Application of the K-Means Algorithm for Patient Clustering Based on Length of Hospital Stay Characteristics at Kutacane Hospital

Maisya Fitri Anugrah, Muhammad Syahputra Novelan

Abstract

Length of hospital stay is an important indicator for evaluating healthcare service efficiency, patient care management, and hospital resource utilization. Variations in hospitalization duration are influenced by several demographic and clinical characteristics, including patient age, gender, medical diagnosis, disease category, and individual health conditions. This study aims to apply the K-Means Clustering algorithm to group patients based on length of hospital stay characteristics at Kutacane Hospital. The research employs a quantitative approach using data mining techniques to identify similarities and differences among inpatient records. The research methodology consists of six stages: data collection, data preprocessing, feature selection and data normalization, K-Means clustering implementation, cluster evaluation, and cluster interpretation. The selected attributes include patient age, gender, medical diagnosis, disease category, admission date, discharge date, and length of hospital stay, depending on the availability and quality of medical records. The Elbow method is utilized to determine the optimal number of clusters by analyzing the Within-Cluster Sum of Squares (WCSS), while the Silhouette Score is employed to evaluate cluster cohesion and separation. The clustering process is designed to identify distinct patient groups characterized by relatively short, moderate, and long hospitalization durations. The resulting cluster profiles are expected to provide information regarding patient distribution and hospitalization patterns. Furthermore, this approach is expected to support hospital management in evaluating inpatient services, planning bed capacity, improving healthcare resource allocation, and supporting data-driven operational decision-making. This study highlights the potential of unsupervised machine learning techniques to transform hospital medical records into meaningful analytical information for improving healthcare service management.

***Keywords**:* K-Means Clustering, Patient Segmentation, Length of Hospital Stay, Data Mining, Machine Learning, Hospital Management.

.

[1]Maisya Fitri Anugrah
[1]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: maisyafitrianugrah@gmail.com [1]

[2]Muhammad Syahputra Novelan
[2]Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: putranovelan@dosen.pancabudi.ac.id [2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

The development of information technology in the healthcare sector has encouraged hospitals to utilize patient data as an important resource for improving healthcare services and supporting strategic decision-making. Hospitals generate large volumes of medical and administrative data every day, including patient demographic information, clinical characteristics, diagnoses, admission dates, discharge dates, and length of hospital stay. These data contain valuable information that can be analyzed to identify patterns of patient care, understand variations in hospitalization duration, and evaluate the efficiency of hospital services. However, the availability of large amounts of patient data does not automatically provide meaningful insights without appropriate data processing and analytical methods. Therefore, the implementation of data mining techniques is essential to transform raw hospital data into useful information that can support healthcare management and operational planning [1] [2].

Length of hospital stay is one of the important indicators used to evaluate hospital service performance and patient care efficiency. The duration of hospitalization may vary depending on several factors, including patient age, medical diagnosis, clinical condition, disease severity, treatment procedures, and response to medical treatment[3] . Patients with similar medical conditions may experience different hospitalization durations due to variations in individual health characteristics and treatment requirements. A long hospital stay may indicate complex medical conditions, complications, or the need for continuous medical monitoring, while a shorter hospitalization period may reflect relatively stable conditions and effective treatment responses. Understanding these variations is important for hospitals because hospitalization duration directly influences bed occupancy, healthcare resource utilization, treatment costs, and the availability of medical facilities for other patients [4] [5].

One of the challenges faced by hospitals is how to identify groups of patients with similar characteristics based on their length of hospital stay. Conventional data analysis often focuses on descriptive statistics and general patient reports, which may not sufficiently reveal hidden patterns within complex healthcare datasets[6] . Patient characteristics are multidimensional, involving demographic attributes such as age and gender, as well as clinical attributes such as diagnosis, disease category, and treatment-related information. Without an appropriate analytical approach, identifying meaningful patient groups based on these variables can become difficult. Therefore, clustering techniques can provide an alternative solution by grouping patients according to similarities in their characteristics without requiring predefined class labels. This approach enables hospitals to discover patient profiles and understand the distribution of hospitalization patterns more systematically [7] [8].

