

Segmentasi Rumah Sakit Indonesia Berbasis Data Menggunakan Faktor Laten dengan Unsupervised Machine Learning

Aisha Gemala Jondya^{1,*}, Reny Fitriani¹, Akbar Shiddiq², Aditya Nugraha Pratama¹

¹Program Studi Teknik Informatika, Fakultas Teknik, Universitas Riau, Indonesia

²Program Profesi Dokter, Fakultas Kedokteran Universitas Riau, Indonesia

Email: ^{1*}aisha.gemala@lecturer.unri.ac.id, ²reny.fitriyani1228@lecturer.unri.ac.id, ³akbar.shiddiq@lecturer.unri.ac.id,

⁴aditya.nugraha0467@student.unri.ac.id

Email Penulis Korespondensi: aisha.gemala@lecturer.unri.ac.id*

Submitted: 29/08/2026; Accepted: 28/09/2026; Published: 30/09/2026

Abstrak – Data rumah sakit yang dipublikasikan melalui Sistem Informasi Rumah Sakit Kementerian Kesehatan (SIRS Kemkes) memuat ribuan rumah sakit dengan ratusan variabel yang saling berkaitan. Meskipun data tersebut berpotensi menggambarkan keragaman karakteristik rumah sakit, segmentasi berbasis banyak indikator dapat menghasilkan representasi yang kompleks dan rentan terhadap nilai ekstrem. Penelitian yang memanfaatkan faktor laten sebagai dasar segmentasi rumah sakit secara nasional juga masih terbatas. Penelitian ini mengisi kesenjangan tersebut melalui integrasi Exploratory Factor Analysis (EFA) dan K-Means clustering untuk membentuk segmentasi rumah sakit berbasis data. Analisis dilakukan pada 3.292 rumah sakit menggunakan 35 skor faktor laten yang diperoleh dari 785 fitur numerik. Analisis awal menunjukkan bahwa nilai ekstrem pada domain layanan menyebabkan segmentasi tidak representatif karena hanya memisahkan satu rumah sakit sebagai outlier. Oleh karena itu, diterapkan winsorization persentil 1%–99% dan standarisasi sebelum pemodelan. Jumlah kluster dievaluasi pada k=2 hingga k=10 menggunakan Elbow Method, Silhouette Score, Calinski-Harabasz, dan Davies-Bouldin. Hasil evaluasi dan pertimbangan substantif menghasilkan tiga kluster, yaitu kluster berkapasitas dasar (64,9%), kluster menengah dengan keunggulan fasilitas penunjang (12,7%), dan kluster dengan layanan serta penunjang operasional intensif (22,4%). Kontribusi penelitian ini adalah menghasilkan kerangka segmentasi berbasis faktor laten yang mengurangi kompleksitas data berdimensi tinggi dan memberikan profil struktural populasi rumah sakit Indonesia untuk mendukung perencanaan sumber daya dan diskusi kebijakan.

Kata Kunci: exploratory factor analysis, klastering, rumah sakit, K-Means, segmentasi

Abstract – Hospital data published through the Hospital Information System of the Indonesian Ministry of Health (SIRS Kemkes) contain thousands of hospitals with hundreds of interrelated variables. Although these data have the potential to describe variations in hospital characteristics, segmentation based on numerous indicators may result in complex representations and be sensitive to extreme values. In addition, studies using latent factors as the basis for nationwide hospital segmentation remain limited. This study addresses this gap by integrating Exploratory Factor Analysis (EFA) and K-Means clustering to develop a data-driven hospital segmentation. The analysis involved 3,292 hospitals using 35 latent factor scores derived from 785 numerical features. Initial analysis showed that extreme values in the service domain resulted in unrepresentative segmentation by separating only one hospital as an outlier. Therefore, 1%–99% percentile winsorization and standardization were applied before clustering. The number of clusters was evaluated from k=2 to k=10 using the Elbow Method, Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The evaluation results and substantive considerations resulted in three clusters: basic-capacity hospitals (64.9%), intermediate hospitals with stronger supporting facilities (12.7%), and hospitals with more intensive services and operational support (22.4%). The contribution of this study is a latent-factor-based segmentation framework that reduces the complexity of high-dimensional data and provides structural profiles of Indonesian hospitals to support resource planning and policy discussions.

Keywords: : clustering, exploratory factor analysis, hospital, K-means, segmentation

1. PENDAHULUAN

Rumah sakit berperan penting dalam sistem pelayanan kesehatan di setiap negara. Kemampuan rumah sakit dalam memberikan pelayanan tidak hanya dipengaruhi oleh jumlah tempat tidur yang tersedia, tetapi juga oleh ketersediaan tenaga kesehatan, fasilitas pelayanan, sarana penunjang, serta berbagai sumber daya lainnya. Kondisi setiap rumah sakit juga tidak sama. Rumah sakit yang memiliki kapasitas tempat tidur yang besar belum tentu memiliki sumber daya dan fasilitas yang sama dengan rumah sakit lainnya [1]. Perbedaan tersebut menyebabkan rumah sakit di Indonesia memiliki karakteristik yang beragam dan perlu dilihat secara lebih menyeluruh berdasarkan data yang tersedia [2].

Pengelompokan rumah sakit di Indonesia pada dasarnya memiliki dasar administratif yang ditetapkan oleh Kementerian Kesehatan [3]. Pendekatan tersebut penting untuk kebutuhan regulasi dan administrasi, tetapi belum sepenuhnya menggambarkan kondisi rumah sakit berdasarkan karakteristik data yang dimilikinya [4]. Rumah sakit dengan klasifikasi yang sama dapat memiliki kondisi kapasitas tempat tidur, jumlah tenaga medis, maupun

ketersediaan sumber daya yang berbeda [5]. Oleh karena itu, diperlukan pendekatan berbasis data yang dapat melihat pola dan karakteristik rumah sakit secara lebih objektif berdasarkan kondisi aktual yang tercatat pada data rumah sakit.

