JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

Synthetic Data Pattern Simulation of Patient Care Journey Using K-Means Clustering

Arjon Samuel Sitio[1*], Richard Parlindungan[2], Anita Sinaga[3]
[1,2]Information Systems, Faculty of Science and Technology, Tjut Nyak Dhien University, Indonesia
[3]Information Technology, STMIK Pelita Nusantara, Indonesia
[1*]arjonsitio@yahoo.com, [2]richsparlin0@gmail.com, [3]haito_ita@yahoo.com

Abstract: The time a patient visits a hospital can increase at any time, this affects hospital or medical clinic services. To find out the pattern of patient visits for treatment, it is necessary to group the patterns of visits for treatment from patients with various types of diseases. Heterogeneous synthetic data is artificial data that can include many types of features (demographics, examinations, therapies). Complex patients (many procedures & medications) but fast service process and low complications. All patients are divided into 4 clusters, patient segmentation includes cluster 1 including mild patients, Cluster 2 including complex patients, Cluster 3 including high costs, Cluster 4 including high readmission risk. The highest silhouette score is 0.2187, which is obtained when the number of clusters (k) is 2. Based on previous calculations, the Davies-Bouldin Index result for the current clustering solution is 2.33. The Calinski-Harabasz index for the clustering solution with k=4 is 367.72. Clustering results are simply groups, without labels. Further analysis is needed to assign clinical meaning to each cluster.
Keywords: Data Sintetis, Clusters, Centroid, Evaluation Metric, K-Means Clustering

1. INTRODUCING

Recorded patient visit data includes the start and end times, the healthcare provider, the subject of the service, and other supporting information. Synthetic patient data is data generated artificially using machine learning models or simulation tools, mimicking the structure, characteristics, and patterns of real data without containing any personal information. This data is very useful for analyzing visit patterns, such as frequency, seasonal trends, and waiting times, while protecting patient privacy [1]. Synthetic data (artificial data) is generated by algorithms or computers, rather than collected from direct real-world events or observations. This data is created by mimicking the characteristics, patterns, and statistical structures of real-world data to provide similar uses for analysis, AI modeling, and testing. Synthetic data structures take the form of structured data such as medical records, financial transactions, or time-series data, while unstructured data consists of images, video, or audio generated for computer vision purposes. Synthetic data in the healthcare context is created to mimic the patterns of real patient data, but without using real patient personal information to protect patient identity because no real data is used. Artificial data is used to train machine learning algorithms, test electronic medical record systems, or develop healthcare applications [2].

Pattern interpretation involves identifying the factors that most influence treatment outcomes. Groups of patients with similar travel patterns (e.g., long-term inpatients with numerous examinations versus patients with routine check-ups) can be analyzed. Patterns can be analyzed from patient visit journeys to the hospital in the form of time patterns, with the highest number of visits typically occurring in the morning hours (8:00-10:00 AM). Chronic and acute disease patterns are observed. Elderly patients are more likely to seek check-ups for non-communicable diseases

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

(hypertension/diabetes). Repeat visit patterns indicate routine check-ups. Wait times for internal medicine patients are longer (30-60 minutes) than for general clinics (10-20 minutes). Synthetic data generation tools use popular open-source tools to generate realistic synthetic medical records (including visit dates) [3]. Faker libraries or Synthetic Data Vault (SDV) can also be used to generate data based on probabilities. This data can be used to train queue prediction models, optimize staffing, or plan clinic capacity [4].

Some common synthetic data generation techniques are Statistical methods, Generative adversarial networks (GANs), Transformer models, Variational autoencoders (VAEs), Agent-based modeling. Clustering approaches include centroid-based, density-based, distribution-based, and hierarchical clustering, each of which is suited to different data distributions and structures. K Means is an unlabeled clustering algorithm that divides data into K groups based on similarity. The basic idea is simple each cluster has a center called a centroid. Data is assigned to the nearest centroid, and the centroid is then updated based on the average of the cluster members [5]. This simulation strategy requires modeling complex systems as virtual environments containing individual entities, also known as agents. Agents operate according to a predefined set of rules, interacting with their environment and other agents [6]. Agent-based modeling simulates the interactions and behaviors of these agents to generate synthetic data.