The K-Means Clustering algorithm is one of the widely used unsupervised machine learning methods for grouping data based on similarities between observations. The algorithm works by dividing a dataset into a predetermined number of clusters, where each observation is assigned to the cluster with the nearest centroid. The clustering process involves selecting the number of clusters, determining initial centroids, calculating distances between data points and centroids, assigning observations to the nearest clusters, and updating centroid positions until the clustering results converge. In the context of hospitalization data, K-Means can be applied to identify patient groups based on demographic and clinical characteristics, particularly their length of hospital stay. The resulting clusters can represent different hospitalization profiles, such as patients with relatively short, moderate, and long stays, depending on the distribution and characteristics of the analyzed dataset. Appropriate data preprocessing, numerical encoding, and feature scaling are important to ensure that the clustering results are not dominated by variables with larger numerical ranges [9] [10].

Based on these considerations, this study aims to apply the K-Means Clustering algorithm to group patients based on length of hospital stay characteristics at Kutacane Hospital. The analysis focuses on identifying patient clusters using relevant demographic and clinical

variables to discover patterns and similarities among hospitalized patients. The research is expected to provide an overview of patient distribution across different hospitalization groups and identify characteristics associated with variations in length of stay. Furthermore, the clustering results are expected to provide useful information for hospital management in supporting inpatient service evaluation, optimizing bed utilization, planning healthcare resources, and improving operational efficiency. This study also contributes to the application of machine learning in healthcare data analytics by demonstrating how unsupervised learning techniques can transform hospital patient records into structured information that supports more data-driven healthcare management[11] .

Literature Review

Data mining is a process of discovering valuable information, patterns, and relationships from large datasets through computational and statistical techniques. In the healthcare sector, data mining has become an important approach for analyzing patient records, identifying disease patterns, evaluating healthcare services, and supporting clinical and administrative decision-making[12] . Hospital information systems generate various types of data, including demographic information, medical diagnoses, treatment histories, and hospitalization duration. Analyzing these data can provide insights into patient characteristics and healthcare service utilization. One of the important applications of data mining in hospital management is patient segmentation, which aims to group patients based on similarities in their demographic and clinical characteristics. Through this process, hospitals can identify distinct patient profiles and obtain a better understanding of variations in healthcare needs and hospitalization patterns [13] [14].

Clustering is an unsupervised machine learning technique used to group data objects based on their similarities without requiring predefined class labels. One of the most widely implemented clustering algorithms is K-Means, which partitions a dataset into a specified number of clusters by minimizing the distance between data observations and their respective cluster centroids. The algorithm generally applies Euclidean distance to measure similarities between data points and cluster centers. Its main stages include determining the number of clusters (K), initializing centroids, calculating distances, assigning data points to the nearest centroids, and updating centroid positions iteratively until the clustering process reaches convergence. K-Means is commonly applied in healthcare analytics because of its computational efficiency, relatively simple implementation, and ability to identify hidden patterns in patient datasets. However, the quality of clustering results depends on the selection of the number of clusters, data preprocessing, feature selection, and appropriate normalization of numerical variables[15] [16].

Patient segmentation based on length of hospital stay is an important analytical approach for understanding variations in inpatient healthcare utilization. Length of stay refers to the total period a patient remains hospitalized, generally calculated from the admission date to the discharge date[17] . This variable can be influenced by several factors, including age, gender, medical diagnosis, disease severity, clinical conditions, and treatment requirements. By integrating demographic and clinical characteristics into the K-Means algorithm, patients can be grouped into clusters representing different hospitalization profiles. These clusters may reveal patterns associated with relatively short, moderate, and long hospitalization durations. The resulting information can support hospital management in evaluating inpatient service efficiency, planning bed capacity, allocating medical resources, and identifying patient groups

that require more intensive healthcare services. Therefore, the application of K-Means Clustering provides an analytical foundation for transforming hospital patient records into meaningful information to support data-driven decision-making at Kutacane Hospital [18].

