YAYASAN AKRAB PEKANBARU
Akrab Juara : Jurnal Ilmu-ilmu Sosial
Volume 11 Nomor 2 Edisi Mei 2026 (485-490)
https://doi.org/10.58487/akrabjuara.v10i1.0000

40

INPATIENT CLUSTER ANALYSIS FOR MEDICAL RESOURCE OPTIMIZATION USING K-MEANS CLUSTERING

Adika May Sari[1], Amas Sari Marthanti[2], Rosmita[3], Dessy Suryani[4], Ahmad Rafik[5]
Universitas Bina Sarana Informatika[12345]
adika.dik@bsi.ac.id, amas.mtm@bsi.ac.id, rosmita.rmt@bsi.ac.id, dessy.dsn@bsi.ac.id, ahmad.aaf@bsi.ac.id
(Naskah diterima: 1 Maret 2026, disetujui: 31 Maret 2026)

Abstract
The objective of this study is to analyze inpatient data in order to segment patients based on specific characteristics such as age, type of illness, medical procedures, and length of stay. The K-Means Clustering method is applied to identify patterns or patient segments that can be utilized in decision-making related to bed management and medical staff allocation more efficiently. The analysis was conducted using the Python programming language for data processing and result visualization. The findings indicate the existence of several groups of patients with distinct characteristics, which can serve as a strategic reference for improving service quality and the operational effectiveness of the hospital.
Keywords—clustering, K-Means, inpatient, length of stay, healthcare resource optimization.

Abstrak
Tujuan dari penelitian ini adalah untuk menganalisis data pasien rawat inap guna mengelompokan pasien berdasarkan karakteristik spesifik seperti usia, jenis penyakit, tindakan medis, dan lama inap. Metode K-Means Clustering digunakan untuk mengidentifikasi pola atau segmen pasien yang dapat yang dapat di manfaatkan dalam pengambilan keputusan terkait pengelolaan tempat tidur dan alokasi tenaga medis secara lebih efisien. Analisis dilakukan dengan menggunakan bahasa pemograman Python untuk pemrosesan data visualisasi hasil. Hasil penelitian menunjukan adanya beberapa kelompok pasien dengan karakteristik berbeda yang dapat menjadikan referensi strategis dalam meningkatkan kualitas layanan dan efektivitas operasional rumah sakit.
Kata kunci—clustering, K-Means, pasien rawat inap, lama inap, optimalisasi sumber daya medis.

I. INTRODUCTION

Hospitals play a vital role in healthcare systems focused on improving the community's quality of life. However, the rising number of patients and the increasing complexity of medical cases necessitate more efficient resource management and medical staff allocation—particularly regarding bed management. Suboptimal management can lead to patient overcrowding, service delays, and excessive workloads for medical personnel (Purba et al., 2023). Therefore, the implementation of appropriate technologies and data analysis methods is crucial for supporting evidence-based decision-making that enhances hospital operational efficiency.

One effective approach is the application of data mining to analyze historical patient data—such as bed occupancy rates, readmission trends, and workloads associated with specific diseases and medical procedures. Studies indicate that data mining methods, such as K-Means Clustering, can be used to group patients based on their medical characteristics, thereby facilitating the design of more efficient hospital resource management strategies (Faizal Rizqi et al., 2024). Furthermore, classification techniques

can be employed to predict the likelihood of patient readmission based on factors such as age, type of illness, and length of stay, helping hospitals better prepare their care capacity (Surya & Maryana, 2020). Data mining is the process of extracting and discovering relevant information and related knowledge from vast datasets using statistical, mathematical, artificial intelligence, and machine learning techniques. Techniques employed in data mining include association, clustering, classification, and prediction.It is expected that this study, analyzed using the Python programming language, will provide valuable information to hospital management for enhancing operational effectiveness and optimizing services.

II. METHOD

1. **K-Means Clustering**

   Objects are grouped in clustering based on data information that clarifies the relationships between them (Tambuwun et al., 2020). The K-Means clustering process is illustrated in Figure 3 (Azhari et al., 2023). K-Means clustering is an unsupervised learning algorithm used to group data based on the similarity of characteristics among data points.