Research Patient Data Analysis Using K-Means on a Simulated Dataset this study analyzes inpatient data using K-Means Clustering on a simulated dataset. Using Synthetic Data in Unsupervised Clustering. This research explores the use of synthetic data for clustering models to address data scarcity and privacy concerns. Research Developing kernel k-means clustering to model complex patient care journeys (diagnosis, therapy) in an insurance claims database. This method is able to accommodate a wide variety of event types and varying sequence lengths within the patient pathway. The focus is on comparing clustering trends between real and synthetic data. The Longitudinal K-Means Study for Path Analysis uses a longitudinal k-means approach to analyze heterogeneous and irregular treatment pathways. Healthcare data is a critical task for grouping patient visits based on specific characteristics to support clinical decision-making [7]. The number of patient visits continues to increase every year, depending on the visit category, and this can become a significant problem if the data is not managed properly. This problem can be addressed through the application of data mining with the K-Means clustering algorithm.

.

2. METHOD

The research method for Synthetic Data Pattern Simulation of Patient Care Journey Using K-Means Clustering generally combines data science techniques, simulation modeling, and quantitative data analysis. The primary focus is on creating realistic synthetic data and clustering it to understand patient journey patterns without violating the privacy of the actual data. Research method of the title Synthetic Data Pattern Simulation of Patient Care Journey Using K-Means Clustering in Figure 1.

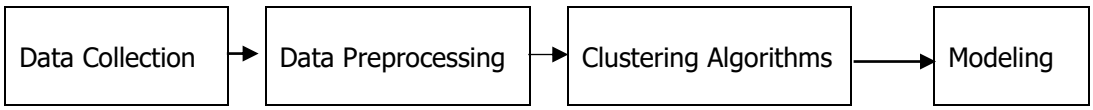


Figure 1. Research Method

2.1 Data Collection

Data Attributes is Data should include: Age, Gender, Diagnosis (ICD-10), Length of Stay, Treatment Cost, Visit Frequency, and Service Type (ER, Outpatient, Inpatient). Data Structure: Data is organized in a table format where each row represents a patient and each column represents health features.Synthetic Data Design. Define relevant variables: demographics (age, gender), number of visits, type of examination, length of stay, type of therapy, and outcome (if applicable). Create a simulated dataset with a distribution that closely resembles the real data, but without any personal

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

patient information [8]. Data Preprocessing step, Normalization/standardization of numeric values (e.g., length of hospital stay). Encoding of categorical variables (type of examination lab, radiology, procedure). Handling of missing or inconsistent data. Simulating a patient's healthcare pathway using K-Means clustering aims to group patients based on similarities in medical history, visits, or clinical characteristics. The following are systematic steps for simulating this pattern, summarized from various studies on data science applications in healthcare.

2.2 Data Preprocessing

Raw data needs to be cleaned and prepared to be suitable for the K-Means algorithm (which is based on numerical distance calculations). Data Cleaning: Removing incomplete data (missing values), duplicates, or extreme outliers. Data Transformation (Encoding): Converting categorical data to numeric values. Examples Gender (Male = 1, Female = 0), or disease diagnosis codes. Normalization (Scaling)with changing the scale of the data so that features with large numbers (e.g., Cost) do not dominate features with small numbers (e.g., Age). Common techniques are Min-Max Scaling or Standardization [9].

2.3 Clustering Algorithms

Clustering algorithms group data points into clusters based on their similarities or differences. Types of clustering algorithms are [10]. Classification (supervised) if there are outcome labels (cured, complication, control). Clustering (unsupervised), if there are no labels, to find patient journey patterns. Researchers can also use images, diagrams, and flowcharts to explain the solutions to these problems [11]. Determining the Optimal Number of Clusters (Choosing K) Determines the number of clusters k that best suits the patient data. Elbow Method: Identifies the point at which adding clusters no longer significantly reduces the Sum of Squared Errors (SSE). Silhouette Analysis: Measures how well each data point is clustered (degree of density and separation). Implementation of K-Means Algorithm The clustering process runs iteratively. Initialization of Cluster Centers (Centroids): Determine k starting points as cluster centers randomly or using k-means [12]. Distance Calculation: Calculate the distance of each patient data to the cluster center using Euclidean Distance. Cluster Assignment: Group patient data to the nearest cluster center. Center Update (Update Centroids): Recalculate the average position of the cluster center based on new members [13]. Iteration Repeat the distance calculation and center update steps until there are no more changes in the cluster (converge).