Perkembangan digitalisasi di bidang kesehatan telah meningkatkan layanan rumah sakit [6] dan menghasilkan data dalam jumlah besar melalui berbagai sistem informasi kesehatan. Salah satu sumber data rumah sakit di Indonesia yang dapat dimanfaatkan adalah Sistem Informasi Rumah Sakit (SIRS) [7] yang dikelola oleh Kementerian Kesehatan Republik Indonesia [8]. Data tersebut mencakup berbagai aspek rumah sakit, mulai dari sumber daya manusia, kapasitas tempat tidur, layanan kesehatan, fasilitas penunjang, hingga infrastruktur kesehatan. Dengan memanfaatkan data rumah sakit, pengelompokan tidak hanya didasarkan pada kategori administratif, tetapi juga dapat menunjukkan pola yang terbentuk secara alami dari data. Pendekatan berbasis data menjadi penting karena dapat memberikan gambaran yang lebih rinci mengenai kesamaan dan perbedaan karakteristik antar rumah sakit. Dengan demikian, pengelompokan berbasis data dapat melengkapi klasifikasi administratif yang telah ada dan menjadi tahap awal dalam penerapan pendekatan data *science* untuk analisis rumah sakit di Indonesia.

Meskipun demikian, pemanfaatan data rumah sakit masih menghadapi tantangan berupa tingginya dimensi data dan banyaknya variabel yang saling berkaitan [9]. Salah satu pendekatan yang umum digunakan untuk menyederhanakan data multidimensi yang saling berkorelasi adalah Exploratory Factor Analysis (EFA), yang mereduksi variabel berjumlah besar menjadi sejumlah kecil faktor laten yang lebih mudah diinterpretasikan [10]. Proses EFA telah dilakukan pada penelitian sebelumnya menggunakan data rumah sakit Indonesia yang bersumber dari Sistem Informasi Rumah Sakit (SIRS) menghasilkan 35 faktor laten dari 785 fitur numerik mentah [11].

Penelitian klasterisasi rumah sakit pada umumnya membentuk kelompok langsung dari variabel administratif tanpa reduksi dimensi awal, sehingga rawan bias oleh korelasi antarvariabel. Penelitian ini memanfaatkan skor faktor hasil EFA sebagai representasi ringkas dari 785 variabel asli sehingga klastering dilakukan pada ruang fitur yang lebih terstruktur. Kontribusi utama penelitian ini adalah melanjutkan hasil EFA sebelumnya ke tahap segmentasi K-Means, termasuk penentuan jumlah klaster berbasis evaluasi multi-metrik dan pengujian stabilitas, yang belum dibahas pada penelitian terdahulu.

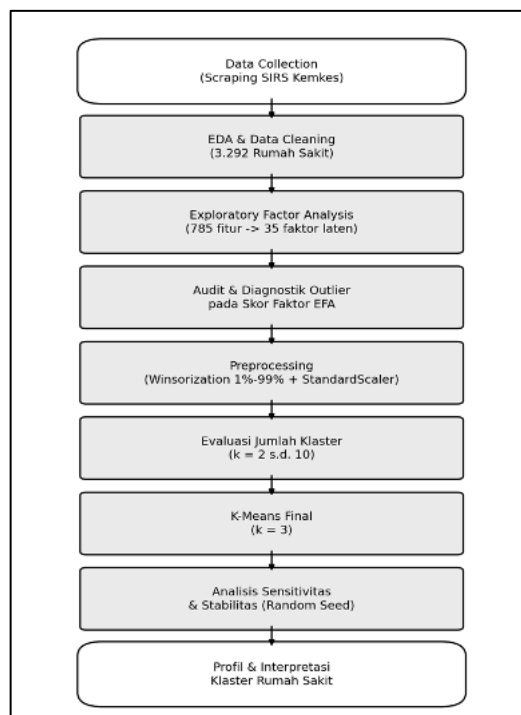
Setelah faktor laten diperoleh melalui Exploratory Factor Analysis (EFA), tahap berikutnya adalah mengelompokkan rumah sakit berdasarkan kemiripan karakteristik yang direpresentasikan oleh skor faktor [12]. Pendekatan clustering dapat digunakan untuk menemukan kelompok rumah sakit yang memiliki pola kapasitas dan sumber daya yang relatif serupa [13] tanpa harus menentukan kelompok tersebut terlebih dahulu [14]. Dengan demikian, hasil pengelompokan diharapkan dapat memberikan segmentasi rumah sakit yang lebih mencerminkan karakteristik data dibandingkan pengelompokan yang hanya didasarkan pada kategori administratif [15]. Dalam penelitian ini, K-Means dipilih sebagai salah satu metode clustering non-hierarkis karena mampu membentuk kelompok berdasarkan tingkat kemiripan antar objek dan relatif sederhana untuk diterapkan pada data berukuran besar [16].

Penggunaan skor faktor sebagai masukan clustering juga memberikan keuntungan karena sejumlah variabel yang saling berkaitan telah diringkas menjadi beberapa faktor yang mewakili informasi utama dari data. Hal ini dapat mengurangi kompleksitas data sekaligus menghasilkan representasi karakteristik rumah sakit yang lebih mudah digunakan dalam proses segmentasi. Namun, sebelum proses clustering dilakukan, karakteristik data perlu diperhatikan agar hasil pengelompokan tidak dipengaruhi oleh perbedaan skala maupun nilai yang terlalu ekstrem. K-Means menggunakan jarak Euclidean dalam menentukan kedekatan antar data, sehingga nilai dengan skala yang lebih besar atau keberadaan outlier dapat memberikan pengaruh yang lebih kuat terhadap pembentukan cluster [17].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk: (1) membentuk segmentasi rumah sakit Indonesia menggunakan algoritma K-Means pada skor faktor hasil EFA, (2) menentukan jumlah klaster optimal melalui evaluasi metrik validasi internal beserta pertimbangan kelayakan substantif, dan (3) menginterpretasikan karakteristik setiap klaster sehingga hasil segmentasi dapat digunakan sebagai bahan diskusi perencanaan sumber daya dan kebijakan rumah sakit.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif untuk melakukan segmentasi rumah sakit berdasarkan karakteristik data yang dimiliki. Tahapan penelitian meliputi penggunaan data dan hasil EFA dari penelitian sebelumnya, audit dan diagnostik data, *preprocessing*, evaluasi jumlah klaster, pemodelan K-Means, analisis sensitivitas dan stabilitas, serta profil dan interpretasi hasil klaster.



