

Penerapan Data Mining Untuk Analisis Pola Penyakit Pasien Rawat Inap Menggunakan Metode K-Means

Application of Data Mining for Analyzing Disease Patterns Among Inpatients Using the K-Means Clustering Method

Ella Kristiantini Susan¹, Arief Hermawan², Donny Avianto³

^{1,2} Magister Teknologi Informasi, Fakultas Pascasarjana, Universitas Teknologi Yogyakarta

³ Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Teknologi Yogyakarta

¹sekarlangit02@gmail.com*, ²ariefdb@uty.ac.id, ³donny@uty.ac.id

Abstrak

Pengelolaan data rekam medis pasien rawat inap menghasilkan data dalam jumlah besar yang belum dimanfaatkan secara optimal untuk mendukung pengambilan keputusan manajerial dan perencanaan layanan kesehatan berbasis data. Rumah sakit menghadapi kesulitan dalam mengidentifikasi pola karakteristik pasien yang berkaitan dengan biaya pelayanan, lama perawatan, penggunaan ICU, dan perbedaan tarif BPJS dengan tarif rumah sakit. Penelitian ini menerapkan teknik data mining menggunakan algoritma K-Means Clustering pada 9.889 data pasien rawat inap. Tahapan penelitian meliputi pembersihan data, normalisasi Z-Score, penentuan jumlah kluster optimal menggunakan Elbow Method dan Silhouette Score, serta analisis karakteristik setiap kluster berdasarkan variabel tarif BPJS, tarif rumah sakit, lama dirawat, lama ICU, dan usia pasien. Hasil penelitian menunjukkan bahwa jumlah kluster optimal adalah $k=3$ dengan Silhouette Score 0,3441. Kluster yang terbentuk terdiri atas Cluster 0 (Pasien Kritis) sebanyak 376 pasien (3,8%) dengan biaya dan penggunaan ICU tertinggi, Cluster 1 (Pasien Menengah) sebanyak 6.625 pasien (67,0%) yang didominasi pasien lansia dengan tingkat perawatan moderat, dan Cluster 2 (Pasien Ringan) sebanyak 2.888 pasien (29,2%) yang didominasi pasien usia muda dengan biaya serta lama perawatan terendah. Selain itu, ditemukan rata-rata selisih tarif rumah sakit dan BPJS sebesar Rp5.602.334 per pasien. Temuan ini menunjukkan bahwa pendekatan clustering mampu mengidentifikasi kelompok pasien dengan karakteristik pemakaian sumber daya dan finansial yang berbeda secara signifikan sehingga dapat digunakan sebagai dasar optimalisasi alokasi sumber daya rumah sakit, perencanaan kapasitas ICU, pengembangan layanan geriatri, serta evaluasi kebijakan pembiayaan dan klaim BPJS pada kelompok pasien dengan beban biaya tertinggi.

Kata kunci: Data Mining, K-Means Clustering, Rekam Medis, Rawat Inap, BPJS.

Abstract

The management of inpatient medical records generates large volumes of data that are often underutilized for managerial decision-making and data-driven healthcare planning. Hospitals frequently encounter difficulties in identifying patient characteristic patterns associated with healthcare costs, length of stay, intensive care unit (ICU) utilization, and discrepancies between Indonesia's National Health Insurance (BPJS) reimbursement rates and hospital service charges. This study applied a data mining approach using the K-Means Clustering

algorithm to analyze 9,889 inpatient records. The research process involved data cleaning, Z-score normalization, determination of the optimal number of clusters using the Elbow Method and Silhouette Score, and cluster characterization based on BPJS reimbursement rates, hospital charges, length of hospitalization, ICU length of stay, and patient age. The results identified three optimal clusters ($k = 3$) with a Silhouette Score of 0.3441. Cluster 0 (Critical Patients) consisted of 376 patients (3.8%) and was characterized by the highest healthcare costs and ICU utilization. Cluster 1 (Intermediate Patients) included 6,625 patients (67.0%), predominantly elderly individuals requiring moderate levels of care. Cluster 2 (Low-Acuity Patients) comprised 2,888 patients (29.2%), mainly younger patients with the lowest costs and shortest hospital stays. The analysis also revealed an average discrepancy of IDR 5,602,334 per patient between hospital charges and BPJS reimbursement rates. These findings demonstrate that clustering can effectively identify patient groups with distinct resource-use and financial characteristics, providing valuable insights for resource allocation, ICU capacity planning, geriatric service development, and evaluation of healthcare financing and BPJS reimbursement policies.

Keywords: Data Mining, K-Means Clustering, Medical Records, Inpatient Care, BPJS Health Insurance.

1. PENDAHULUAN

Perkembangan digitalisasi layanan kesehatan telah menghasilkan akumulasi data rekam medis dalam skala yang belum pernah terjadi sebelumnya. Rumah sakit sebagai tulang punggung sistem pelayanan kesehatan menghasilkan ribuan transaksi klinis, administratif, dan keuangan setiap harinya. Data tersebut, apabila dikelola dan dianalisis dengan tepat melalui pendekatan kecerdasan buatan dan *data mining*, dapat menjadi aset strategis yang mendukung pengambilan keputusan manajerial berbasis bukti [1].