Research Methodology

This study employs a quantitative research approach using data mining techniques to analyze and group patient characteristics based on length of hospital stay at Kutacane Hospital. The primary method used is the K-Means Clustering algorithm, which aims to identify groups of patients based on similarities in their demographic and clinical characteristics. The research dataset consists of inpatient medical records obtained from Kutacane Hospital. The attributes analyzed are determined according to the availability and quality of the collected data, potentially including patient age, gender, medical diagnosis, disease category, admission date, discharge date, and length of hospital stay. Length of hospital stay is calculated based on the difference between the patient's discharge date and admission date. The application of K-Means is intended to identify patient groups with similar hospitalization characteristics without relying on predefined class labels[19], [20] .

The research procedure is organized into six systematic stages, beginning with data collection and ending with the interpretation of clustering results. Each stage is designed to ensure that the collected patient data are properly processed, the clustering algorithm is implemented consistently, and the resulting groups can be evaluated objectively. Data preprocessing, feature selection, categorical encoding, and normalization are conducted to improve data quality and ensure that each selected attribute can contribute appropriately to the clustering process. The K-Means algorithm is subsequently implemented to generate patient clusters, while the Elbow method and Silhouette Score are used to determine the appropriate number of clusters and evaluate clustering quality. The overall research methodology is presented in six stages, as described below [21], [22] .



Figure 1. Research Methodology

1. Data Collection
   The data collection stage involves obtaining inpatient medical records from Kutacane Hospital as the primary dataset for this study. The collected records provide the empirical foundation for identifying patterns and similarities among patients based on their hospitalization characteristics. Data collection focuses on obtaining relevant demographic and clinical attributes, including patient age, gender, medical diagnosis, disease category, admission date, discharge date, and length of hospital stay. The selection of these attributes considers their relevance to the research objectives and their availability within the hospital information system. Patient records are subsequently examined to ensure that the collected information is sufficiently complete and suitable for further data processing and clustering analysis. Patient confidentiality is maintained throughout the research process by excluding personally identifiable information from the analytical dataset.
2. Data Preprocessing
   Data preprocessing is conducted to improve the quality, consistency, and reliability of the patient dataset before applying the K-Means Clustering algorithm. This stage includes examining missing values, identifying duplicate records, correcting inconsistent data formats, and detecting potentially invalid values. Particular attention is given to admission and discharge dates because these attributes are used to calculate the length of hospital stay.

Records containing incomplete or inconsistent hospitalization information are handled using appropriate data-cleaning procedures. Furthermore, irrelevant attributes, such as patient names and identification numbers, are excluded from the analytical dataset to maintain data privacy and prevent unnecessary variables from influencing the analysis. This preprocessing stage is essential because inaccurate or inconsistent data may affect the clustering process and reduce the quality of the resulting patient groups.

3. Feature Selection and Data Normalization
   Data preprocessing is conducted to improve the quality, consistency, and reliability of the patient dataset before applying the K-Means Clustering algorithm. This stage includes examining missing values, identifying duplicate records, correcting inconsistent data formats, and detecting potentially invalid values. Particular attention is given to admission and discharge dates because these attributes are used to calculate the length of hospital stay. Records containing incomplete or inconsistent hospitalization information are handled using appropriate data-cleaning procedures. Furthermore, irrelevant attributes, such as patient names and identification numbers, are excluded from the analytical dataset to maintain data privacy and prevent unnecessary variables from influencing the analysis. This preprocessing stage is essential because inaccurate or inconsistent data may affect the clustering process and reduce the quality of the resulting patient groups.
4. K-Means Clustering Implementation
   At this stage, the K-Means Clustering algorithm is implemented to group patients according to similarities in their demographic and clinical characteristics. The algorithm begins by determining the number of clusters and initializing the cluster centroids. Each patient record is subsequently assigned to the nearest centroid based on the similarity of its selected attributes. After the assignment process, the centroid positions are recalculated based on the average characteristics of the patients within each cluster. These iterations continue until the cluster memberships stabilize or the predefined stopping criterion is achieved. The resulting clusters are expected to represent different patient hospitalization profiles, particularly groups characterized by relatively short, moderate, and long lengths of hospital stay. Through this process, the algorithm identifies similarities among patients and provides an overview of the distribution of hospitalization characteristics at Kutacane Hospital.
5. Cluster Evaluation
   The clustering results are evaluated to determine the quality, consistency, and suitability of the identified patient groups. The Elbow method is used to examine the relationship between the number of clusters and the Within-Cluster Sum of Squares (WCSS). This method helps identify an appropriate number of clusters by observing the point at which the reduction in WCSS begins to decrease significantly. Furthermore, the Silhouette Score is applied to evaluate the similarity of patients within the same cluster compared with patients assigned to different clusters. Higher Silhouette Score values generally indicate better-defined and more clearly separated clusters, while values close to zero indicate overlapping groups. Negative values may suggest inappropriate cluster assignments. The evaluation results provide a quantitative basis for selecting the most suitable clustering configuration and determining whether the resulting patient groups demonstrate meaningful separation.
6. Cluster Interpretation and Reporting
   The final stage involves interpreting the characteristics of each patient cluster based on the distribution of records and the average values of the selected demographic and clinical attributes. The analysis focuses on identifying the number of patients in each cluster, the average length of hospital stay, age distribution, diagnosis characteristics, and other relevant variables. Each cluster is subsequently characterized according to its dominant attributes, allowing the identification of patient groups with relatively short, moderate, or long hospitalization durations. The findings are presented through statistical tables, graphical visualizations, and descriptive explanations to facilitate interpretation of the clustering