Gambar 1. Alur Penelitian Klasterisasi K-Means dengan Skor Faktor Hasil EFA

2.1 Data Penelitian

Data penelitian dikumpulkan dengan Teknik *web scraping* dari Sistem Informasi Rumah Sakit (SIRS) Kementerian Kesehatan Republik Indonesia dan mencakup rumah sakit yang tersebar di seluruh Indonesia. Tahap pengumpulan data, EDA, *data cleaning*, dan Exploratory Factor Analysis (EFA) telah dilakukan pada penelitian sebelumnya. Dari proses tersebut, data awal yang terdiri atas 785 fitur numerik direduksi menjadi 35 skor faktor laten. Skor faktor tersebut mewakili lima domain karakteristik rumah sakit, yaitu Sumber Daya Manusia (SDM), Layanan, Kapasitas Tempat Tidur, Fasilitas Penunjang, dan Penunjang Operasional. Web scraping dilakukan Desember 2025 melalui Dashboard RS Online SIRS Kemkes; RS tidak aktif/tidak lengkap dikeluarkan sehingga diperoleh 3.292 observasi.

Pada penelitian ini, 35 skor faktor hasil EFA digunakan sebagai input utama dalam proses segmentasi menggunakan K-Means [18]. Penentuan jumlah faktor dilakukan menggunakan *Kaiser Criterion* dengan mempertahankan faktor yang memiliki nilai eigenvalue lebih besar dari 1. Selain itu, scree plot digunakan sebagai pendukung untuk mengidentifikasi titik *elbow* yang menunjukkan jumlah faktor yang paling representatif dalam menjelaskan struktur data. Kombinasi kedua pendekatan tersebut digunakan untuk memperoleh jumlah faktor yang optimal pada masing-masing domain analisis. Proses tersebut telah dilakukan pada penelitian sebelumnya. Dengan demikian, penelitian ini melanjutkan proses analisis dari hasil EFA tanpa mengulang kembali tahapan pembentukan faktor. Data identitas rumah sakit, seperti jenis, kelas, kepemilikan, provinsi, dan kabupaten/kota, tidak digunakan dalam perhitungan jarak pada K-Means. Informasi tersebut akan digunakan pada tahap akhir untuk membantu menggambarkan karakteristik dan persebaran rumah sakit pada setiap kluster.

2.2 Diagnostik Data

Sebelum proses klasterisasi dilakukan, 35 skor faktor hasil EFA melalui tahap audit dan diagnostik data [19]. Pemeriksaan dilakukan untuk memastikan kondisi data dan mengidentifikasi kemungkinan adanya nilai ekstrem (*outlier*) yang dapat memengaruhi proses pengelompokan. Pemeriksaan mencakup nilai yang hilang (*missing value*), nilai *infinite*, standar deviasi faktor, duplikasi data, serta distribusi dan rentang nilai pada masing-masing faktor.

Tahap ini menjadi penting karena K-Means menggunakan jarak Euclidean untuk menentukan kedekatan antarobservasi. Nilai ekstrem dapat memberikan pengaruh yang besar terhadap perhitungan jarak dan posisi centroid sehingga berpotensi menghasilkan kluster yang kurang mewakili kondisi mayoritas rumah sakit. Oleh karena itu, hasil audit dan diagnostik digunakan sebagai dasar untuk menentukan perlakuan terhadap nilai ekstrem sebelum model K-Means final dibangun.

2.3 Preprocessing

Tahap *preprocessing* dilakukan untuk menyiapkan data faktor sebelum digunakan dalam proses klasterisasi. Penanganan nilai ekstrem dilakukan menggunakan metode *winsorization* pada persentil 1%–99%. Metode ini

digunakan untuk membatasi pengaruh nilai yang terlalu tinggi atau terlalu rendah tanpa menghapus rumah sakit dari dataset. Dengan demikian, seluruh rumah sakit tetap dipertahankan dalam proses analisis.

Setelah proses *winsorization*, data distandardisasi menggunakan *StandardScaler*. Standardisasi dilakukan agar setiap faktor memiliki skala yang sebanding sehingga faktor tertentu tidak memberikan pengaruh yang lebih besar hanya karena memiliki rentang nilai yang lebih tinggi [20]. Data hasil *preprocessing* selanjutnya digunakan sebagai input pada proses evaluasi jumlah kluster dan pemodelan K-Means.

2.4 Evaluasi Jumlah Kluster

Evaluasi jumlah kluster dilakukan untuk menentukan jumlah kelompok rumah sakit yang paling sesuai dengan karakteristik data. Pengujian dilakukan pada rentang $k=2$ sampai $k=10$. Beberapa metode validasi internal digunakan, yaitu *Within-Cluster Sum of Squares* (Inertia) melalui *Elbow Method*, *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index* [21].

Silhouette Score digunakan untuk melihat seberapa dekat suatu rumah sakit dengan anggota klusternya dibandingkan dengan kluster lain. Nilai yang semakin mendekati 1 menunjukkan pemisahan kluster yang semakin baik. *Calinski-Harabasz Index* digunakan untuk membandingkan penyebaran data antar-kluster dan di dalam kluster, sedangkan *Davies-Bouldin Index* digunakan untuk melihat tingkat kemiripan antar-kluster. Pada *Elbow Method*, perubahan nilai *inertia* diamati untuk melihat titik ketika penambahan jumlah kluster tidak lagi memberikan penurunan yang berarti [22].