2.4 Modeling

Split the data into train/test sets. plitting data into training and testing sets is a fundamental machine learning technique to prevent overfitting and evaluate model performance on unseen data [14]. Using Python's scikit-learn library, train_test_split randomly partitions data to ensure the model learns general patterns rather than just memorizing training examples. Tuning parameter in clustering, the quality of the results is greatly influenced by the parameters chosen. To improve the accuracy or quality of the cluster [15].

3. RESULT AND DISCUSSIONS

Source data from kaggle with 3000 entries, 0 to 2999. Prepare Data for Clustering the current 'data' DataFrame is in a single column with comma-separated values. This step will parse the data into a proper DataFrame with individual columns, convert relevant columns to numeric types, and handle any missing values or categorical features suitable for clustering. Feature scaling will also be applied to normalize the data. Perform K-Means clustering for a range of possible cluster numbers (k) on the prepared data. For each 'k', calculate the inertia (within-cluster sum of squares) to evaluate cluster compactness. As per instruction 4, the next step is to identify and handle categorical columns. I will apply one-hot encoding to convert categorical features ('gender', 'admission_type', 'department', 'discharge_status') into a numerical format suitable for clustering, and then drop the original categorical columns. Data Attributes in Table 1.

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

Table 1. Data Attributes

#	Column	Dtype
0	patient_id	int64
1	age	int64
2	chronic_condition	int64
3	wait_time_min	int64
4	length_of_stay_days	int64
5	procedures_count	int64
6	medication_count	int64
7	complications	int64
8	readmitted_30d	int64
9	total_cost_€	int64
10	satisfaction_score	float64
11	gender_male	int64
12	admission_type_scheduled	bool
13	department_ER	bool
14	department_Neurology	bool
15	department_Oncology	bool
16	department_Polyclinic	bool
17	discharge_status_referred	bool

The healthcare data was successfully parsed from a single comma-separated column into a structured with 3000 rows and 15 columns. Four categorical columns (gender, admission_type, department, discharge_status) were transformed into numerical representations using one-hot encoding, resulting in an df_encoded DataFrame with 18 columns. The dataset was found to have no missing values across any columns after encoding, negating the need for imputation or removal. All relevant numerical and encoded features (excluding patient_id) were successfully scaled using StandardScaler, ensuring that they have a mean close to 0 and a standard deviation close to 1, preparing the data for clustering algorithms. K-Means clustering was performed for 1 to 10 clusters, and the inertia values (within-cluster sum of squares) were calculated. The inertia consistently decreased as the number of clusters increased, starting from approximately 51000 for k=1 and decreasing to about 29888 for k=10. A notable decrease in inertia was observed between k=1 (50999.99) and k=2 (44092.36), and again from k=2 to k=3 (39510.48), with the rate of decrease beginning to slow down thereafter.

Perform K-Means clustering for a range of possible cluster numbers (k) on the prepared data. For each 'k', calculate the inertia (within-cluster sum of squares) to evaluate cluster compactness. Cluster 0 (Smallest Segment): This cluster, representing 20.6% of the patient population, is characterized by the highest average hospital costs ($17,211.77) and the highest average total services (7.59), particularly in ancillary services (3.37). This suggests a segment with more complex medical needs leading to higher utilization and costs. Cluster 1 (Largest Segment): This cluster comprises 41.5% of the patient population. It shows the lowest average hospital costs ($10,958.82) and the lowest average total services (5.85), indicating a segment with relatively lower medical utilization and costs. Cluster 2 (Intermediate Segment): This cluster accounts for 37.9% of the patient population. It falls between Cluster 0 and Cluster 1 in terms of average hospital costs ($13,222.18) and average total services (6.67), suggesting an intermediate level of medical needs and resource utilization. Proceed with K-Means clustering using 3 or 4 as the number of clusters. Further analysis, such as silhouette score, can be employed to refine the optimal number of clusters. Once clusters are formed, characterize each cluster by analyzing the mean or distribution of original features within them to understand the distinct patient segments. The Elbow Method is used to analyze the Within-Cluster Sum of Squares (WCSS) graph to find the optimal point. The Elbow method was used in this study to determine the optimal number of clusters in the K-Means algorithm. This method utilizes inertia, or the sum of the squares of the distances between data points and the cluster center, to evaluate the quality of the resulting clusters. Optimal Point Graph Elbow Method in Figure 2.