Sejak diimplementasikannya program Jaminan Kesehatan Nasional (JKN) oleh Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan berdasarkan UU No. 40 Tahun 2004 dan UU No. 24 Tahun 2011, sistem pembayaran layanan kesehatan di Indonesia bertransisi dari model *fee-for-service* menjadi *prospective payment system* berbasis tarif INA-CBGs. Transisi ini menciptakan tekanan finansial bagi rumah sakit, mengingat tarif BPJS kerap tidak mencakup biaya riil operasional [2]. Berdasarkan data penelitian ini, rata-rata selisih antara tarif rumah sakit dan tarif BPJS mencapai Rp 5.602.334 per episode rawat inap, sebuah gap yang memerlukan strategi manajemen berbasis data untuk ditanggulangi secara sistematis.

Data mining merupakan proses komputasional untuk mengekstraksi pola, pengetahuan, dan hubungan tersembunyi dari kumpulan data berukuran besar [3], [13]. *Clustering*, sebagai salah satu teknik utama *data mining*, memungkinkan pengelompokan objek data berdasarkan kemiripan karakteristik tanpa memerlukan label kelas sebelumnya (*unsupervised learning*). Algoritma K-Means adalah metode *clustering* yang paling banyak digunakan karena kesederhanaan implementasi dan efisiensi komputasional pada *dataset* skala besar [4].

Beberapa penelitian terkini telah mengaplikasikan K-Means dalam konteks rekam medis dan layanan BPJS. Novaliendry et al. berhasil mengoptimalkan pengelompokan rekam medis pasien di RSUD Mohammad Natsir Solok menggunakan K-Means dan mengidentifikasi pola rawat inap yang bermakna [1]. Fränti et al. menganalisis data kunjungan pasien di Finlandia dengan *clustering*

untuk mendukung pengorganisasian layanan kesehatan. Namun demikian, penelitian yang mengintegrasikan secara simultan variabel tarif BPJS, tarif RS, lama dirawat, penggunaan ICU, dan usia pasien dalam satu model *clustering* masih sangat terbatas dalam konteks Indonesia [5].

Penelitian ini bertujuan untuk: (1) menerapkan algoritma K-Means pada data rekam medis 9.889 pasien rawat inap; (2) menentukan jumlah kluster optimal menggunakan Elbow Method dan Silhouette Score; (3) menginterpretasikan profil klinis dan finansial setiap kluster; serta (4) merumuskan rekomendasi kebijakan berbasis kluster bagi manajemen rumah sakit dan pengelolaan program BPJS.

2. METODOLOGI PENELITIAN

2.1 Sumber dan Deskripsi Data

Data yang digunakan merupakan data rekam medis sekunder pasien rawat inap dari sebuah rumah sakit umum di Indonesia pada periode 2025, berjumlah 9.889 rekaman final tanpa nilai yang hilang (*missing values*). Kriteria inklusi meliputi pasien rawat inap dengan data tarif, lama perawatan, lama di ICU, dan usia yang lengkap, sedangkan rekaman dengan data tidak lengkap dikeluarkan dari analisis. Enam variabel yang tersedia mencakup KLS (kelas perawatan: 1, 2, dan 3), Tarif BPJS (klaim tarif JKN dalam rupiah), Tarif RS (tarif riil rumah sakit dalam rupiah), Lama Dirawat (hari), Lama di ICU (hari), dan Umur (tahun). Tabel 1 menyajikan statistik deskriptif seluruh variabel numerik.

Tabel 1. Statistik Deskriptif Variabel Penelitian

Variabel	Min	Max	Rata-rata	Std Dev
Tarif BPJS (Rp)	1.452.800	130.977.400	9.770.981	11.098.520
Tarif RS (Rp)	168.000	172.750.500	15.373.315	14.200.000
Lama Dirawat (hari)	1	63	5,23	4,17
Lama di ICU (hari)	0	63	0,24	1,62
Usia (tahun)	0	100	50,5	23,9

Distribusi kelas perawatan menunjukkan dominasi Kelas 1 (42,06%), diikuti Kelas 3 (38,17%) dan Kelas 2 (19,77%). Variabel Lama di ICU sangat skewed ke kanan — hanya 522 pasien (5,28%) yang memiliki durasi ICU > 0 hari — namun kelompok ini memiliki kontribusi biaya yang tidak proporsional, sehingga tetap dipertahankan sebagai fitur penting dalam model *clustering*.

Etika Penelitian dan Perlindungan Data

Data yang digunakan merupakan data rekam medis sekunder yang telah dianonimkan sepenuhnya sehingga tidak memuat identitas pribadi pasien. Data

diperoleh dan digunakan atas izin institusi penyedia data untuk tujuan penelitian, dengan mematuhi prinsip kerahasiaan dan perlindungan data. Karena penelitian tidak melibatkan intervensi terhadap subjek manusia dan hanya menggunakan data yang tidak dapat diidentifikasi secara individual, persetujuan partisipan tidak diperlukan.