Selain nilai dari keempat metrik tersebut, ukuran masing-masing kluster juga diperiksa. Hal ini dilakukan agar jumlah kluster yang dipilih tidak hanya memberikan nilai evaluasi yang baik, tetapi juga menghasilkan kelompok yang memiliki jumlah anggota yang cukup representatif. Hasil dari seluruh evaluasi tersebut digunakan sebagai dasar dalam menentukan jumlah kluster untuk model K-Means final.

2.5 K-Means Final dan Validasi Stabilitas

Model K-Means final diterapkan setelah jumlah kluster ditentukan melalui tahap evaluasi. Data yang digunakan berupa 35 skor faktor yang telah melalui proses *winsorization* dan standardisasi. K-Means mengelompokkan rumah sakit berdasarkan kemiripan karakteristik faktor dengan menggunakan jarak Euclidean terhadap centroid masing-masing kluster.

Model dijalankan menggunakan parameter $n_init=50$ dan $random_state=42$. Parameter n_init digunakan agar algoritma melakukan beberapa inisialisasi centroid dan memilih hasil dengan performa terbaik, sedangkan $random_state$ digunakan untuk menjaga agar proses dapat direproduksi. Hasil pemodelan berupa label kluster untuk setiap rumah sakit yang selanjutnya digunakan pada tahap validasi dan interpretasi.

Validasi stabilitas dilakukan untuk mengetahui apakah struktur kluster yang terbentuk konsisten terhadap perubahan inisialisasi acak pada algoritma K-Means. Pengujian dilakukan dengan mempertahankan jumlah kluster dan data input yang sama, kemudian menjalankan kembali K-Means menggunakan 10 nilai *random seed*, yaitu 0, 1, 7, 21, 42, 99, 123, 2026, 2027, dan 2028. Setiap pengulangan menggunakan $n_init=50$ sehingga perbedaan hasil yang diamati terutama merepresentasikan pengaruh perubahan *random seed* terhadap proses inisialisasi model. Hasil pengelompokan pada $random\ seed=42$ digunakan sebagai label referensi.

Konsistensi hasil pengelompokan antar-*seed* dievaluasi menggunakan Adjusted Rand Index (ARI) dengan membandingkan label pada masing-masing *seed* terhadap label referensi pada $random\ seed=42$. ARI digunakan karena dapat mengukur tingkat kesamaan dua hasil pengelompokan tanpa bergantung pada penamaan numerik kluster. Nilai ARI yang semakin mendekati 1 menunjukkan bahwa struktur dan keanggotaan kluster semakin konsisten terhadap perubahan inisialisasi, sedangkan nilai yang lebih rendah menunjukkan adanya perubahan anggota kluster antar-*seed*. Dengan demikian, stabilitas model dinilai berdasarkan konsistensi label kluster terhadap variasi *random seed*, sedangkan kualitas internal kluster tetap dievaluasi menggunakan metrik validasi yang telah dilakukan pada tahap sebelumnya.

3. HASIL DAN PEMBAHASAN

3.1 Penentuan Jumlah Kluster

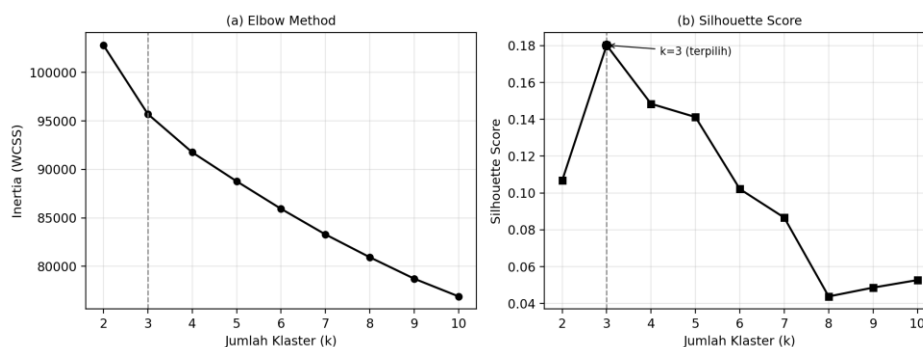
Evaluasi dilakukan untuk menentukan jumlah kluster yang paling sesuai pada rentang $k=2$ hingga $k=10$. Empat metrik digunakan sebagai pertimbangan, yaitu *Inertia* (WCSS), *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index*. Selain kualitas internal kluster, proporsi anggota pada kluster terkecil juga diperhatikan untuk menghindari terbentuknya kelompok yang terlalu kecil.

Tabel 1. Metrik Evaluasi Jumlah Kluster K-Means ($k=2$ s.d. 10)

k	Inertia (WCSS)	Silhouette	Calinski-Harabasz	Davies-Bouldin	Klaster Terkecil (%)
2	102.803,92	0,1068	397,35	2,5193	48,72
3*	95.705,08	0,1802	335,32	2,3881	12,67
4	91.755,61	0,1484	280,28	2,6778	10,69
5	88.771,71	0,1412	244,83	2,4994	3,07
6	85.930,24	0,1021	224,01	2,3990	3,07
7	83.289,27	0,0866	209,90	2,2723	1,15
8	80.932,36	0,0438	198,76	2,2413	1,18
9	78.721,17	0,0486	190,27	2,1629	1,12
10	76.892,26	0,0527	181,77	2,1781	1,12

*k=3 dipilih sebagai model final.

