at k=4 and Experiment III also produced an elbow at k=4, such that the final models for Experiments II and III were executed with k=4.

III. RESULTS AND DISCUSSION

A. Comparison of Evaluation Results from the Three Experiments

This study assessed the quality of each experimental run using three standard internal clustering metrics: Silhouette Score, Calinski-Harabasz Index (CH), and Davies-Bouldin Index (DBI). These metrics are widely adopted in clustering research as they evaluate quality on the basis of intra-cluster cohesion and inter-cluster separation. A comparative summary of the metric results across the three experiments is presented in Table 4.

Table 4. Comparison of Clustering Evaluation Metrics Across Three Experiments

Experiment	Number of Clusters (k)	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
I	3	0.4372	5220.4149	0.8431
II	4	0.5209	7909.5754	0.6997
III	4	0.6558	8368.0807	0.8188

Experiment I, which incorporated the Date feature, produced the lowest clustering quality among all three runs. The Date attribute is sequential in nature and does not represent any intrinsic patient characteristic; its inclusion therefore introduced noise into the inter-data distance computation and disrupted the formation of cohesive clusters. A Silhouette Score below 0.5 indicates that cluster boundaries are poorly defined.

Experiment II, which replaced the Date feature with Age_Numeric, demonstrated significant improvement across all metrics. This confirms that representing patient age as a continuous numeric variable contributes more meaningfully to the clustering process than temporal date information.

Experiment III yielded the best results across all three metrics, establishing it as the final model configuration for this study. Although Age_Numeric was not explicitly included as a feature, age-related information remained implicitly represented through the Date of Birth attribute, which is processed internally by PyCaret's pipeline, thus preventing any loss of meaningful information. This outcome is consistent with the principle that selecting only substantively relevant features and excluding those that introduce noise, produces clusters that are more cohesive and more clearly separated.

B. Evaluation of Clustering Quality in the Experiment

Table 5 presents the three internal metric evaluation results for the final experimental run at k=4.

Table 5. Results of the Quality Metrics Evaluation for the Final Experimental Cluster

Evaluation Metrics	Value	Interpretation
Silhouette Score	0.6558	Good - high internal cohesion, fairly clear separation between clusters
Calinski-Harabasz Index	8,368.0807	Clusters are dense and well-separated (highest value = best)
Davies-Bouldin Index	0.8188	Low similarity between clusters (closer to 0 = better)

A Silhouette Score of 0.6558 indicates that objects within each cluster bear strong similarity to one another while differing substantially from objects in other clusters. This value is considered satisfactory, as it surpasses the commonly cited threshold of 0.5, above which clustering quality is deemed acceptable; values exceeding 0.7 are generally regarded as indicators of excellent clustering. The Calinski-Harabasz Index of 8,368.0807, the highest value obtained across all experimental runs, signals that the clusters exhibit high internal density alongside large distances between their respective centroids, which represents an ideal outcome for clustering quality. The Davies-Bouldin Index of 0.8188 reflects a low average similarity between each cluster and its nearest neighboring cluster, indicating that the boundaries between clusters are well-defined. Taken together, these three metrics consistently confirm that the final model yields a clustering outcome of good quality.

C. Clustering Visualization

To confirm the validity of the chosen number of clusters and visualize the quality of the resulting clustering, three types of plots were generated using PyCaret's built-in visualization features: the Elbow Plot, the Silhouette Plot, and the 2D Cluster PCA Plot. These three visualizations complement each other: the Elbow Plot determines the number of clusters, the Silhouette Plot measures the quality of separation between clusters, and the PCA Plot shows the spatial distribution of clusters in two-dimensional space.

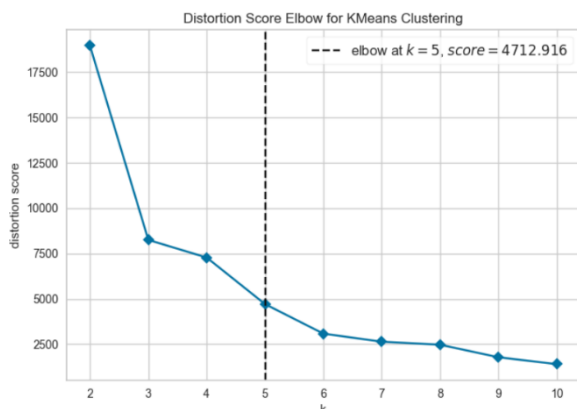


Fig 1. Elbow Method

Figure 1 plots the Distortion Score (WCSS) against values of k from 2 to 10. The curve descends steeply between $k=2$ and $k=4$, then flattens markedly, indicating that each additional cluster beyond $k=4$ yields only marginal reduction in distortion relative to the added model complexity. The inflection point falls clearly at $k=4$ (Distortion Score = 8,647.790), as indicated by the dashed vertical line, establishing $k=4$ as the optimal cluster count for this study.