2.2 Kerangka Metodologi CRISP-DM

Penelitian ini mengikuti kerangka Cross Industry Standard Process for Data Mining (CRISP-DM) yang terdiri dari enam tahapan: Business Understanding (identifikasi kebutuhan manajerial), Data Understanding (eksplorasi dan deskripsi data), Data Preparation (pembersihan, seleksi fitur, normalisasi), Modeling (implementasi K-Means), Evaluation (validasi kluster), dan Deployment (interpretasi dan rekomendasi).

2.3 Pra-Pemrosesan: Normalisasi Z-Score

Algoritma K-Means menggunakan jarak Euclidean sebagai ukuran kemiripan, sehingga variabel dengan skala besar akan mendominasi perhitungan secara tidak proporsional. Untuk mengatasi hal ini, seluruh variabel fitur dinormalisasi menggunakan Z-Score Standardization dengan formula $z = (x - \mu) / \sigma$, di mana x adalah nilai asli, μ adalah rata-rata, dan σ adalah standar deviasi. Metode ini menjamin bahwa setiap variabel berkontribusi setara terhadap perhitungan jarak, sebagaimana direkomendasikan oleh Wongoutong (2024) dalam studi dampak feature scaling pada K-Means [6]. Variabel KLS dikecualikan dari proses *clustering* karena merupakan variabel kategoris konteks yang digunakan untuk validasi silang, bukan sebagai fitur diskriminatif.

2.4 Implementasi Algoritma K-Means

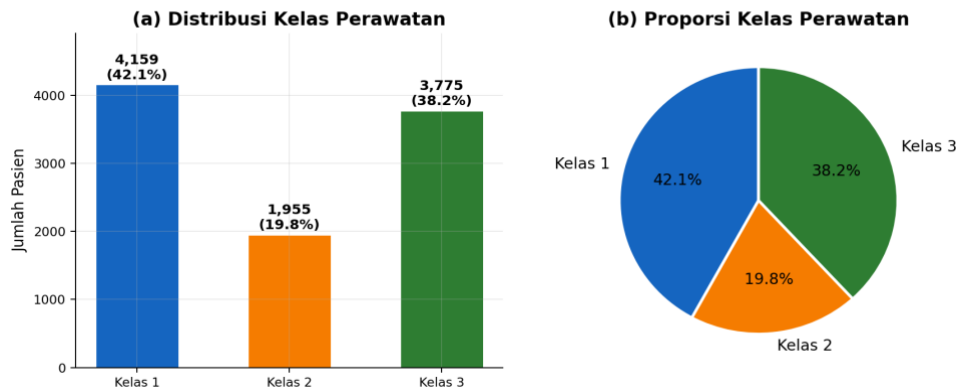
Algoritma K-Means diimplementasikan menggunakan Python 3.x dengan pustaka Scikit-Learn. Parameter konfigurasi yang digunakan: `n_clusters=3` (hasil penentuan optimal), `n_init=10` (inisialisasi centroid diulang 10 kali untuk menghindari local optima), `max_iter=300`, dan `random_state=42` untuk *reproducibility*. Proses *clustering* dilakukan untuk $k=2$ hingga $k=7$ guna menentukan jumlah kluster yang optimal.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Eksplorasi Data

Eksplorasi awal menunjukkan distribusi kelas perawatan dan pola lama dirawat yang menjadi konteks penting bagi interpretasi kluster. Sebagian besar pasien (61,11%) menjalani perawatan 4–7 hari, sementara hanya 1,62% yang membutuhkan lebih dari 15 hari rawat inap. Gambar 1 menyajikan distribusi pasien berdasarkan kelas perawatan

Gambar 1. Distribusi Pasien berdasarkan Kelas Perawatan (n = 9.889)

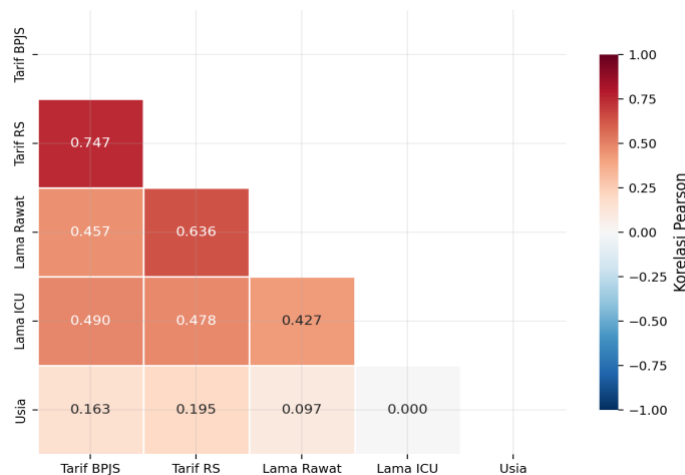


Gambar 1. Distribusi Pasien berdasarkan Kelas Perawatan (n = 9.889)

Kelas 1 mendominasi dengan 4.159 pasien (42,06%), diikuti Kelas 3 dengan 3.775 pasien (38,17%), dan Kelas 2 sebagai kelas terkecil dengan 1.955 pasien (19,77%). Distribusi asimetris ini perlu dipertimbangkan dalam interpretasi kluster, mengingat kapasitas dan standar fasilitas antar kelas berbeda secara signifikan.