Hasil pada Tabel 1 menunjukkan bahwa nilai *Inertia* menurun seiring bertambahnya jumlah klaster. *Silhouette Score* memberikan hasil yang lebih informatif. Nilai tertinggi diperoleh pada k=3, yaitu 0,1802, sedangkan nilai tersebut menurun pada k≥4. Pada saat yang sama, k=3 menghasilkan klaster terkecil sebesar 12,67% atau 417 rumah sakit, sehingga tidak membentuk kelompok yang terlalu kecil



Gambar 2. (a) Kurva Elbow Method dan (b) Silhouette Score untuk k=2 s.d. 10

Kurva *Elbow* tidak menunjukkan titik tekukan yang sangat tegas. Kondisi ini menunjukkan bahwa penambahan jumlah klaster secara bertahap memang mengurangi variasi dalam klaster, tetapi *Elbow Method* belum memberikan pilihan k yang cukup jelas. Oleh karena itu, pemilihan jumlah klaster tidak didasarkan pada *Inertia* saja. Meskipun k=2 menghasilkan distribusi yang lebih seimbang, solusi tersebut hanya membentuk dua kelompok dan memberikan pemisahan yang lebih sederhana. Berdasarkan kombinasi kualitas pemisahan, ukuran klaster, dan pertimbangan interpretabilitas, k=3 dipilih sebagai jumlah klaster final. Dengan demikian, penelitian selanjutnya menggunakan tiga kelompok rumah sakit sebagai dasar untuk menganalisis karakteristik masing-masing segmen.

Berdasarkan kombinasi hasil evaluasi tersebut, k=3 dipilih sebagai model final. Pemilihan ini terutama didasarkan pada *Silhouette Score* yang paling tinggi di antara kandidat, ukuran setiap klaster yang masih memadai, serta jumlah kelompok yang memungkinkan karakteristik rumah sakit dianalisis secara lebih rinci. Keputusan ini mengikuti aturan multi-metrik dengan *Silhouette Score* sebagai kriteria utama, sedangkan *Calinski-Harabasz* dan *Davies-Bouldin* sebagai pendukung. *Calinski-Harabasz* lebih tinggi pada k=2, tetapi k=2 hanya pemisahan biner yang kurang informatif; perbaikan *Davies-Bouldin* pada k besar juga tidak diikuti kenaikan *Silhouette Score*, mengindikasikan klaster semakin terfragmentasi. Karena itu, k=3 dipilih sebagai kompromi terbaik. Nilai *Silhouette Score* yang relatif rendah juga menunjukkan bahwa terdapat sebagian tumpang tindih karakteristik antar-rumah sakit, sehingga hasil klaster lebih tepat dipahami sebagai segmentasi berdasarkan kemiripan profil, bukan sebagai kelompok yang memiliki batas karakteristik yang sepenuhnya terpisah.

3.2 Stabilitas dan Sensitivitas Model

Setelah jumlah kluster ditetapkan, dilakukan pengujian untuk memastikan bahwa solusi $k=3$ tidak terbentuk hanya karena pengaruh inisialisasi acak K-Means. Model dijalankan menggunakan 10 random seed dengan konfigurasi yang sama, yaitu $n_init=50$ dan $k=3$. Hasil setiap pengulangan dibandingkan dengan hasil model referensi pada $random\ seed=42$ menggunakan Adjusted Rand Index (ARI). Hasil pengujian menunjukkan nilai ARI berada pada rentang 0,9928–1,0000 (rata-rata $[ISI] \pm SD [ISI]$). Nilai yang sangat mendekati 1 menunjukkan bahwa keanggotaan rumah sakit pada ketiga kluster hampir tidak berubah meskipun $random\ seed$ diubah. Dengan demikian, struktur tiga kluster dapat dikatakan stabil terhadap variasi inisialisasi K-Means.

Selain stabilitas model, dilakukan analisis sensitivitas terhadap tahap preprocessing. Hasil perbandingan menunjukkan bahwa penggunaan data tanpa penanganan nilai ekstrem cenderung menghasilkan pengelompokan yang dipengaruhi oleh observasi ekstrem. Setelah dilakukan *winsorization* pada persentil 1–99 dan standarisasi, distribusi anggota kluster menjadi lebih proporsional dan hasilnya lebih sesuai untuk interpretasi segmentasi populasi. Oleh karena itu, konfigurasi preprocessing tersebut digunakan pada model final. Hasil ini memberikan dasar bahwa model yang dipilih tidak hanya memiliki kualitas internal yang memadai, tetapi juga relatif konsisten terhadap perubahan inisialisasi dan tidak terlalu bergantung pada keberadaan nilai ekstrem.

3.3 Karakteristik Kluster Final

Model K-Means dengan $k=3$ menghasilkan segmentasi terhadap **3.292 rumah sakit** ke dalam tiga kelompok. Untuk memudahkan interpretasi, nomor kluster diurutkan berdasarkan rata-rata skor total faktor dari nilai terendah hingga tertinggi. Skor total faktor dihitung sebagai rata-rata dari 35 skor faktor terstandarisasi (z-score) hasil EFA untuk setiap rumah sakit, sehingga nilai positif menunjukkan profil di atas rata-rata populasi rumah sakit secara keseluruhan, sedangkan nilai negatif menunjukkan sebaliknya. Distribusi dan karakteristik umum masing-masing kluster ditunjukkan pada Tabel 2.