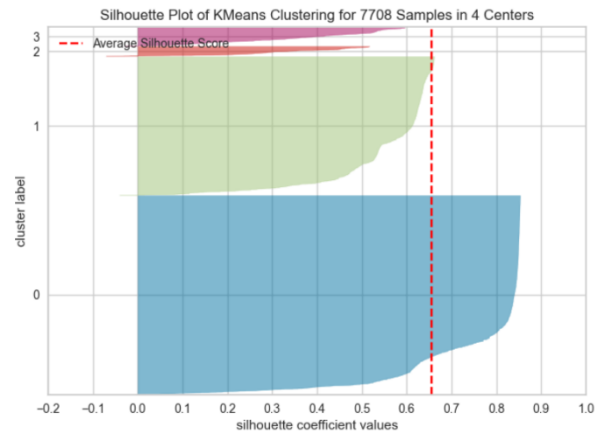


Fig 2. Silhouette Score

Figure 2 presents the silhouette plot for the $k=4$ model applied to 7,691 post-outlier-removal samples. All four clusters produce silhouette blades whose tips clear the overall mean line (0.6558, marked by the red dashed line), confirming that the majority of members in each cluster are more similar to their own cluster than to any other [18]. Clusters 0 and 3 exhibit broader blades, reflecting stronger internal cohesion, while Clusters 1 and 2 show narrower but still above-average blades consistent with their greater intra-cluster diversity. No cluster exhibits a dominant proportion of negative coefficients, indicating that misassigned data points are negligible across all groups.

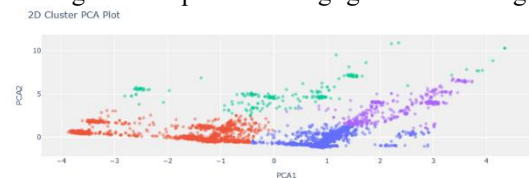


Fig 3. PCA Cluster Plot

Figure 3 displays the two-dimensional space produced by the PCA reduction step in the pipeline, the actual space in which K-Means operated on this dataset. PC1 (horizontal axis) captures the primary variation in patient age and service unit, while PC2 (vertical axis) captures variation in demographic-administrative profile, namely gender and insurance type.

The four clusters are distributed diagonally across the plot. Cluster 1 (green) occupies the upper-right region, distinctly separated from the others owing to the unique demographic profile of MCH patients. Cluster 0 (red) dominates the central area with a dense, compact distribution consistent with its status as the largest segment. Cluster 3 (purple) stretches horizontally along the lower area, reflecting wide age variation within a

homogeneous service profile. Cluster 2 (blue) appears more diffusely spread on the left, mirroring its high intra-cluster heterogeneity.

The partial overlap between Clusters 0 and 3 in the central zone is analytically expected: both share identical non-age attributes (BPJS insurance, returning visits, general clinic), and age, their sole differentiator, is a continuous variable that yields no sharp boundary after PCA projection. Despite this, the overall spatial separation across all four clusters is consistent with the Silhouette Score of 0.6558, confirming that PCA preserved the meaningful grouping structure of the original feature space.

D. Characteristics of Each Cluster

Table 6 summarizes the dominant characteristics of the four clusters produced in the final experiment.

Table 6. Dominant Characteristics of Each Cluster in the Final Experiment

Cluster	Insurance	Type of Visit	Clinic/Room	Age Group	Case Type
0	BPJS	Long-term	General	Adults (± 41 years)	New
1	BPJS	Long-term	KIA	Young (± 18 years)	New
2	General (Independent)	New/Existing	Emergency & Dental	Various (± 42 years)	New
3	BPJS	Existing	General	Elderly (± 64 years)	New

Cluster 0 - Adult BPJS Patients with Chronic Diseases

This is the largest cluster, comprising 3,490 patients, all BPJS members with returning-patient status (97%) accessing the general clinic. Women account for 73% of the group (mean age: ± 41 years). Diagnostically, the cluster is dominated by noncommunicable diseases: essential hypertension (36%), type 2 diabetes mellitus, and, notably, schizophrenia as a significant diagnosis, indicating a chronic-disease segment requiring long-term monitoring.