results. The identified patterns are then discussed in relation to inpatient service utilization and hospital operational conditions. The results are expected to provide useful information for Kutacane Hospital management in evaluating patient hospitalization patterns, supporting bed capacity planning, optimizing healthcare resource allocation, and improving the effectiveness of inpatient services.

Results

This section presents the analytical framework for applying the K-Means Clustering algorithm to group inpatient characteristics based on length of hospital stay at Kutacane Hospital. The analysis focuses on examining the structure of the patient dataset, identifying relevant demographic and clinical attributes, determining the optimal number of clusters, and interpreting the characteristics of the resulting patient groups. The variables considered include patient age, gender, medical diagnosis, disease category, and length of hospital stay, depending on the availability of medical record data. The analysis is designed to identify similarities and differences among patients and provide a structured understanding of hospitalization patterns.

Before implementing the K-Means algorithm, the collected patient records are subjected to preprocessing to ensure data consistency and reliability. The process includes missing-value identification, duplicate-record detection, categorical encoding, feature selection, and numerical normalization. These procedures are necessary to ensure that the selected attributes contribute appropriately to the clustering process. Subsequently, the Elbow method and Silhouette Score are applied to evaluate candidate cluster configurations and identify the most appropriate number of clusters.

To provide a systematic presentation of the analytical results, this section is organized into three main tables covering data preparation, cluster configuration evaluation, and patient cluster profiles. Graphical visualizations are also required to illustrate the Elbow curve and Silhouette Score distribution. Since the actual Kutacane Hospital dataset has not been provided for this document, numerical results are presented as fields to be completed rather than fabricated experimental findings. All final values and graphs must be calculated from the actual patient records before publication.

Table 1. Patient Data Preparation and Feature Summary

No.	Processing Item	Purpose	Result to Report
1	Initial dataset	Identify the total number of patient records	Actual record count
2	Patient identification	Identify individual records	Anonymous patient codes
3	Missing-value assessment	Identify incomplete patient information	Missing-value count
4	Missing-value treatment	Handle incomplete observations	Treatment method
5	Duplicate detection	Prevent repeated patient records	Duplicate record count
6	Age validation	Ensure patient age is within a valid range	Valid age records
7	Gender classification	Standardize demographic categories	Male and female distribution
8	Diagnosis classification	Group medical diagnoses into relevant categories	Diagnosis categories
9	Admission date validation	Check hospitalization entry dates	Valid admission records
10	Discharge date validation	Verify patient discharge information	Valid discharge records

11	Length of stay calculation	Determine hospitalization duration	Length of stay in days
12	Feature selection	Retain relevant demographic and clinical variables	Final selected attributes
13	Categorical encoding	Transform categorical attributes into numerical representations	Encoded dataset
14	Data normalization	Standardize numerical measurement scales	Normalized dataset
15	Final analysis matrix	Prepare data for K-Means implementation	Final rows and columns

Table 1 presents the 15 data preparation activities required before implementing the K-Means Clustering algorithm. The preparation process begins with examining the number and completeness of patient records and continues with validating demographic and clinical attributes. Particular attention is given to admission dates and discharge dates because these variables determine the length of hospital stay. The calculated hospitalization duration becomes one of the important attributes in identifying patient groups. Furthermore, categorical encoding and numerical normalization are performed to ensure that the selected variables can be processed consistently. The final analysis matrix represents the dataset that will be used in the clustering process.