2. **The steps for clustering using the K-Means method are as follows:**

   a) Determine the value of *k* (the number of clusters to be formed).

   b) Determine the initial center point for each cluster.

   c) Calculate the distance from each input data point to each centroid using the Euclidean distance formula to identify the nearest centroid for each data point. The Euclidean distance formula is: $D(x,y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_1)^2 + (X_3 - Y_1)^2}$, where D = Distance, x = Data point, and y = Centroid. d) Classify the data based on proximity to the centroids. e) Recalculate the cluster centers based on the current cluster members; the cluster center is the average value of all data objects within a specific cluster. f) Recalculate the position of each object relative to the new cluster centers; if the cluster centers do not change, the clustering process is complete. Alternatively, repeat the process starting from step (c) until the cluster centers no longer change.

III. RESULT

1. Data and Data Sources

This study utilizes a simulated hospital patient dataset obtained from the Kaggle platform (https://www.kaggle.com/datasets/blueblushed/hospital-dataset-for-practice), a prominent open data source for analysis and machine learning purposes. The dataset contains key information such as patient age, costs, medical conditions, medical procedures, length of stay, satisfaction levels, and readmission status.

The dataset is used to identify hospitalization patterns and segment patients based on their medical characteristics to support strategies for optimizing hospital medical resources. Data analysis is conducted using the Python programming language, supported by libraries such as pandas (for data manipulation), scikit-learn (for implementing the K-Means Clustering algorithm), and matplotlib and seaborn (for exploratory visualization).

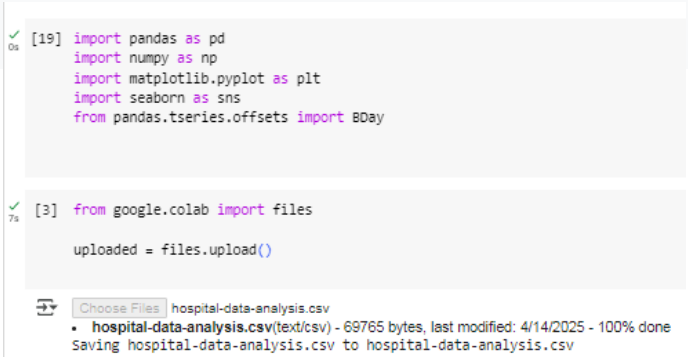
2. Research Stages

The methodology employed in this study comprises several key stages, namely:

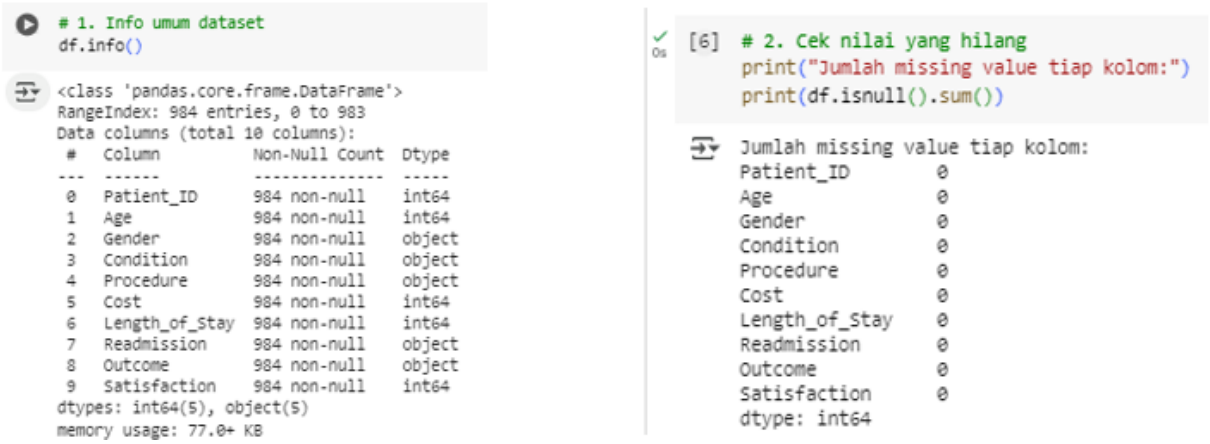
3. Data Collection and Cleaning

The simulated hospitalization dataset is downloaded from Kaggle to initiate the data collection process. For subsequent analysis, the dataset is imported into a Python environment using packages such as pandas. To ensure data quality and consistency, the data undergoes a cleaning procedure prior to analysis. This cleaning step addresses issues such as duplicate records, formatting errors, and missing values.