Cluster 1 - Maternal and Child Health Patients

This cluster comprises 1,295 patients (79% female) accessing MCH and maternity services. The low mean age (± 18 years) reflects the inclusion of infants and children brought by their parents. Approximately 45% of visits involve pregnancy, delivery, or infant checkups, while 24% of patients are self-paying, the second-highest proportion across all clusters, suggesting a segment of pregnant women not yet covered under JKN.

Cluster 2 - Unplanned Care Patients and Self-Referred General Checkups

With 1,422 patients spanning ages 0–94 (mean: ± 34 years), this cluster has the most balanced gender split (52% female, 47% male) and the highest proportion of self-paying patients at 72%. Visit types are distributed across general clinics (63%), emergency (22%), dental (12%), and inpatient (3%). Most visits are elective or preventive in nature; however, pulmonary tuberculosis appears as a notable diagnosis, requiring specific attention from an infectious disease control perspective.

Cluster 3 - Elderly BPJS Patients with High Comorbidity (23.3%).

This cluster consists of 1,885 elderly patients (mean age: ± 64 years), all BPJS members with a near-total returning-visit rate (99%) and relatively balanced gender distribution (54% female, 45% male). Its diagnostic profile carries the heaviest disease burden: essential hypertension (34%), congestive heart failure (5%), type 2 diabetes (5%), and dual hypertension-diabetes comorbidity. Over 68% of visits are related to degenerative noncommunicable conditions, positioning Cluster 3 as the highest-risk, most clinically complex segment.

Cross-Cluster Patterns

Several cross-cluster patterns merit particular attention from a facility management perspective. First, essential hypertension is the leading diagnosis in three of the four clusters (0, 1, and 3), identifying it as the highest-burden condition at Puskesmas Kupang and indicating the need for a comprehensive, cross-segment hypertension control program. Second, the overall gender distribution is markedly skewed, 66% of patients are female, partly reflecting high MCH utilization (Cluster 1) and women's greater engagement with both preventive and curative health services. Third, while 84% of patients hold BPJS membership, the 16% self-paying patients are concentrated in Cluster 2, pointing to a population segment not yet covered by JKN and potentially more exposed to access barriers. Fourth, 95% of all visits across clusters are returning visits, indicating that Puskesmas Kupang primarily serves patients requiring long-term care continuity rather than episodic or first-time attendees.

IV. CONCLUSION

This study confirms that K-Means effectively identifies structured patient visit patterns at Puskesmas Kupang, yielding four coherent and operationally interpretable segments. Among the three feature engineering configurations tested, the six-feature setup excluding temporal attributes produced the strongest model, a Silhouette Score of 0.6558 at $k=4$, with two-component PCA incorporated to mitigate redundancy across the mixed-type feature set prior to clustering. The four clusters capture meaningfully distinct patient profiles: adult BPJS enrollees with chronic conditions accessing the general clinic; reproductive-age patients utilizing maternal and child health services; patients with unplanned care needs and predominantly self-pay insurance; and elderly patients with degenerative comorbidities and high return-visit frequency, a structured patient profile that manual recording processes cannot readily yield.

The findings carry direct value across three dimensions. Practically, the segmentation enables Puskesmas Kupang to align medication stocking, staffing schedules, and service program design with the actual clinical and demographic composition of each patient segment, including Prolanis prioritization for Clusters 0 and 3, midwifery reinforcement for Cluster 1, and affordable screening outreach for Cluster 2. Methodologically, the iterative three-experiment protocol provides empirical evidence that feature selection substantively affects clustering outcomes, while the PCA-integrated pipeline offers a replicable framework for similar analyses at other primary care facilities. Academically, the study contributes an administrative-attribute perspective on patient segmentation, distinct from the disease-type focus that dominates comparable Indonesian research, and generates descriptive evidence on comorbidity patterns, insurance coverage gaps, and care continuity trends at the primary care level.

This study is limited to data from a single facility over nine months, which warrants caution before generalizing the findings, and the cluster labels remain descriptive pending clinical validation with local healthcare personnel. Future work should compare K-Means against alternative algorithms such as K-Medoids or DBSCAN across multi-facility datasets, and examine the feasibility of embedding the segmentation output into real-time health management information systems as an operational decision-support feature.