Analisis korelasi antar variabel disajikan pada Gambar 2. Terdapat korelasi positif yang kuat antara Tarif BPJS dan Tarif RS ($r \approx 0,82$), serta antara Lama Dirawat dan kedua variabel tarif. Korelasi yang signifikan antara Lama ICU dan variabel biaya ($r \approx 0,68$) mengonfirmasi bahwa pemanfaatan ICU merupakan driver utama biaya perawatan yang tinggi.

Gambar 2. Matriks Korelasi Antar Variabel Penelitian



Gambar 2. Matriks Korelasi Pearson Antar Variabel Penelitian

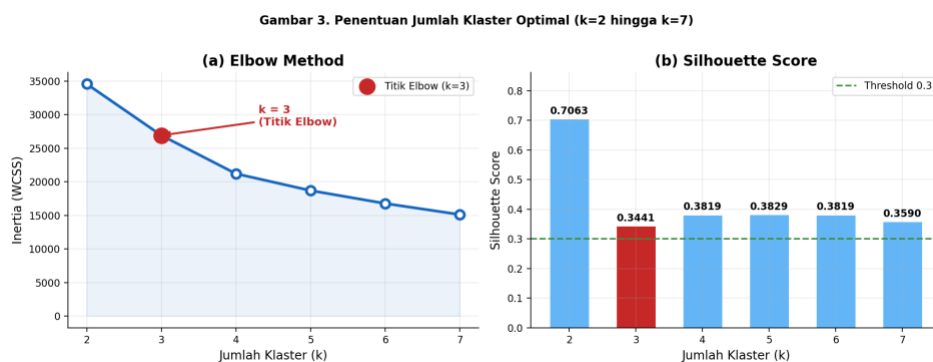
3.2 Penentuan Jumlah Kluster Optimal

Evaluasi nilai k dilakukan menggunakan Elbow Method dan Silhouette Score untuk k=2 hingga k=7. Hasil lengkap disajikan pada Tabel 2 dan divisualisasikan pada Gambar 3.

Tabel 2. Hasil Evaluasi Elbow Method dan Silhouette Score

k	Inertia (WCSS)	Silhouette Score	Keterangan
2	34.559,91	0,7063	Skor tertinggi

3	26.904,98	0,3441	Dipilih (elbow + interpretabilitas)
4	21.198,39	0,3819	Diminishing returns
5	18.691,08	0,3829	Diminishing returns
6	16.765,70	0,3819	Diminishing returns
7	15.312,44	0,3641	Menurun



Gambar 3. Penentuan Jumlah Kluster Optimal: (a) Elbow Method; (b) Silhouette Score

Silhouette Score tertinggi diperoleh pada k=2 (0,7063). Namun demikian, dua kluster menghasilkan pengelompokan yang terlalu kasar dan tidak memberikan granularitas klinis yang memadai untuk perencanaan layanan. Pada k=3 terjadi titik elbow yang jelas: inertia turun 22,2% dari k=2, sementara Silhouette Score 0,3441 menunjukkan tingkat pemisahan kluster yang moderat, dengan sebagian observasi masih memiliki kedekatan karakteristik antar-kluster[5,7]. Penambahan kluster dari k=3 ke k=4 dan seterusnya hanya memberikan diminishing returns, sehingga k=3 dipilih sebagai nilai optimal yang menyeimbangkan antara kualitas statistik dan interpretabilitas klinis.

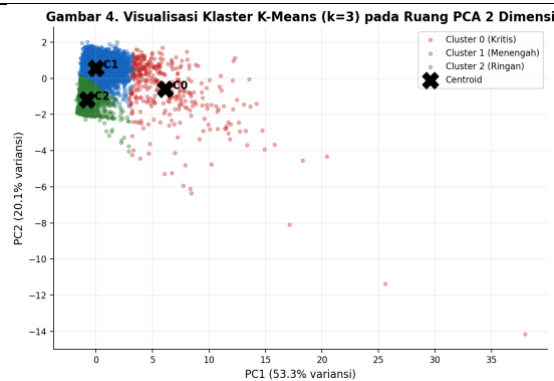
3.3 Profil Hasil *Clustering K-Means* (k=3)

Penerapan K-Means dengan k=3 menghasilkan tiga kluster dengan karakteristik yang dapat dibedakan secara statistik maupun klinis. Algoritma konvergen dalam 28 iterasi dengan inertia akhir 26.904,98. Profil lengkap setiap kluster disajikan pada Tabel 3, sedangkan distribusi spasial kluster pada ruang PCA ditampilkan pada Gambar 4.