This cleaning process is essential because poor data quality can lead to biased and misleading analysis results. As noted by Ortiz et al. (2024), data preprocessing—including data cleaning—is crucial for ensuring data reliability in analyses involving artificial intelligence and machine learning, particularly regarding health data collected from wearable devices. Handling missing values, removing outliers, and standardizing data are critical steps in enhancing data quality before the data is used in predictive models (Ortiz et al., 2024).



Picture 2.1 import data to phyton



Picture 2.2 Data Cleansing

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is employed in this study. EDA provides the initial framework for analysis and is conducted at the beginning of the analytical process. This stage involves examining and understanding the data to uncover patterns and trends. According to Joseph Santoso (2023), EDA helps data scientists familiarize themselves with the data before undertaking more complex analyses. The primary objectives of EDA are to detect errors and identify patterns within the data. EDA also assists in identifying and handling missing values, duplicates, outliers, and data entry errors (Joseph M. Tandiallo, 2024). The EDA stage is carried out to understand the structure and characteristics of the data prior to applying more complex analytical techniques, such as clustering. This process includes:

- Analyzing data distributions—such as patient age and length of stay—to understand the spread of values and detect outliers.
- Identifying correlations between variables, for instance, the relationship between the type of disease and the medical procedures administered.
- Visualizing data using charts such as histograms, boxplots, and heatmaps to identify patterns, trends, and anomalies within the dataset.

The EDA process helps reveal hidden information within the data and provides crucial preliminary insights for decision-making.

The data used in this study comprises records for 1,000 hospitalized patients, covering attributes such as name, gender, medical condition, procedure, cost, length of stay, readmission, outcome, and customer satisfaction. The attributes selected for clustering are patient age, cost, length of stay, and satisfaction.

Patient segmentation using K-Means Clustering enables the hospital to optimize resource management, improve service quality, and gain a deeper understanding of patient needs.

Importing Libraries
The necessary libraries are imported to facilitate data preparation, modeling and evaluation (K-Means Clustering), and result visualization.
Import the required libraries.

- Impor library

```
[26] import pandas as pd
     import numpy as np
     from sklearn.cluster import KMeans
     from sklearn.preprocessing import StandardScaler
     from sklearn.metrics import silhouette_score
     import matplotlib.pyplot as plt
```

- Preprocessing data, K- Means *(Age, Cost, Length_of_Stay, Satisfaction).*

```
[78] numeric_features = ['Age', 'Cost', 'Length_of_Stay', 'Satisfaction']
     data = df[numeric_features]
```

- Normalisation data /

```
[79] scaler = StandardScaler()
     scaled_data = scaler.fit_transform(data)
```

Picture 2.3 import libraries

2. Determining the Number of Clusters
To find the value of *k* in the k-means clustering process, the silhouette score and the within-cluster sum of squares (WSS) must be calculated and visualized. Determining the optimal number of clusters (Elbow Method).

```
wcss = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_data)
    wcss.append(kmeans.inertia_)

# Plot Elbow Method
plt.plot(range(1, 11), wcss, marker='o')
plt.xlabel('Jumlah Kluster (k)')
plt.ylabel('WCSS')
plt.title('Elbow Method')
plt.show()
```

```
# Plot Silhouette Score
plt.figure(figsize=(10, 6))
plt.plot(range(2, 11), silhouette_scores, marker='o', linestyle='--', color='b')
plt.xlabel('Jumlah Kluster (k)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score untuk Menentukan Jumlah Kluster Optimal')
plt.xticks(range(2, 11))
plt.grid(True)
plt.show()
```

Picture 2.4 Determining clusters

3. Distance from Object to Centroid
The researcher calculates the distance of each data point to each centroid. The figure below illustrates the calculation of the distance from each object to the centroids and the assignment of the object to the nearest centroid.