Tabel 2. Distribusi dan Karakteristik Dominan Tiap Kluster

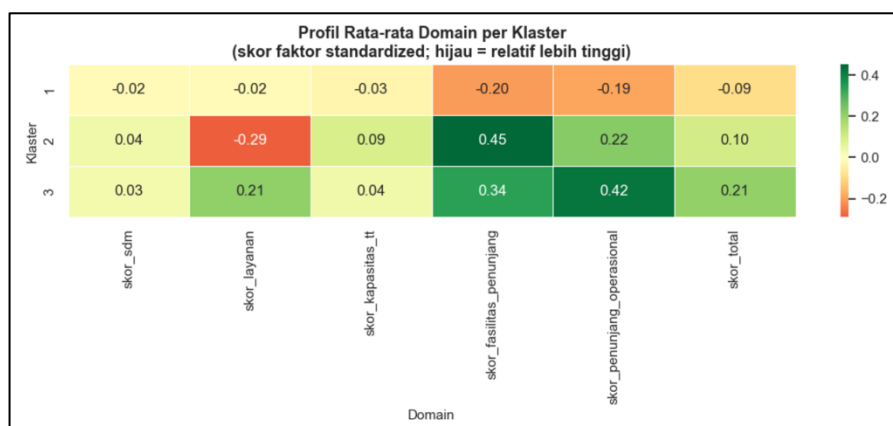
Kluster	Jumlah RS	Persentase	Kelas Dominan	Jenis Dominan	Kepemilikan Dominan
1	2.138	64,95%	C (57,9%)	RS Umum (81,2%)	Swasta/Lainnya (27,8%)
2	417	12,67%	C (47,0%)	RS Umum (88,5%)	Swasta/Lainnya (29,3%)
3	737	22,39%	C (50,1%)	RS Umum (91,7%)	Swasta/Lainnya (29,3%)

Kluster 1 merupakan kelompok terbesar yang terdiri atas 2.138 rumah sakit (64,95%). Kelompok ini didominasi oleh rumah sakit kelas C sebesar 57,9% dan rumah sakit umum sebesar 81,2%. Kluster 2 terdiri atas 417 rumah sakit (12,67%), dengan dominasi rumah sakit kelas C sebesar 47,0% dan rumah sakit umum sebesar 88,5%. Sementara itu, Kluster 3 mencakup 737 rumah sakit (22,39%), dengan rumah sakit kelas C sebesar 50,1% dan rumah sakit umum sebesar 91,7%.

Distribusi tersebut menunjukkan bahwa ketiga kluster memiliki karakteristik administratif yang relatif serupa, terutama karena rumah sakit kelas C dan rumah sakit umum tetap mendominasi seluruh kelompok. Dengan demikian, perbedaan antarkluster tidak dapat dijelaskan hanya berdasarkan kelas atau jenis rumah sakit. Perbedaan yang lebih jelas perlu dilihat melalui profil domain faktor yang digunakan sebagai dasar clustering.

3.4 Profil Domain Antar Kluster

Setelah terbentuk 3 kluster hasil dari proses klustering dengan K-Means, dilakukan analisis kluster terhadap tiap domain. Hal ini untuk melihat seperti apa sebaran kluster terhadap domain SDM, layanan, kapasitas tempat tidur, fasilitas penunjang dan penunjang operasional. Gambar 3 memperlihatkan profil rata-rata domain (standardized) per kluster.



Gambar 3. Profil Rata-rata Domain Per Klaster

Klaster 1 memiliki skor negatif pada seluruh domain dan skor total sebesar -0,092. Nilai terendah terdapat pada domain Fasilitas Penunjang (-0,205) dan Penunjang Operasional (-0,190). Pola tersebut menunjukkan bahwa Klaster 1 memiliki profil relatif lebih rendah dibandingkan rata-rata populasi pada hampir seluruh domain yang dianalisis. Karena mencakup 64,95% rumah sakit, kelompok ini dapat dipandang sebagai profil kapasitas dasar dalam konteks indikator yang digunakan penelitian.

Berbeda dengan Klaster 1, Klaster 2 memiliki skor positif pada sebagian besar domain dengan skor total 0,102. Nilai tertinggi terdapat pada Fasilitas Penunjang (0,452) dan Penunjang Operasional (0,224). Namun, domain Layanan justru memiliki skor paling rendah di antara ketiga klaster, yaitu -0,291. Pola ini menunjukkan bahwa Klaster 2 memiliki karakteristik yang menonjol pada aspek fasilitas dan penunjang operasional, tetapi tidak menunjukkan profil layanan yang sama tingginya. Oleh karena itu, kelompok ini lebih tepat disebut sebagai profil fasilitas penunjang, bukan sekadar kelompok dengan tingkat kapasitas menengah.

Klaster 3 memiliki skor total tertinggi, yaitu 0,208, dengan seluruh domain berada pada nilai positif. Skor tertinggi terdapat pada Penunjang Operasional (0,424) dan Fasilitas Penunjang (0,340), diikuti oleh Layanan (0,210). Pola tersebut menunjukkan bahwa Klaster 3 memiliki profil yang relatif lebih tinggi dan lebih merata pada berbagai domain, terutama pada aspek layanan dan penunjang operasional. Kelompok ini selanjutnya dapat disebut sebagai profil layanan dan operasional relatif tinggi. Hasil tersebut menunjukkan bahwa ketiga klaster tidak hanya berbeda berdasarkan tingkat skor total, tetapi juga memiliki pola domain yang berbeda. Hal ini menjadi penting karena rumah sakit dengan skor total yang serupa dapat memiliki kombinasi karakteristik yang berbeda apabila dilihat pada masing-masing domain.

3.5 Interpretasi dengan Variabel Asli

Untuk melihat apakah pola yang diperoleh dari skor faktor dapat ditemukan kembali pada data awal, dilakukan pemeriksaan terhadap beberapa variabel asli, yaitu total tempat tidur, total layanan, dan total SDM. Median digunakan untuk menggambarkan nilai tipikal setiap kelompok karena data rumah sakit memiliki rentang nilai yang cukup lebar. Tabel 4 menunjukkan median masing-masing variabel terhadap klaster hasil yang terbentuk.

Tabel 4. Median Variabel Asli per Klaster

Klaster	Median Total TT	Median Total Layanan	Median Total SDM
1	95,5	30	204
2	150,0	82	413
3	124,0	67	344

Pola pada variabel asli konsisten dengan interpretasi skor faktor: Klaster 1 memiliki median tempat tidur, layanan, dan SDM paling rendah; Klaster 2 memiliki median tempat tidur dan SDM tertinggi, konsisten dengan keunggulannya pada domain Fasilitas Penunjang dan Penunjang Operasional; sementara Klaster 3 memiliki

median total layanan yang lebih tinggi dibanding Klaster 1, sejalan dengan skor domain Layanan yang lebih tinggi pada klaster tersebut.

