



Comparative Analysis of DBSCAN and K-Means Algorithms in Inpatient Patient Segmentation Using the RFMT Model Approach

Analisis Komparasi Algoritma DBSCAN dan K-Means dalam Pemetaan Segmentasi Pasien Rawat Inap Menggunakan Model RFMT

Mohamad Shodikin^{1*}, Nungky Taniasari²

¹Program Studi Rekayasa Perangkat Lunak, Universitas Anwar Medika, Indonesia

²Program Studi Administrasi Kesehatan, Universitas Anwar Medika, Indonesia

E-Mail: ¹m.shodikin@gmail.com, ²taniasarinungky@gmail.com

Received Jul 31th 2025; Revised Sep 12th 2025; Accepted Sep 20th 2025; Available Online Oct 30th 2025

Corresponding Author: Mohamad Shodikin

Copyright © 2025 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Hospitals play a strategic role in providing quality healthcare services, particularly inpatient care. Accurate patient segmentation models are needed to support strategic decision-making and enhance patient loyalty. This study proposes a comparative analysis of two clustering algorithms, DBSCAN and K-Means, for inpatient segmentation using the Recency, Frequency, Monetary, and Interpurchase Time (RFMT) model. Unlike previous studies that focused on retail and marketing, this study applies RFMT specifically to hospital inpatient data. The dataset, obtained from Hospital X in Sidoarjo for the period (January–October 2022), was preprocessed and converted into RFMT format. The DBSCAN and K-Means algorithms were evaluated using internal cluster validity indices: the Silhouette Index (SI) and the Calinski-Harabasz Index (CHI). Experimental results show that DBSCAN achieves its best SI of 0.384 ($Eps=0.9$, $MinPts=24$) but produces many noisy points, while K-Means performs better with $SI=0.399$ and $CHI=14,625,319$ at $k=7$ clusters. These findings highlight that K-Means produces more stable and valid clusters in this context, while DBSCAN struggles due to the dense distribution of the dataset. This study contributes by demonstrating the application of RFMT-based clustering to hospital data and comparing the strengths and limitations of two widely used algorithms.

Keyword: Clustering, DBSCAN, K-Means, Patient Segmentation, RFMT

Abstrak

Rumah sakit memainkan peran strategis dalam memberikan layanan kesehatan berkualitas, khususnya layanan rawat inap. Model segmentasi pasien yang akurat diperlukan untuk mendukung pengambilan keputusan strategis dan meningkatkan loyalitas pasien. Studi ini mengusulkan analisis komparatif dua algoritma *clustering*, DBSCAN dan K-Means, untuk segmentasi rawat inap menggunakan model *Recency, Frequency, Monetary, and Interpurchase Time* (RFMT). Berbeda dengan studi sebelumnya yang berfokus pada ritel dan pemasaran, penelitian ini menerapkan RFMT secara spesifik pada data rawat inap rumah sakit. *Dataset* yang diperoleh dari Rumah Sakit X di Sidoarjo periode (Januari–Oktober 2022), telah diproses terlebih dahulu dan diubah ke dalam format RFMT. Algoritma DBSCAN dan K-Means dievaluasi menggunakan indeks validitas kluster internal: *Silhouette Index* (SI) dan *Calinski-Harabasz Index* (CHI). Hasil eksperimen menunjukkan bahwa DBSCAN mencapai SI terbaiknya sebesar 0,384 ($Eps=0,9$, $MinPts=24$) tetapi menghasilkan banyak titik *noise*, sementara K-Means berkinerja lebih baik dengan $SI=0,399$ dan $CHI=14.625,319$ pada $k=7$ klaster. Temuan ini menyoroti bahwa K-Means menghasilkan klaster yang lebih stabil dan valid dalam konteks ini, sementara DBSCAN mengalami kesulitan karena distribusi kepadatan *dataset*. Studi ini berkontribusi dengan menunjukkan penerapan *clustering* berbasis RFMT pada data rumah sakit dan membandingkan kekuatan dan keterbatasan dua algoritma yang banyak digunakan.