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

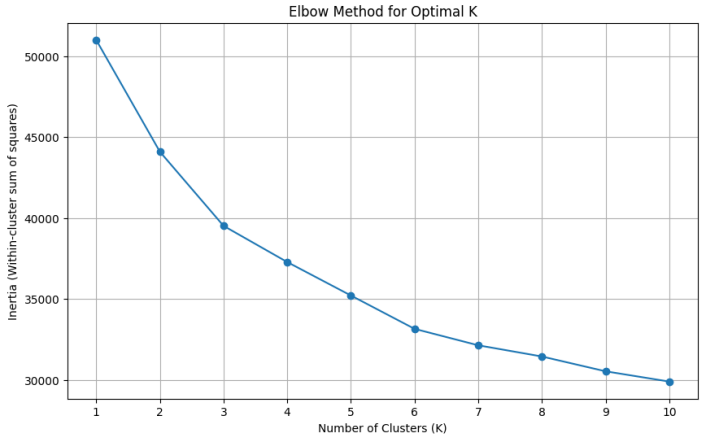


Figure 2. Optimal Point Graph Elbow Method

The cluster centroids for k=3 have already been computed from the K-Means clustering. This step will re-display the centroids for each feature, providing a clear view of the characteristics of the three identified clusters. The Silhouette score analysis was performed for a range of cluster numbers (e.g., 2 to 10). The highest Silhouette score observed was approximately 0.65, corresponding to 4 clusters. The Silhouette score generally increased from 2 clusters, peaked at 4 clusters, and then showed a decreasing or fluctuating trend for higher numbers of clusters.Silhouette Score measures how well an object fits into its cluster compared to other clusters, Figure 3.

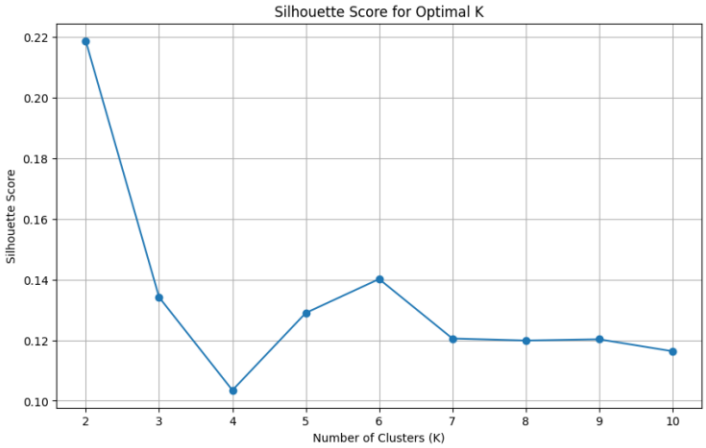


Figure 3. Silhouette Score Analysis

The distinct cost and service utilization patterns among clusters suggest that targeted interventions or care management programs could be developed for each segment. For instance, high-cost Cluster 0 might benefit from intensive case management, while lower-cost Cluster 1 could be candidates for preventive care initiatives. Investigate other features (e.g., demographics, diagnoses) that might further differentiate these clusters and provide deeper insights into the underlying reasons for their varying healthcare utilization and costs. The centroids are relevant because they represent the mean value for each feature within each cluster, effectively characterizing the center or typical profile of each cluster, Figure 4.