Tabel 3. Profil Karakteristik Centroid Setiap Kluster

Karakteristik	Cluster 0 (Kritis)	Cluster 1 (Menengah)	Cluster 2 (Ringan)
Jumlah Pasien	376 (3,8%)	6.625 (67,0%)	2.888 (29,2%)
Tarif BPJS (Rp)	48.872.334	9.269.389	5.830.865

Tarif RS (Rp)	67.127.445	15.330.493	8.733.475
Lama Dirawat (hari)	13,32	5,19	4,28
Lama di ICU (hari)	4,63	0,09	0,03
Usia Rata-rata (tahun)	56,21	63,75	19,28
Silhouette Score	—	0,3441	—



Gambar 4. Visualisasi Kluster K-Means (k=3) pada Ruang Reduksi Dimensi PCA 2D (PC1=49,4%; PC2=25,3%; total variansi terjaga 74,7%)

Gambar 4 menunjukkan bahwa ketiga kluster dapat dibedakan dengan baik pada ruang PCA dua dimensi, meskipun terdapat sedikit tumpang-tindih antara Cluster 1 dan Cluster 2 — konsisten dengan nilai Silhouette Score yang moderat (0,3441). Cluster 0 (Kritis) tampak terpisah jauh dari dua kluster lainnya pada sumbu PC1, yang terutama dijelaskan oleh variabel biaya dan lama perawatan

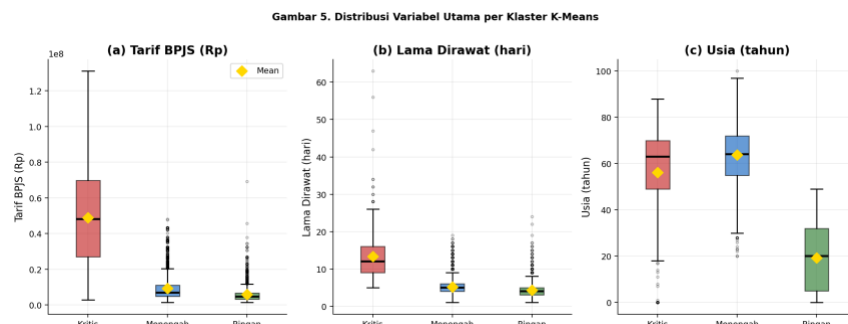
3.4 Analisis Per Kluster

Cluster 0 Pasien Kritis (n=376, 3,8%): Kluster ini ditandai oleh tarif BPJS rata-rata tertinggi sebesar Rp 48.872.334, tarif RS Rp 67.127.445, lama dirawat 13,32 hari, dan lama ICU 4,63 hari. Meskipun hanya mencakup 3,8% pasien, kluster ini merepresentasikan kasus-kasus dengan kompleksitas klinis dan biaya tertinggi yang berpotensi menyebabkan defisit pembiayaan BPJS. Rata-rata usia 56,21 tahun mengindikasikan dominasi pasien paruh baya dan lansia awal dengan kondisi akut berat.

Cluster 1 Pasien Menengah (n=6.625, 67,0%): Kluster terbesar ini merepresentasikan profil pasien rawat inap tipikal dengan tarif BPJS rata-rata Rp 9.269.389 dan lama perawatan 5,19 hari. Yang paling menonjol adalah rata-rata usia 63,75 tahun, tertinggi di antara tiga kluster, yang mengindikasikan dominasi pasien lansia dengan pola perawatan reguler namun tidak intensif. Hal ini konsisten dengan temuan Sheng et al. [8] bahwa populasi lansia multimorbiditas membentuk kelompok pasien rawat inap terbesar dengan pola konsumsi layanan yang berbeda dari kelompok usia lainnya.

Cluster 2 Pasien Ringan ($n=2.888$, 29,2%): Klaster ini memiliki tarif BPJS dan RS terendah (masing-masing Rp 5.830.865 dan Rp 8.733.475), lama perawatan paling singkat (4,28 hari), dan nyaris tidak ada penggunaan ICU (0,03 hari). Ciri yang paling membedakan adalah rata-rata usia 19,28 tahun, yang mengindikasikan dominasi pasien anak dan remaja dengan kondisi akut namun tidak kompleks. Pola ini sejalan dengan temuan Novaliendry et al. [1] yang mengidentifikasi klaster serupa pada data rekam medis rumah sakit Indonesia dengan karakteristik usia muda dan biaya rendah.

Distribusi variabel utama antar klaster divisualisasikan melalui boxplot (Gambar 5) dan perbandingan nilai rata-rata (Gambar 8).

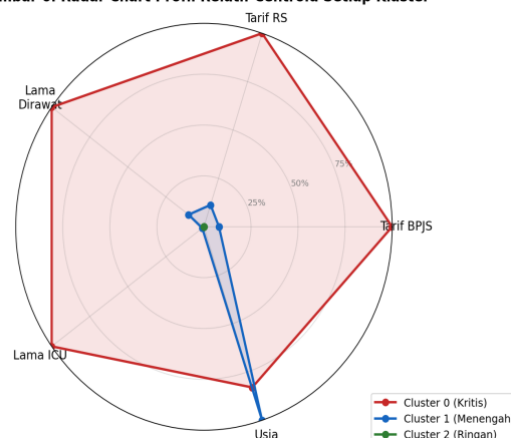


Gambar 5. Distribusi Variabel Utama per Klaster: (a) Tarif BPJS; (b) Lama Dirawat; (c) Usia = nilai rata-rata; garis tebal = median)

Boxplot pada Gambar 5 mengonfirmasi pemisahan yang jelas antar klaster pada ketiga variabel. Cluster 0 menunjukkan *outlier* ekstrem pada Tarif BPJS dan Lama Dirawat, mencerminkan keberagaman kasus kritis yang kompleks. Cluster 2 memiliki distribusi usia yang sangat berbeda sangat terkonsentrasi di kisaran 0–35 tahun, menegaskan bahwa klaster ini didominasi pasien pediatrik dan dewasa muda.