REFERENCES

- [1] Q. Cheng *et al.*, “Determinants of healthcare utilization under the Indonesian national health insurance system – a cross-sectional study,” *BMC Health Serv. Res.*, vol. 25, no. 1, pp. 48, Jan. 2025, doi: 10.1186/s12913-024-11951-8.
- [2] S. Maulia, B. S. Ginting, and A. Sihombing, “IMPLEMENTASI DATA MINING PENGELOMPOKAN JENIS PENYAKIT PASIEN MENGGUNAKAN METODE CLUSTERING (STUDI KASUS: PUSKESMAS SAMBIREJO),” *Jurnal Informatika Kaputama*, vol. 5, no. 1, 2021.
- [3] L. Mayola, M. Hafizhb, and H. S. Syahputra, “Klasterisasi Rumah Sakit berdasarkan Kunjungan Pasien menggunakan Algoritma K-Means: Data 2019-2023,” *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 7, no. 1, pp. 15–21, Jan. 2025, doi: 10.47233/jteksis.v7i1.1703.
- [4] M. H. Ziqrullah, S. D. Purnamasari, and I. Z. Yadi, “Clustering Data Penyakit Pasien Pada Puskesmas Mulyaguna Menggunakan Algoritma K-Means,” *Jurnal Media Informatika*, vol. 7, no. 1, pp. 608-616, Des. 2024.
- [5] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: A comprehensive survey and performance evaluation,” *Electronics*, vol. 9, no. 8, Art. no. 1295, Aug. 2020, doi: 10.3390/electronics9081295.
- [6] W. C. Yang, J. P. Lai, Y. H. Liu, Y. L. Lin, H. P. Hou, and P. F. Pai, “Using Medical Data and Clustering Techniques for a Smart Healthcare System,” *Electronics*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/electronics13010140.
- [7] N. L. Hasanah, A. A. Ningrum, and F. D. Marleny, “PENERAPAN K-MEANS CLUSTERING MENENTUKAN LOKASI PROMOSI PENERIMAAN MAHASISWA BARU UNIVERSITAS MUHAMMADIYAH BANJARMASIN,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 12, no. 2, Jun. 2025, doi: 10.35957/jatisi.v12i2.11488.
- [8] M. Noval, F. D. Marleny, and W. Windarsyah, “Implementasi Algoritma K-Means Untuk Analisis Pola Penjualan Pada Toko Monisa,” *Jurnal Media Informatika*, vol. 6, no. 3, Aug. 2025.
- [9] A. Aljohani, “Optimizing Patient Stratification in Healthcare: A Comparative Analysis of Clustering Algorithms for EHR Data,” *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, pp. 173, Jul. 2024, doi: 10.1007/s44196-024-00568-8.
- [10] A. H. Lutfiannisa, M. Maimunah, and P. Sukmasetya, “Clustering Data Pasien Berdasarkan Usia di Puskesmas Menerapkan Metode K-Means,” *Journal of Information System Research (JOSH)*, vol. 5, no. 2, pp. 639–647, Jan. 2024, doi: 10.47065/josh.v5i2.4755.
- [11] B. H. Prakoso, E. Rachmawati, D. R. P. Mudiono, V. Vestine, and G. E. J. Suyoso, “Klasterisasi Puskesmas dengan K-Means Berdasarkan Data Kualitas Kesehatan Keluarga dan Gizi Masyarakat,” *Jurnal Buana Informatika*, vol. 14, no. 01, pp. 60–68, Apr. 2023, doi: 10.24002/jbi.v14i01.7105.
- [12] D. Z. Ardini and R. Rhodiyah, “Analisis Pola Penyakit Kronis pada Lansia Menggunakan K-Means Clustering di Puskesmas Kelurahan Sempur Barat,” *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, vol. 6, no. 3, pp. 1842–1853, Sep. 2025, doi: 10.63447/jimik.v6i3.1595.
- [13] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y.-I. Cho, “Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms,” *Mathematics*, vol. 12, no. 16, p. 2553, Aug. 2024, doi: 10.3390/math12162553.
- [14] P. P. Pane, Y. Ramadhan Nasution, and Mhd. Furqan, “Implementasi Data Mining dengan K-Means Clustering untuk Memprediksi Pengadaan Obat,” *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 2, pp. 286–296, Feb. 2024, doi: 10.47065/josyc.v5i2.4920.
- [15] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, “A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, no. 1, pp. 31, Dec. 2021, doi: 10.1186/s13638-021-01910-w.
- [16] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana, “Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta,” *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.