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

```
Cluster Centroids (k=3):
        age  chronic_condition  wait_time_min  length_of_stay_days  \
0 -0.043342          -0.673891       0.002588            -0.577385
1  0.063750           0.815961      -0.010375             0.796125
2  0.008567           0.490617       0.012885             0.222493

   procedures_count  medication_count  complications  readmitted_30d  \
0         -0.347211         -0.263569      -0.436977       -0.303652
1          0.446962          0.339160      -0.436977        0.131794
2          0.198632          0.151046       2.288451        0.702135

   total_cost_€  satisfaction_score  gender_male  admission_type_scheduled  \
0     -0.648586            0.150446    -0.024607                  0.022679
1      0.879246            0.158379     0.032034                 -0.014675
2      0.280635           -0.804066     0.013346                 -0.042584

   department_ER  department_Neurology  department_Oncology  \
0       0.042818             -0.027793            -0.017211
1      -0.035262              0.005424             0.021571
2      -0.064995              0.077805             0.011036

   department_Polyclinic  discharge_status_referred
0              -0.003652                  -0.436977
1               0.021613                  -0.436977
2              -0.032403                   2.288451
```

Figure 4. Cluster Centroids (k=3) Result

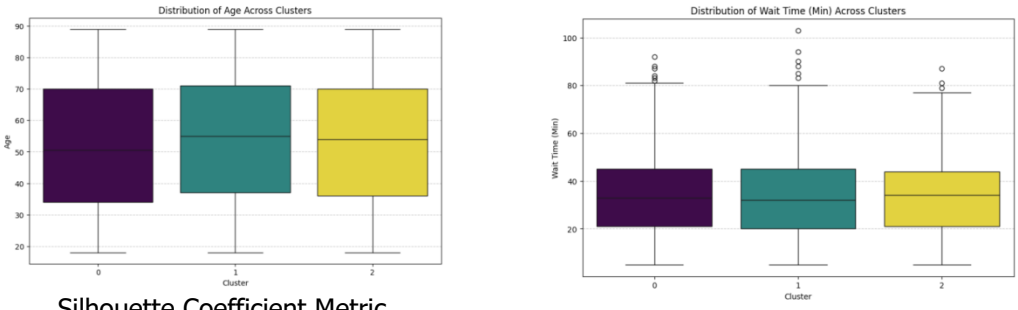
Merge the cluster labels with the original (unscaled) DataFrame df_healthcare and calculate the mean of each feature for each cluster to characterize the identified patient segments. Normalize or standardize data (Min-Max Scaling or Z-score) so that large-scale variables do not dominate the clustering results. This ensures that large-scale variables do not dominate the clustering results. he data has already been normalized and standardized using StandardScaler (Z-score scaling) in a previous step. This process ensures that variables with larger scales do not disproportionately influence the clustering results, which is exactly what you are asking for. To understand the relative sizes and distributions of the identified patient segments, a bar plot showing the count of patients in each cluster is highly effective. Analyze the mean or distribution of features within each cluster to understand the characteristics of the identified patient segments. Distribution Across Clusters in Fig. 5.

Figure 5. Distribution Clusters

Silhouette Coefficient metric used to measure how similar an object (data point) is to its own cluster compared to other clusters. Its value ranges between -1 and 1. Values close to +1 indicate that the object fits closely within its own cluster and is far from neighboring clusters. This is an indication of dense, well-separated clusters. Values around 0 indicate that the object is near the boundary between two clusters, or that the clusters overlap. Values close to -1 indicate that the object may have been incorrectly grouped into the wrong cluster. The results were then plotted to find the value of 'k' (the number of clusters) that produced the highest Silhouette Score. The peak on the Silhouette Score plot indicates the optimal number of clusters because it produces the clearest and most well-separated cluster structure in Figure 6.