Table 2. Candidate Cluster Configuration and Evaluation

No.	Number of Clusters (K)	WCSS	Silhouette Score	Assessment
1	2	Calculate	Calculate	Evaluate initial patient segmentation
2	3	Calculate	Calculate	Examine three-group segmentation
3	4	Calculate	Calculate	Evaluate additional patient profiles
4	5	Calculate	Calculate	Examine cluster separation
5	6	Calculate	Calculate	Evaluate possible over-segmentation
6	7	Calculate	Calculate	Compare cluster consistency
7	8	Calculate	Calculate	Examine additional cluster structures
8	Selected K	Actual result	Actual result	Determine optimal configuration
9	Distance metric	Euclidean	N/A	Maintain consistent distance measurement
10	Initialization	Record setting	N/A	Document centroid initialization
11	Iteration count	Record setting	N/A	Document clustering iterations
12	Convergence	Record setting	N/A	Verify clustering stability
13	Feature scaling	Normalized data	N/A	Ensure comparable variable scales

14	Cluster stability	Repeated evaluation	Optional metric	Examine consistency across runs
15	Evaluation decision	Final WCSS	Final score	Justify the selected cluster number

Table 2 provides a comparison framework for evaluating different cluster configurations using the K-Means algorithm. The candidate values of K range from 2 to 8 to examine how changes in the number of clusters influence the grouping of patient records. WCSS is used to observe the reduction in within-cluster variation, while the Silhouette Score evaluates the separation and consistency of the resulting groups. A suitable cluster configuration should demonstrate a reasonable balance between compactness, separation, and practical interpretability. The final number of clusters must be determined from the actual computational results rather than assuming that three clusters will automatically provide the best segmentation.

Table 3. Patient Cluster Profile and Interpretation

No.	Cluster Profile Attributes	Cluster 1	Cluster 2	Cluster 3
1	Number of patients	Calculate	Calculate	Calculate
2	Percentage of patients	Calculate	Calculate	Calculate
3	Average patient age	Calculate	Calculate	Calculate
4	Median patient age	Calculate	Calculate	Calculate
5	Male patient distribution	Calculate	Calculate	Calculate
6	Female patient distribution	Calculate	Calculate	Calculate
7	Average length of stay	Calculate	Calculate	Calculate
8	Median length of stay	Calculate	Calculate	Calculate
9	Minimum length of stay	Calculate	Calculate	Calculate
10	Maximum length of stay	Calculate	Calculate	Calculate
11	Most frequent diagnosis	Calculate	Calculate	Calculate
12	Dominant disease category	Calculate	Calculate	Calculate
13	Average normalized age	Calculate	Calculate	Calculate
14	Cluster centroid characteristics	Calculate	Calculate	Calculate
15	Dominant hospitalization pattern	Interpret	Interpret	Interpret

Table 3 presents the framework for interpreting the characteristics of each patient cluster generated by the K-Means algorithm. The analysis includes the number and percentage of patients, average age, gender distribution, hospitalization duration, and the most frequent medical diagnoses within each group. These attributes are used to identify the distinguishing characteristics of each cluster and understand the distribution of patient hospitalization patterns. The average and median lngth of stay are particularly important because they provide an overview of the hospitalization duration within each group. The final cluster descriptions should be based on the actual centroid values and statistical distributions obtained from the clustering process.