	Patient_ID	Cluster	Jarak_Terdekat_Ke_Centroid
0	1	2	2.37
1	2	1	2.83
2	3	2	1.90
3	4	1	2.44
4	5	1	2.83
5	6	0	1.95
6	7	0	1.93
7	8	2	1.74
8	9	1	2.16
9	10	2	1.97

```
Cluster 0: Rata-rata jarak = 1.24
Cluster 1: Rata-rata jarak = 1.95
Cluster 2: Rata-rata jarak = 1.50
```

```
Outlier dengan jarak terbesar:
Patient_ID                    995
Condition                  Cancer
Jarak_Terdekat_Ke_Centroid    2.894924
Name: 978, dtype: object
```

Picture 2.5 centroid

4. Based on the calculated distances of objects to the centroid, Cluster 0 has the shortest average distance (1.24),

Cluster 1 has an average distance of 1.95, and Cluster 2 has an average distance of 1.50. The patient with ID 995 (cancer condition) is the closest to the centroid.
4.4 K-means Clutering Results Presented below is the sequence of the analysis and the K-means clustering results for partitioning the data into three groups.

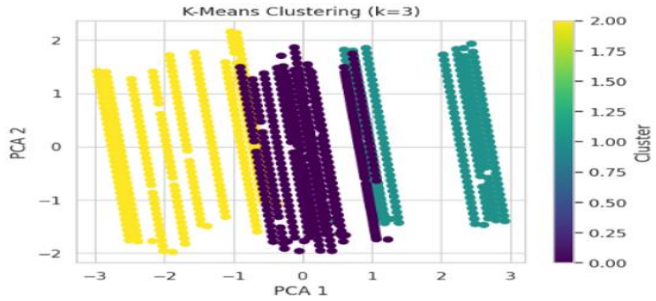
- Data Preprocessing

Select relevant data, specifically numerical and categorical variables.

```
df.groupby('Cluster').agg({
    'Age': 'mean',
    'Cost': 'mean',
    'Length_of_Stay': 'mean',
    'Satisfaction': 'mean',
    'Readmission': 'mean', # Kolom Readmission ditambahkan
}).round(2).rename(columns={'Readmission': 'Readmission (%)'})
)
```

Cluster	Age	Cost	Length_of_Stay	Satisfaction	Readmission (%)
0	57.30	4389.26	37.15	3.64	0.15
1	32.63	4002.50	34.47	4.55	0.14
2	65.48	17882.15	41.01	2.77	0.56

Picture 2.6 K-Means Result



Picture 2.7 ClusteringResult

The K-Means clustering process (with n_clusters=3) divided patients into three groups based on similarities in the features Age, Cost, Length_of_Stay, Satisfaction, and Readmission. The interpretations are as follows:

1. **Cluster 0:**
   - Age: Average age of 57.30 years; patients ranging from adulthood to approaching old age.
   - Cost: 4,389.26 → Moderate treatment cost.
   - Length_of_Stay: 37.15 → Relatively long hospital stay.
   - Satisfaction: 3.64/5 → Moderate satisfaction.
   - Readmission (%): 0.15 (15%) → Low readmission rate.

This cluster represents adult patients with long hospital stays and moderate treatment costs. Although satisfaction levels are not particularly high, readmission remains low, indicating that the treatment process is effective in preventing patient readmission.

2. **Cluster 1:**
   - Age: 32.63 years → Young patients.
   - Cost: 4,002.50 → Lowest treatment cost.
   - Length_of_Stay: 34.47 days → Moderate hospital stay.
   - Satisfaction: 4.55 → Highest satisfaction among all clusters.
   - Readmission (%): 0.14 (14%) → Lowest readmission rate.

This cluster consists of young patients with economical treatment costs and the highest service satisfaction. This reflects efficient and effective service; they recover quickly and rarely require readmission.

3. **Cluster 2:**
   - Age: 65.48 years → Elderly patients.
   - Cost: 17,882.15 → Highest treatment cost.
   - Length_of_Stay: 41.01 days → Longest hospital stay.
   - Satisfaction: 2.77 → Lowest satisfaction. • Readmission (%): 0.56 (56%) → Highest readmission rate.