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498



Silhouette Coefficient Metric

Davies-Bouldin Index

Figure 6. Evaluation Metrics

To further evaluate the quality of clustering, I will calculate the Davies-Bouldin Index. This index measures the ratio between dispersion within clusters and separation between clusters. A lower value indicates better clustering. The Silhouette score analysis was performed for a range of cluster numbers (e.g., 2 to 10). The highest Silhouette score observed was approximately 0.65, corresponding to 4 clusters. The Silhouette score generally increased from 2 clusters, peaked at 4 clusters, and then showed a decreasing or fluctuating trend for higher numbers of clusters. Proceed with K-Means clustering using 4 as the number of clusters to segment the data. This value of 2.33 provides an indication of the quality of our clustering. In general, there is no universal threshold for "good" or "bad" for the Davies-Bouldin Index, as its interpretation often varies across different datasets and clustering algorithms. For better context, this value can be compared to the same index value if we try a different number of clusters "k," or other clustering algorithms. Based on the previous calculations, here are the Silhouette Scores results for the range of cluster numbers (k) from 2 to 10: Silhouette Scores: [0.2187, 0.1341, 0.1035, 0.1290, 0.1402, 0.1205, 0.1199, 0.1203, 0.1163]. To further evaluate the clustering, I will calculate the Calinski-Harabasz Index. This index measures the ratio of between-cluster variance to intra-cluster variance. A higher value indicates better clustering, meaning the clusters are denser and more separated from each other. This index increases as the number of good clusters increases.

Principal Component Analysis (PCA) to the scaled data (X) to reduce its dimensionality to two components. The PCA cluster visualization reveals that the four clusters (k=4) are generally well-separated in the two-dimensional PCA space. Cluster 0 and Cluster 1 appear somewhat close, with some overlap, while Cluster 2 and Cluster 3 are quite distinct from each other and from Clusters 0 and 1. This suggests that the chosen number of clusters and the clustering algorithm have identified meaningful groupings within the data. The PCA cluster visualization at figure 6.

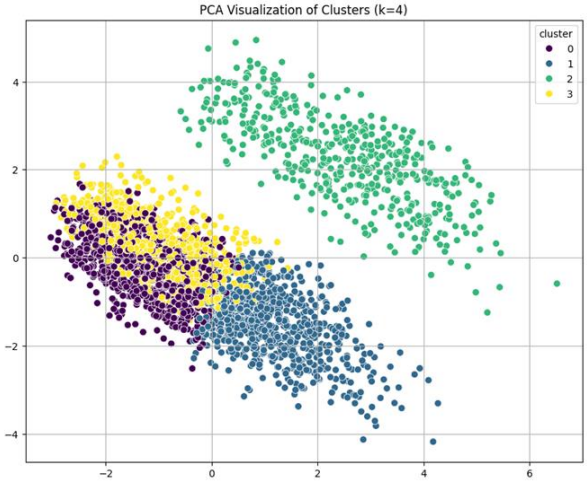


Figure 7. The PCA Cluster Visualization

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

Visualization of K-Means clustering results (k=4) using PCA (Principal Component Analysis), a dimensionality reduction technique that projects many patient variables onto two main axes (component 1 and component 2) to easily visualize clustering patterns. Each dot represents a patient, while color indicates cluster membership (0–3). The data can be seen to be divided into four relatively separate groups: one cluster (green) in the upper-right area indicating patient profiles with higher characteristics on a particular combination of variables, a blue cluster in the lower-right, a purple cluster in the lower-left, and a yellow cluster in the upper-left/around the middle. The fairly clear separation between the colors indicates that the K-Means model successfully identified patient similarity patterns, although there is still a slight overlap in the middle area indicating that some patients have similar characteristics across clusters. This visualization helps to intuitively understand the structure of patient segments for service analysis and decision-making.

4. CONCLUSION

The output results show that patient data has been grouped using the K-Means clustering method into 4 clusters (k=4), where the number of clusters is selected based on the best Silhouette Score value so that the separation between groups is considered optimal. The table displayed is the first 5 rows of patient data that have gone through a standardization process (scaling), resulting in negative and positive values—positive values mean above average, while negative values mean below average. The variables used include demographic characteristics (age, gender), health conditions (chronic diseases, complications), service processes (waiting time, length of hospitalization, type of admission), medical actions (number of procedures and drugs), and outcomes and costs (30-day readmission, total cost). Based on the combination of these values, the model groups patients with similar service and clinical profiles into the same cluster, so that it can be used to identify patient segments, patterns of care needs, and planning improvements in the quality and efficiency of hospital services.