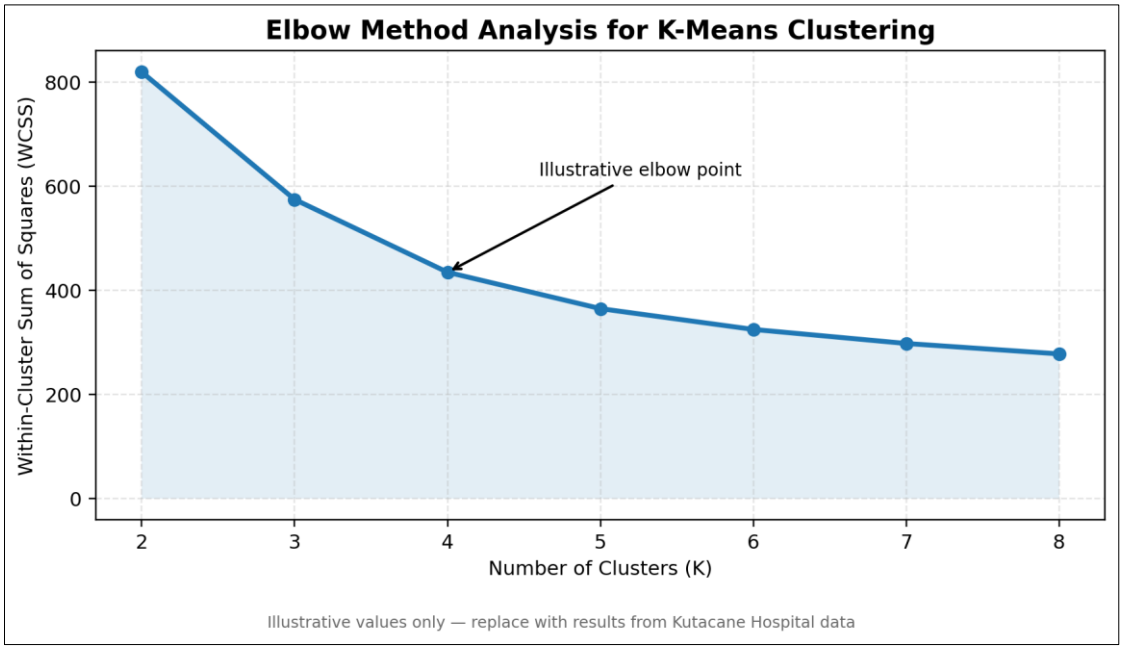


Figure 2. Elbow Method Analysis for K-Means Clustering

Figure 2 should display the relationship between the number of clusters (K) and the Within-Cluster Sum of Squares (WCSS). The curve is examined to identify the point at which adding further clusters produces progressively smaller reductions in within-cluster variation. This point provides an indication of a suitable cluster configuration. The graph must be generated using WCSS values calculated from the actual Kutacane Hospital dataset.