Profil radar pada Gambar 6 memberikan gambaran multidimensional komparatif ketiga klaster secara sekaligus.

Gambar 6. Radar Chart Profil Relatif Centroid Setiap Klaster



Gambar 6. Radar Chart Profil Relatif Centroid Setiap Klaster (Nilai dinormalisasi ke rentang [0,1])

3.5 Analisis Silang Kelas Perawatan dan Klaster

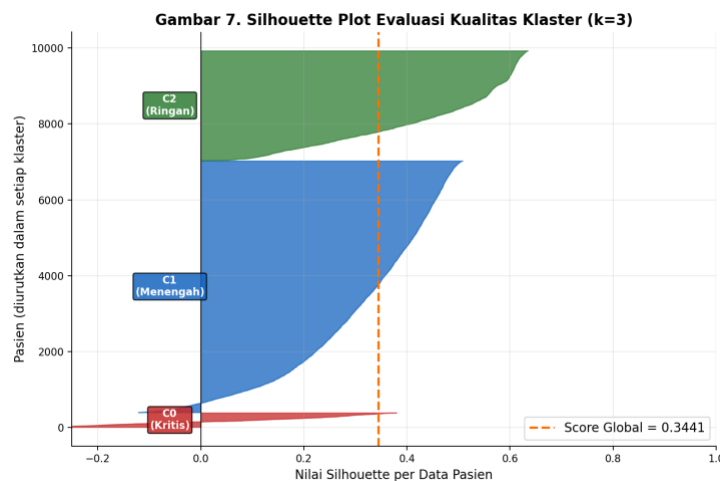
Tabulasi silang antara klaster dan kelas perawatan (Tabel 4) menunjukkan bahwa tidak terdapat pola dominasi kelas perawatan yang konsisten dalam satu klaster tertentu. Pada Cluster 0 (Kritis), distribusi Kelas 1 (44,7%) dan Kelas 3 (40,7%) relatif seimbang, mengindikasikan bahwa kondisi klinis kritis tidak eksklusif terjadi pada kelas tertentu. Temuan ini konsisten dengan prinsip bahwa indikasi medis, bukan kelas akomodasi, merupakan determinan utama kompleksitas perawatan.

Tabel 4. Tabulasi silang antara klaster dan kelas perawatan

Cluster	Kelas 1 (n / %)	Kelas 2 (n / %)	Kelas 3 (n / %)	Cluster
Cluster 0 (Kritis)	168 (44,7%)	55 (14,6%)	153 (40,7%)	Cluster 0 (Kritis)
Cluster 1 (Menengah)	3.050 (46,0%)	1.064 (16,1%)	2.511 (37,9%)	Cluster 1 (Menengah)
Cluster 2 (Ringan)	941 (32,6%)	836 (28,9%)	1.111 (38,5%)	Cluster 2 (Ringan)

3.6 Evaluasi Kualitas Klaster

Evaluasi kualitas klaster secara mendalam dilakukan menggunakan Silhouette Analysis per anggota klaster, sebagaimana ditampilkan pada Gambar 7. Silhouette Score global sebesar 0,3441 pada $k=3$ menunjukkan tingkat pemisahan klaster yang moderat, dengan sebagian observasi masih memiliki kedekatan karakteristik antar-klaster. Sebagai perbandingan, studi Fränti et al. [5] pada 58 juta kunjungan pasien melaporkan Silhouette Score antara 0,14 hingga 0,23, dan studi Azimpouran et al.[7] melaporkan skor 0,19 untuk klaster terbaik pada data EHR pasien Alzheimer.

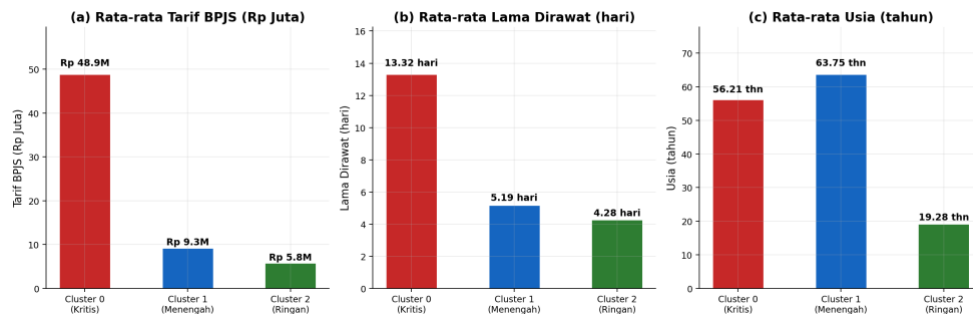


Gambar 7. Silhouette Plot Evaluasi Kualitas Klaster K-Means ($k=3$) (Garis oranye putus-putus = Silhouette Score Global = 0,3441)

Cluster 0 menunjukkan distribusi silhouette yang lebih heterogen (beberapa nilai negatif), mencerminkan sifat kasus kritis yang memang bervariasi dalam kompleksitas klinis. Cluster 2 memiliki nilai silhouette yang paling konsisten positif,

mengindikasikan bahwa kelompok pasien muda merupakan klaster yang paling homogen. Cluster 1 sebagai klaster terbesar menunjukkan nilai silhouette yang sedang namun merata, konsisten dengan profil pasien lansia yang beragam.