3.6 Pembahasan

Hasil penelitian menunjukkan bahwa rumah sakit di Indonesia memiliki profil struktural dan layanan yang heterogen, meskipun sebagian besar kelompok tetap didominasi oleh rumah sakit kelas C dan rumah sakit umum. K-Means menghasilkan tiga profil utama, yaitu profil kapasitas dasar pada 64,95% rumah sakit, profil fasilitas penunjang pada 12,67% rumah sakit, dan profil layanan-operasional relatif tinggi pada 22,39% rumah sakit. Perbedaan tersebut menunjukkan bahwa klasifikasi administratif belum sepenuhnya menggambarkan variasi karakteristik rumah sakit ketika dilihat dari kombinasi sumber daya, layanan, kapasitas, dan fasilitas. Disamping itu, penelitian terkait segmentasi rumah sakit dengan metode *factor analysis* masih sangat minim dilakukan di Indonesia.

Menariknya, hasil pemeriksaan terhadap variabel asli menunjukkan bahwa Klaster 2 memiliki median tempat tidur, layanan, dan SDM tertinggi, tetapi bukan merupakan klaster dengan skor total faktor tertinggi. Sebaliknya, Klaster 3 memiliki skor total faktor tertinggi meskipun median ketiga variabel tersebut berada di bawah Klaster 2. Temuan ini menunjukkan bahwa segmentasi yang dihasilkan tidak hanya mengurutkan rumah sakit berdasarkan ukuran absolut. Perbedaan tersebut muncul karena clustering dilakukan menggunakan skor faktor yang merepresentasikan beberapa domain secara simultan. Dengan demikian, rumah sakit dengan jumlah sumber daya yang besar belum tentu memiliki profil faktor yang paling tinggi pada seluruh dimensi. Hal ini terjadi karena skor faktor merepresentasikan posisi relatif tiap rumah sakit terhadap populasi, sedangkan median variabel asli menggambarkan nilai absolut. Klaster 2 unggul absolut, tetapi Klaster 3 memiliki proporsi layanan-operasional lebih tinggi relatif terhadap kapasitasnya, sehingga label "profil layanan-operasional intensif" merujuk pada keunggulan relatif, bukan nilai absolut tertinggi.

Dari perspektif segmentasi, Klaster 1 menunjukkan kelompok terbesar dengan skor relatif rendah pada hampir seluruh domain. Klaster 2 memiliki karakteristik yang lebih spesifik karena menunjukkan skor tinggi pada Fasilitas Penunjang dan Penunjang Operasional tetapi memiliki skor Layanan paling rendah. Sementara itu, Klaster 3 menunjukkan profil yang relatif tinggi dan lebih merata, terutama pada Layanan, Fasilitas Penunjang, dan Penunjang Operasional. Perbedaan pola tersebut menunjukkan bahwa segmentasi berbasis data dapat memberikan informasi yang lebih rinci dibandingkan pengelompokan yang hanya menggunakan satu karakteristik rumah sakit.

Pendekatan EFA yang dilanjutkan dengan K-Means juga memberikan keuntungan dalam mengurangi kompleksitas data sebelum proses clustering. Sebanyak 35 skor faktor digunakan untuk merepresentasikan karakteristik rumah sakit dalam beberapa domain utama. K-Means kemudian mengelompokkan rumah sakit berdasarkan kemiripan profil tersebut. Dengan pendekatan ini, segmentasi tidak hanya menjawab "berapa besar rumah sakit", tetapi juga "pada aspek apa rumah sakit tersebut memiliki karakteristik yang relatif lebih tinggi atau lebih rendah". Hal ini menjadi dasar yang lebih informatif untuk melakukan analisis lanjutan mengenai karakteristik setiap kelompok rumah sakit.

Meskipun demikian, hasil segmentasi perlu dipahami sesuai dengan cakupan data yang digunakan. Klaster yang terbentuk merepresentasikan profil struktural, sumber daya, dan layanan berdasarkan variabel penelitian, bukan peringkat mutu atau kinerja rumah sakit. Oleh karena itu, rumah sakit yang berada pada klaster dengan skor faktor lebih tinggi tidak dapat secara langsung dianggap memiliki kinerja pelayanan yang lebih baik.

Penelitian ini juga masih menggunakan K-Means sebagai satu metode clustering. Meskipun model menunjukkan stabilitas yang sangat tinggi terhadap perubahan *random seed*, K-Means memiliki karakteristik tertentu dalam membentuk kelompok berdasarkan jarak Euclidean dan cenderung sesuai untuk struktur klaster yang relatif kompak. Penelitian selanjutnya dapat membandingkan hasil tersebut dengan metode clustering lain serta menggunakan indikator mutu dan utilisasi yang independen sebagai validasi eksternal. Dengan demikian, segmentasi yang dihasilkan dapat diuji lebih lanjut apakah memiliki hubungan dengan performa rumah sakit.