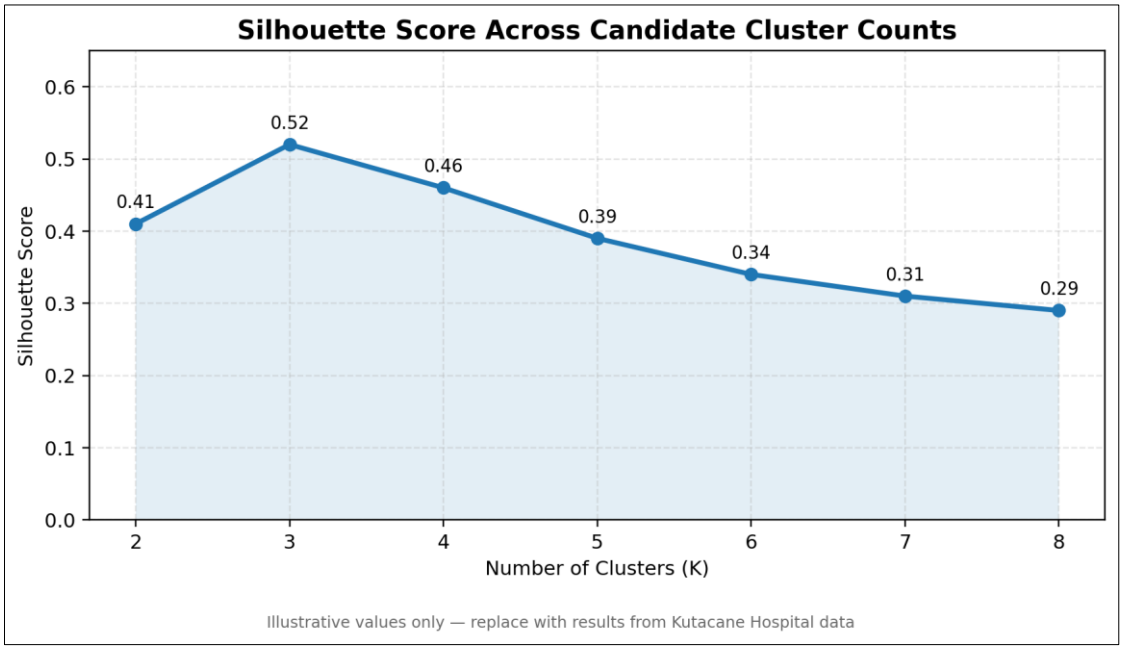


Figure 3. Silhouette Score Analysis for K-Means Clustering

Figure 3 should present the Silhouette Score for each candidate number of clusters. Higher values generally indicate that patients are more similar to members of their assigned cluster than to patients in neighboring clusters. Values close to zero may indicate overlapping groups, while negative values may indicate inappropriate assignments. The final cluster

configuration should consider this evaluation together with the Elbow method and the interpretability of patient profiles

Conclusion

This study presents the application of the K-Means Clustering algorithm to group inpatient characteristics based on length of hospital stay at Kutacane Hospital. The proposed approach consists of several stages, including data collection, preprocessing, feature selection, categorical encoding, data normalization, clustering implementation, and cluster evaluation. By utilizing demographic and clinical attributes, such as patient age, gender, medical diagnosis, and length of hospital stay, the K-Means algorithm provides an approach for identifying groups of patients with similar hospitalization characteristics. The Elbow method and Silhouette Score are employed to determine the appropriate number of clusters and evaluate the quality of the resulting patient segmentation.

The application of K-Means Clustering is expected to provide meaningful insights into patient hospitalization patterns and differences in healthcare service utilization. The resulting cluster profiles can support hospital management in understanding patient characteristics, evaluating inpatient service efficiency, planning bed capacity, and optimizing healthcare resource allocation. However, the reliability of the clustering results depends on the quality of the medical records, the selection of relevant attributes, and the evaluation of cluster performance using actual patient data. Future research should prioritize validating the proposed model using real inpatient datasets, examining the clinical relevance of each cluster with healthcare professionals, and comparing K-Means with alternative clustering algorithms to obtain more accurate and meaningful patient segmentation results.

References

[1] H. W. Dhany, F. Izhari, and S. Komputer, “Jurnal ICT : Information and Communication Technologies, 15 (2) (2024) 48-54 Prediction analysis condition animal use algorithm (SVM+KNN),” 2024. [Online]. Available: www.ejournal.marqchainstitute.or.id/index.php/JICT

[2] Z. Sitorus, D. Prayogi, M. A. Rizko, A. G. Suteja, and M. R. Harahap, “Implementation of the Insertion Sort Algorithm to Sort Positive Integers in Ascending Order Using Flowgorithm,” *Journal of Information Technology, computer science and Electrical Engineering (JITCSE)*, vol. 1, no. 3, pp. 323–328, 2024, doi: 10.30596/jitcse.

[3] L. Cao, “On Machine Learning and Public Administration,” *Frontiers in Management Science*, vol. 1, no. 1, Aug. 2022, doi: 10.56397/fms.2022.08.01.

[4] K. Khairul, A. H. Nasyuha, A. Ikhwan, M. H. Aly, and A. Ahyanuardi, “Implementation of Multiple Linear Regression to Estimate Profit on Sales of Screen Printing Equipment,” *JURNAL INFOTEL*, vol. 15, no. 2, pp. 55–61, Jun. 2023, doi: 10.20895/infotel.v15i2.934.

[5] M. S. Novelan and S. Aryza, “OPTIMIZATION CVRP WITH MACHINE LEARNING FOR IMPROVED CLASSIFICATION OF IMBALANCED DATA FOOD DISTRIBUTION,” *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 4, pp. 917–925, Jun. 2025, doi: 10.33480/jitk.v10i4.6467.

[6] N. Sekar Ramadhanti and W. Ananta Kusuma, "OPTIMASI DATA TIDAK SEIMBANG PADA INTERAKSI DRUG TARGET DENGAN SAMPLING DAN ENSEMBLE SUPPORT VECTOR MACHINE," vol. 7, no. 6, 2020, doi: 10.25126/jtiik.202072857.