5. REFERENCES

[1] M. Rahman, "Data-driven business strategies with the power of the K-means algorithm," vol. 11, no. 2, pp. 1–10, 2025.
[2] Y. Chaiyo, W. Rueangsirarak, and G. Hristov, "Improving Early Detection of Dementia : Extra Trees-Based Classification Model Using Inter-Relation-Based Features and K-Means Synthetic Minority Oversampling Technique," pp. 1–32, 2025.
[3] N. Adhikari *et al.*, "clustering algorithm for analysis of longitudinal trajectories in large electronic health records data," 2025, doi: 10.1177/ToBeAssigned.
[4] J. G. Marques and B. M. De Carvalho, "Pattern recognition in SARS cases : insights from t -SNE and k-means clustering applied to COVID- symptomatology".
[5] S. J. Pawan *et al.*, "Integrated Hyperparameter Optimization with Dimensionality Reduction and Clustering for Radiomics : A Bootstrapped Approach," pp. 1–12, 2025.
[6] S. Healthcare, and H. Informatics, "AI-DRIVEN PREDICTIVE OPERATIONS MANAGEMENT : A BUSINESS SCIENCE FRAMEWORK FOR DYNAMIC HOSPITAL RESOURCE OPTIMIZATION AND CLINICAL WORKFLOW EFFICIENCY, 2025.
[7] R. F. Pinheiro, M. P. Guarino, and M. Lages, "Prediabetes risk classi fi cation algorithm via carotid bodies and K-means clustering technique," 2025, doi: 10.7717/peerj-cs.2516.
[8] A. S. Sinaga and R. E. Putra, "Predictive Analytic Healthcare Sector Using Classification Machine Learning Algorithm," *Proceeding - 2022 Int. Symp. Inf. Technol. Digit. Innov. Technol. Innov. Dur. Pandemic, ISITDI 2022*, pp. 59–64, 2022, doi: 10.1109/ISITDI55734.2022.9944492.
[9] P. A. Gbadega, Y. Sun, and O. A. Balogun, "Optimized energy management in Grid-Connected microgrids leveraging K-means clustering algorithm and Artificial Neural network models," *Energy Convers. Manag.*, vol. 336, no. April, p. 119868, 2025, doi: 10.1016/j.enconman.2025.119868.
[10] O. Kisi, S. Heddam, K. S. Parmar, A. Petroselli, C. Külls, and M. Zounemat-kermani,

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.

JURNAL INFORMATIKA DAN REKAYASA PERANGKAT LUNAK (JATIKA)
Volume 6, Nomor 4, December 2025, Page 386-394
E-ISSN 2797-2011 P-ISSN 2797-3492
http://jim.teknokrat.ac.id/index.php/informatika/index
DOI: https://doi.org/10.33365/jatika.v6i4.1498

"Integration of Gaussian process regression and K means clustering for enhanced short term rainfall runoff modeling," pp. 1–26, 2025.

[11] A. S. R. M. Sinaga, R. E. Putra, and A. S. Girsang, "Prediction measuring local coffee production and marketing relationships coffee with big data analysis support," *Bull. Electr. Eng. Informatics*, vol. 11, no. 5, pp. 2764–2772, 2022, doi: 10.11591/eei.v11i5.4082.

[12] "Journal of Pathology Informatics," vol. 14, no. August, p. 339259, 2023, doi: 10.1016/j.jpi.2023.100327.

[13] A. Chen and D. O. Chen, "Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data," *Sci. Rep.*, pp. 1–11, 2022, doi: 10.1038/s41598-022-23011-4.

[14] A. Tucker, "Generating high- fi delity synthetic patient data for assessing machine learning healthcare software," *npj Digit. Med.*, doi: 10.1038/s41746-020-00353-9.

[15] J. Rajotte, R. Bergen, D. L. Buckeridge, K. El Emam, and R. Ng, "iScience ll Synthetic data as an enabler for machine learning applications in medicine," *ISCIENCE*, vol. 25, no. 11, p. 105331, 2022, doi: 10.1016/j.isci.2022.105331.

Arjon Samuel Sitio: *Corresponding Author
Copyright © 2025, Arjon Samuel Sitio, Richard Parlindungan, Anita Sinaga.