This cluster comprises elderly patients with complex medical needs. Prolonged hospital stays and high costs indicate severe cases. However, low satisfaction and high readmission rates highlight the need for service improvements and stricter monitoring for this group.

IV. CONCLUTION

In conclusion, this study successfully applied the K-Means Clustering method to segment inpatients based on several key parameters: age, treatment cost, length of stay, satisfaction level, and readmission. The clustering results divided patients into three distinct groups with unique characteristics:
Cluster 1 consists of young patients with low treatment costs, short hospital stays, and high satisfaction levels. This group tends to have a low incidence of readmission.
Cluster 0 represents adult patients with moderate conditions. While readmission rates are low, there is room for improvement regarding length of stay and satisfaction levels.
Cluster 2 comprises elderly patients with high medical needs, prolonged hospital stays, and significant readmission rates. Satisfaction with services is also lowest within this cluster.
This analysis demonstrates that patient segmentation using K-Means can serve as an effective approach to support data-driven decision-making in hospitals—particularly regarding bed management, medical staff allocation, and the planning of more targeted care interventions.

BIBLIOGRAPHY

Faizal Rizqi, M., Martanto, M., & Hayati, U. (2024). Clustering Kunjungan Pasien Menggunakan Algoritma K-Means Pada Rumah Sakit Di Wilayah Bekasi. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 230–237. https://doi.org/10.36040/jati.v8i1.8327

Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., Ziegenbein, C., Singhal, A., Choi, S. W., & Tewari, M. (2024). Data Preprocessing Techniques for Artificial Intelligence (AI)/Machine Learning (ML)-Readiness: Systematic Review of Wearable Sensor Data in Cancer Care (Preprint). *JMIR MHealth and UHealth*, *September*. https://doi.org/10.2196/59587

Prakoso, B. H., Rachmawati, E., Mudiono, D. R. P., Vestine, V., & Suyoso, G. E. J. (2023). Klasterisasi Puskesmas dengan K-Means Berdasarkan Data Kualitas Kesehatan Keluarga dan Gizi Masyarakat. *Jurnal Buana Informatika*, *14*(01), 60–68. https://doi.org/10.24002/jbi.v14i01.7105

Surya, & Maryana, N. F. (2020). Faktor Faktor Yang Berhubungan Dengan Pengetahuan Keluarga Pasien Mengenai Cuci Tangan. *Jurnal Penelitian Perawat Profesional*, *2*(5474), 1333–1336.

Purba, Windania, Gamaliel Armando Sembiring, Andreas Saputra, Theresia Turnip, Ben Jua, Ivand Manihuruk, Fakultas Sains, and Dan Teknologi. 2023. "Penerapan Data Mining Untuk Pengelolaan Data Rekam Medis Menggunakan Metode K-Means Clustering Pada Rumah Sakit Royal Prima Medan." Jurnal TEKINKOM 6(1):158–68. doi: 10.37600/tekinkom.v6i1.857.

Joseph Santoso. (2023). Ilmu data (Data Science) (Muhammad Sholikan (ed.); Pertama). Yayasan Prima Agus Teknik.

Joseph M. Tandiallo. (2024). Exploratory Data Analysis (EDA) Using Python. In kaggle. https://medium.com/@teppan_noodle/exploratory-data-analysis-eda-using-python-f85938cb1810

Tambuwun, C. H., Langi, Y. A. R., & Rindengan, A. J. (2020). Estimasi Bobot Parameter M Pada Fuzzy C-Means Menggunakan Analisis Robust Dengan Simulasi Data Spasial. D'CARTESIAN, 9(1), 50. https://doi.org/10.35799/dc.9.1.2020.27600

Azhari, R., Hartama, D., Lubis, M. R., Nasution, D. F., & Windarto, A. P. (2023). Analisis Penerapan Data Mining Terhadap Kasus Positif Covid-19 Menggunakan Metode K-Means Clustering. Journal of Informatics, Electrical and Electronics Engineering, 3(2), 221–235. https://doi.org/10.47065/jieee.v3i2.1760