[7] R. Farta Wijaya, F. Kurniawan, R. R. Putrai, A. Alvin, F. Sains, and D. Teknologi, "OPTIMIZING THE MANAGEMENT OF VILLAGE ACTIVITIES THROUGH INFORMATION SYSTEMS FOR TRANSPARENCY AND ACCOUNTABILITY IN PERTUMBUKAN VILLAGE, WAMPU DISTRICT," 2024.

[8] M. M. Maulana, A. I. Suroso, Y. Nurhadryani, and K. B. Seminar, "The Smart Governance Framework and Enterprise System's Capability for Improving Bio-Business Licensing Services," *Informatics*, vol. 10, no. 2, Jun. 2023, doi: 10.3390/informatics10020053.

[9] A. Putera Utama Siahaan, N. Azizah Harahap, R. Yuni Simanullang, and P. Wanny, "Analysis of Inpatient Data Using Cluster Analysis on Simulation Dataset," *Bulletin of Information Technology (BIT)*, vol. 6, no. 1, pp. 33–39, 2025, doi: 10.47065/bit.v5i2.1830.

[10] M. S. Novelan, S. Efendi, P. Sihombing, and H. Mawengkang, "VEHICLE ROUTING PROBLEM OPTIMIZATION WITH MACHINE LEARNING IN IMBALANCED CLASSIFICATION VEHICLE ROUTE DATA," *Eastern-European Journal of Enterprise Technologies*, vol. 5, no. 3(125), pp. 49–56, 2023, doi: 10.15587/1729-4061.2023.288280.

[11] M. Iqbal and S. Efendi, "Data-Driven Approach for Credit Risk Analysis Using C4.5 Algorithm," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 14, no. 1, pp. 11–20, May 2023, doi: 10.21512/comtech.v14i1.8243.

[12] L. Zhu, Y. Zhang, Z. Pan, R. Wang, S. Kwong, and Z. Peng, "Binary and multi-class learning based low complexity optimization for HEVC encoding," *IEEE Transactions on Broadcasting*, vol. 63, no. 3, pp. 547–561, Sep. 2017, doi: 10.1109/TBC.2017.2711142.

[13] W. Cetera, W. Gogołek, A. Żołnierski, and D. Jaruga, "Potential for the use of large unstructured data resources by public innovation support institutions," *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00610-6.

[14] G. Rajeshkumar *et al.*, "Smart office automation via faster R-CNN based face recognition and internet of things," *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100719.

[15] M. Overton, S. Larson, L. J. Carlson, and S. Kleinschmit, "Public data primacy: the changing landscape of public service delivery as big data gets bigger," *Global Public Policy and Governance*, vol. 2, no. 4, pp. 381–399, Dec. 2022, doi: 10.1007/s43508-022-00052-z.

[16] J. Ilmiah, E. & Erudisi, A. Zahara, M. I. Kabullah, and R. E. Putera, "JIEE: Effectiveness of the OSSRBA (Online Single Submission Risk Based Approach) System in Business Licensing Services in Payakumbuh City DPMPTSP," vol. 3, pp. 22–40, 2023.

[17] L. Von Rueden *et al.*, "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 614–633, Jan. 2023, doi: 10.1109/TKDE.2021.3079836.

[18] W. Jin, "Challenges and Innovative Countermeasures Faced by Public Administration in the Context of Big Data and Internet of Things," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/8949365.

[19] S. D. A. Bujang *et al.*, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[20] M. Genkin, "Zero-Shot Machine Learning Technique for Classification of Multi-User Big Data Workloads," in *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 5046–5055. doi: 10.1109/BigData50022.2020.9378023.

[21] P. Hasan Putra, M. Syahputra Novelan, and M. Rizki, "ANALYSIS K-NEAREST NEIGHBOR METHOD IN CLASSIFICATION OF VEGETABLE QUALITY BASED ON COLOR."

[22] Ill.) IEEE Technology and Engineering Management Society. Conference (2018 : Evanston, Institute of Electrical and Electronics Engineers, and IEEE Technology and Engineering Management Society, *TEMSCON 2018 : 2018 IEEE Technology & Engineering Management Conference (TEMSCON).*