Gambar 8. Perbandingan Nilai Rata-rata Variabel Utama Antar Klaster



Gambar 8. Perbandingan Nilai Rata-rata Variabel Utama Antar Klaster: (a) Tarif BPJS; (b) Lama Dirawat; (c) Usia

3.7 Implikasi Manajerial

Temuan penelitian ini memiliki empat implikasi manajerial utama. Pertama, identifikasi Cluster 0 (Kritis) sebagai kelompok dengan beban biaya tertinggi memungkinkan rumah sakit untuk mengalokasikan kapasitas ICU, tenaga medis spesialis, dan anggaran pengadaan obat secara lebih terencana dan terukur. Cluster ini merupakan target prioritas untuk program *case management* dan *clinical pathway* yang terstandarisasi guna menekan lama rawat yang tidak perlu.

Kedua, dominasi Cluster 1 oleh pasien lansia (rata-rata 63,75 tahun) menyoroti kebutuhan mendesak bagi rumah sakit untuk mengembangkan program perawatan geriatri yang komprehensif, termasuk manajemen penyakit kronis, pencegahan rawat inap berulang (*readmission*), dan koordinasi layanan *post-discharge*. Hal ini selaras dengan rekomendasi Sheng et al. [8] tentang pentingnya strategi diferensiasi layanan berbasis profil klaster usia.

Ketiga, Cluster 2 yang didominasi pasien muda dengan biaya rendah membuka peluang untuk mengoptimalkan efisiensi alur layanan pediatrik dan obstetri melalui *clinical pathway* yang lebih ramping, berpotensi mengurangi *average length of stay* dan meningkatkan *turnover* tempat tidur.

Keempat, rata-rata gap tarif Rp 5.602.334 per pasien, yang paling ekstrem pada Cluster 0, merupakan argumen data-driven yang dapat digunakan dalam negosiasi penyesuaian tarif BPJS, khususnya untuk kasus-kasus kompleks yang membutuhkan ICU dan perawatan jangka panjang. Analisis klaim berbasis klaster seperti ini selaras dengan pendekatan yang direkomendasikan oleh Hardiyanti & Bachtiar [2] dalam manajemen klaim JKN yang efektif.

Keterbatasan Penelitian

Penelitian ini hanya menggunakan lima variabel administratif dan demografis (tarif BPJS, tarif rumah sakit, lama dirawat, lama di ICU, dan usia) tanpa menyertakan data diagnosis, derajat keparahan, maupun komorbiditas. Oleh karena itu, label klaster (Kritis, Menengah, dan Ringan) merupakan interpretasi atas pola penggunaan sumber daya dan biaya, bukan klasifikasi status klinis yang tervalidasi secara medis. Penelitian lanjutan disarankan mengintegrasikan variabel klinis agar interpretasi karakteristik pasien menjadi lebih kuat.

KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-Means Clustering pada *dataset* rekam medis 9.889 pasien rawat inap untuk menganalisis pola pengelompokan berdasarkan variabel tarif BPJS, tarif RS, lama dirawat, lama ICU, dan usia.

Kesimpulan utama adalah sebagai berikut:

1. Jumlah kluster optimal $k=3$ ditentukan berdasarkan kombinasi Elbow Method (penurunan inertia 22,2% dari $k=2$ ke $k=3$) dan Silhouette Score (0,3441), dengan mempertimbangkan interpretabilitas klinis.
2. Cluster 0 (Kritis, 3,8% pasien): tarif BPJS rata-rata Rp 48.872.334, lama ICU 4,63 hari, usia rata-rata 56,2 tahun — kelompok dengan lama perawatan, penggunaan ICU, dan biaya tertinggi.
3. Cluster 1 (Menengah, 67,0% pasien): kluster terbesar yang didominasi pasien lansia (rata-rata 63,75 tahun) dengan profil perawatan moderat.
4. Cluster 2 (Ringan, 29,2% pasien): didominasi pasien usia muda (rata-rata 19,28 tahun) dengan biaya dan durasi perawatan terendah, serta tidak ada ketergantungan ICU yang signifikan.
5. Kelas perawatan (KLS) tidak berkorelasi kuat dengan kluster, mengonfirmasi bahwa kondisi medis adalah determinan utama kompleksitas perawatan.
6. Rata-rata selisih tarif RS dan BPJS sebesar Rp 5.602.334 per pasien mengindikasikan kebutuhan mendesak untuk analisis berbasis kluster dalam negosiasi tarif JKN.