4. KESIMPULAN

Penelitian ini berhasil melakukan segmentasi rumah sakit di Indonesia menggunakan pendekatan *unsupervised machine learning* dengan K-Means berdasarkan 35 skor faktor hasil *Exploratory Factor Analysis* (EFA). Hasil evaluasi menunjukkan bahwa tiga klaster merupakan solusi yang paling sesuai dengan mempertimbangkan *Silhouette Score*, ukuran klaster, dan interpretabilitas hasil. Model final menghasilkan tiga kelompok rumah sakit dengan karakteristik yang berbeda, yaitu kelompok dengan profil kapasitas dasar, kelompok yang relatif menonjol pada fasilitas penunjang, serta kelompok dengan profil layanan dan operasional yang relatif lebih tinggi. Hasil pengujian stabilitas dengan 10 *random seed* juga menunjukkan nilai ARI sebesar 0,9928–1,0000, sehingga struktur klaster yang dihasilkan relatif konsisten terhadap variasi inisialisasi. Hasil penelitian menunjukkan bahwa karakteristik rumah sakit tidak hanya dapat dibedakan berdasarkan kelas atau ukuran rumah sakit, tetapi juga

berdasarkan kombinasi multidimensi antara SDM, layanan, kapasitas tempat tidur, fasilitas penunjang, dan penunjang operasional. Penggunaan EFA sebelum K-Means memungkinkan sejumlah indikator rumah sakit direpresentasikan dalam domain faktor yang lebih ringkas dan kemudian digunakan untuk membentuk segmentasi berbasis kemiripan profil. Dengan demikian, pendekatan ini dapat menjadi alternatif berbasis data yang melengkapi klasifikasi administratif rumah sakit, sekaligus memberikan gambaran awal mengenai keragaman karakteristik rumah sakit di Indonesia. Namun, segmentasi yang dihasilkan belum merepresentasikan mutu atau kinerja pelayanan karena belum mencakup indikator seperti utilisasi dan kualitas layanan, sehingga pengembangan selanjutnya perlu mempertimbangkan indikator tersebut serta membandingkan metode clustering lainnya.

REFERENCES

- [1] E. Rachmawati et al., "Analysis Of The Market Structure of Hospital Industry in Indonesia", *Indonesian Journal of Health Administration*, vol. 12, no. 1, 2024.
- [2] D. E. et al., "Measuring the inequalities in healthcare resource in facility and workforce: A longitudinal study in China," *National Library of Medicine*, 2023.
- [3] K. K. Republik Indonesia, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 3 Tahun 2020," April 2020. [Online]. Available: <https://peraturan.bpk.go.id/Details/152506/permenkes-no-3-tahun-2020>. [Accessed 2026].
- [4] H. A. R., "Perkembangan Rumah Sakit di Indonesia di Era Disruptif," *Jurnal Keperawatan Merdeka*, vol. 2, no. 1, pp. 1-6, 2022.
- [5] M. A., S. V. and I. M., "Machine Learning in Biomedical Informatics: Optimizing Resource Allocation and Energy Efficiency in Public Hospitals," *IEEE Access*, 2025.
- [6] A. G. Jondya, R. A. M. and S. T., "Enhancing Public Health Records: Development of Death Certificate Management System," in *4th International Conference on Electrical Engineering and Informatics (ICon EEI)*, Pekanbaru, 2024.
- [7] Direktorat Jenderal Kesehatan Lanjutan, "Dashboard RS Online," December 2025. [Online]. Available: https://sirs.kemkes.go.id/fo/home/dashboard_rs?id=0.
- [8] D. J. K. Lanjutan, "Retrieved from Sistem Informasi Rumah Sakit," Dashboard RS Online, [Online]. Available: https://sirs.kemkes.go.id/fo/home/dashboard_rs?id=0. [Accessed April 2026].
- [9] R. P., R. S. S. and B. T., "Various dimension reduction techniques for high dimensional data analysis: a review," *Artificial Intelligence Review* 3473–3515, p. 3473–3515, 2021.
- [10] S. L., Y. İ. and M. A., "Exploratory factor analysis (EFA) in quantitative researches and practical considerations," *Gümüşhane Üniversitesi Sağlık Bilimleri Dergisi*, vol. 13, no. 2, pp. 947-965, 2022.
- [11] A. G. Jondya, R. F. Yani, A. Shiddiq, L. Oktaviana and A. N. Pratama, "Pengembangan Representasi Data Rumah Sakit di Indonesia Menggunakan Exploratory Factor Analysis," *Jurnal Sistem Informasi (JUSIN)*, vol. 7, no. 1, 2026.
- [12] L. Chen et al., "Evaluation of standard use of intravenous infusion in 25 general hospitals in Shanxi province based on factor analysis and cluster analysis," *Frontiers in Pharmacology*, vol. 16, 2025.
- [13] H. Hu et al., "An Effective and Adaptable K-means Algorithm for Big Data Cluster Analysis," *Pattern Recognition*, vol. 139, 2023.
- [14] T. K., V. S. and R. V., "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, 2022.
- [15] L. X., G. R., J. O. and C. J., "Evaluating medical service performance of hospitals in Sichuan Province, China: exploratory factor analysis and hierarchical clustering analysis based on diagnosis-related groups," *Frontiers in Public Health*, vol. 13, 2025.
- [16] S. M. et al., "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, 2023.
- [17] E. U. Oti et al., "Comprehensive Review of K-Means Clustering Algorithms," *International Journal of Advances in Scientific Research and Engineering (ijasre)*, vol. 7, no. 8, 2021.
- [18] K. E. Setiawan et al., "Clustering models for hospitals in Jakarta using fuzzy c-means and k-means," in *7th International Conference on Computer Science and Computational Intelligence*, 2022.
- [19] P. S. et al., "Evaluation of disease burden by (separate) patient-reported, clinical, and laboratory factor scores in patients with established rheumatoid arthritis: A factor analysis replication," *Journal of Rheumatology*, vol. 49, no. 4, 2022.
- [20] A. M. H. et al., "An Exploratory Clustering Analysis of Digital Readiness and Maturity for Electronic Medical Record Implementation in Indonesia," *Asia Pacific Journal of Health Management*, 2026.
- [21] S. N. T. M. and G. A. A. S., "Discovering the optimal number of crime cluster using Elbow, Silhouette, Gap Statistics, and NbClust methods," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 13, no. 1, pp. 1-10, 2022.
- [22] S. C. et al., "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP Journal on Wireless Communications and Networking*, no. 31, 2021.