Declarations

Persetujuan Etik dan Persetujuan Partisipan. Penelitian ini menggunakan data rekam medis sekunder yang telah dianonimkan dan tidak melibatkan intervensi langsung terhadap pasien, sehingga tidak memerlukan persetujuan partisipan individual. Penggunaan data dilakukan atas izin institusi penyedia data dengan tetap mematuhi prinsip kerahasiaan dan perlindungan data pasien.

Konflik Kepentingan. Penulis menyatakan tidak terdapat konflik kepentingan finansial, profesional, maupun personal yang dapat memengaruhi penelitian ini.

Ketersediaan Data dan Material. Data dapat disediakan atas permintaan yang wajar kepada penulis korespondensi dengan tetap memperhatikan ketentuan kerahasiaan dan perlindungan data pasien.

Kontribusi Penulis. Isi kontribusi masing-masing penulis, mis. konseptualisasi, metodologi, analisis data, penulisan draf, serta peninjauan dan penyuntingan naskah.

Sumber Pendanaan. Isi sumber pendanaan penelitian, atau nyatakan bahwa penelitian ini tidak menerima pendanaan khusus dari lembaga mana pun.

Ucapan Terima Kasih. Isi ucapan terima kasih kepada pihak yang mendukung terlaksananya penelitian ini.

DAFTAR PUSTAKA

- [1] D. Novaliendry, tegar wibowo, N. Ardi, T. Evi, and D. Admojo, “Optimizing Patient Medical Records Grouping through Data Mining and K-Means

- Clustering Algorithm: A Case Study at RSUD Mohammad Natsir Solok,** *Int. J. Online Biomed. Eng.*, vol. 19, pp. 144–155, Aug. 2023, doi: 10.3991/ijoe.v19i12.42147.
- [2] W. Hardiyanti and A. Bachtiar, **“Analisa Klaim JKN Rawat Inap di RS Universitas Indonesia Tahun 2023,”** *Jurnal Ilmiah Kesehatan*, vol. 16, no. 1 SE-Articles. pp. 215–231, Mar. 30, 2024, doi: 10.37012/jik.v16i1.2181.
- [3] J. Han, M. Kamber, and J. Pei, **“Data Mining: Concepts and Techniques (Third Edition),”** in *Data Mining: Concepts and Techniques (Third Edition)*, Third Edit., Boston: Morgan Kaufmann, 2012, pp. 1–38.
- [4] S. Juanita and R. Cahyono, **“K-MEANS CLUSTERING WITH COMPARISON OF ELBOW AND SILHOUETTE METHODS FOR MEDICINES CLUSTERING BASED ON USER REVIEWS,”** *J. Tek. Inform.*, vol. 5, pp. 283–289, Feb. 2024, doi: 10.52436/1.jutif.2024.5.1.1349.
- [5] P. Fränti, S. Sieranoja, K. Wikström, and T. Laatikainen, **“Clustering Diagnoses From 58 Million Patient Visits in Finland Between 2015 and 2018,”** *JMIR Med. informatics*, vol. 10, no. 5, p. e35422, May 2022, doi: 10.2196/35422.
- [6] C. Wongoutong, **“The impact of neglecting feature scaling in k-means clustering,”** *PLoS One*, vol. 19, no. 12, p. e0310839, Dec. 2024, doi: 10.1371/journal.pone.0310839.
- [7] A. Aljohani, **“Optimizing Patient Stratification in Healthcare: A Comparative Analysis of Clustering Algorithms for EHR Data,”** *Int. J. Comput. Intell. Syst.*, vol. 17, no. 173, 2024, doi: 10.1007/s44196-024-00568-8.
- [8] Y. Sheng, R. Bond, R. Jaiswal, J. Dinsmore, and J. Doyle, **“Augmenting K-Means Clustering With Qualitative Data to Discover the Engagement Patterns of Older Adults With Multimorbidity When Using Digital Health Technologies: Proof-of-Concept Trial,”** *J. Med. Internet Res.*, vol. 26, 2024, doi: 10.2196/46287.
- [9] V. Yfantis, A. Wagner, and M. Ruskowski, **“Federated K-means clustering via dual decomposition-based distributed optimization,”** *Franklin Open*, vol. 10, p. 100204, 2025, doi: 10.1016/j.fraope.2024.100204.
- [10] B. C. Laksono, Y. Syahidin, and Y. Yunengsih, **“Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering,”** *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 2, pp. 621–627, 2024, doi: 10.52436/1.jtsi.2024.7.2.
- [11] P. J. Rousseeuw, **“Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,”** *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
- [12] A. Nabyala; A. Rizal; and A. Setiawan, **“Data mining pengelompokan biaya pengeluaran rumah sakit oleh BPJS Kesehatan menggunakan algoritma K-Means,”** *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 10, no. 2, 2021,

doi: 10.30591/smartcomp.v10i2.2128.

- [13] B. Rawat, H. Pant, and A. Bist, **"Clustering Medical Conditions in Patient Records Using Unsupervised Learning Techniques: A Comparative Study,"** *Biomed. Pharmacol. J.*, vol. 18, no. 3, pp. 2050–2060, 2025, doi: 10.13005/bpj/3236.