Kata Kunci: Clustering, DBSCAN, K-Means, Segmentasi Pasien, RFMT

1. PENDAHULUAN

Rumah sakit sebagai entitas fasilitas pelayanan kesehatan didukung oleh sumber daya kesehatan profesional, dimana memiliki fasilitas pelayanan medis permanen, serta menyelenggarakan pelayanan kesehatan secara menyeluruh, meliputi diagnosis, pengobatan, dan perawatan berkesinambungan bagi pasien

[1]. Rumah sakit sebagai penyedia layanan kesehatan memiliki tantangan dalam memahami karakteristik pasien rawat inap untuk meningkatkan kepuasan dan mempertahankan loyalitas pasien. Segmentasi pasien yang akurat dapat membantu manajemen dalam merancang strategi pelayanan dan pemasaran yang lebih efektif. Segmentasi pelanggan dapat dilakukan melalui pendekatan berbasis data demografis, geografis, perilaku, maupun psikografis [6]. Namun, pendekatan konvensional berbasis demografi seringkali kurang mampu menggambarkan pola perilaku pasien secara komprehensif.

Perkembangan *machine learning* memberikan solusi melalui teknik *clustering* yang dapat mengelompokkan pasien berdasarkan kesamaan pola kunjungan dan penggunaan layanan. Tujuan *clustering* yaitu pengelompokan data menjadi sejumlah kluster sehingga karakter objek di satu kluster mempunyai tingkat kesamaan tinggi, sementara perbedaan antar kluster tetap terjaga [11]. Menurut Suyanto (2020), algoritma DBSCAN memiliki keunggulan dalam membentuk kluster dengan bentuk yang tidak beraturan (*non-spherical*) serta mampu menangani keberadaan *noise* atau data yang menyimpang secara lebih fleksibel. DBSCAN juga efektif dalam mengidentifikasi dan memisahkan data *noise*, serta dapat mendeteksi kluster dalam berbagai bentuk geometri yang tidak terdefinisi secara ketat [12]. K-Means adalah algoritma *clustering* dikenal efisien untuk menangani *dataset* berukuran cukup besar. K-Means bekerja dengan cara mengelompokkan data berdasarkan tingkat kemiripan dan perbedaan antar objek, serta memiliki kemampuan komputasi yang cepat dalam memproses data, termasuk dalam menghadapi *outlier* [14]. Algoritma *clustering* seperti DBSCAN dan K-Means banyak digunakan karena mampu memetakan kelompok dengan karakteristik berbeda tanpa memerlukan label awal.

Urgensi penelitian ini terletak pada kebutuhan rumah sakit dalam memahami perilaku pasien rawat inap secara lebih akurat melalui pendekatan berbasis data (*data-driven*). Penelitian ini bertujuan untuk melakukan analisis komparatif terhadap kinerja algoritma DBSCAN dan K-Means dalam segmentasi pasien dengan menggunakan model *Recency*, *Frequency*, *Monetary*, dan *Interpurchase Time* (RFMT). RFMT menggambarkan perilaku konsumen dengan mengukur empat dimensi utama, yaitu: waktu sejak transaksi terakhir dilakukan (*Recency*), frekuensi atau jumlah transaksi (*Frequency*), total nilai pembelian dalam periode tertentu (*Monetary*) dan interval waktu antar dua transaksi berturut-turut (*Interpurchase Time*). Temuan yang dihasilkan dari penelitian ini diharapkan dapat menjadi dasar bagi manajemen rumah sakit dalam merumuskan strategi retensi pasien yang lebih efektif dan berorientasi pada peningkatan kualitas layanan.

Penelitian terdahulu secara luas telah mengimplementasikan model RFM pada sektor ritel dan perbankan, namun penerapannya dalam konteks layanan rumah sakit masih relatif terbatas. Kontribusi utama penelitian ini adalah pengembangan model RFM menjadi RFMT dengan penambahan variabel *Interpurchase Time* untuk memberikan gambaran perilaku pasien yang lebih komprehensif. Selain itu, penelitian ini menyajikan analisis komparatif antara dua algoritma *clustering* yang banyak digunakan, yakni DBSCAN dan K-Means. Perbedaan mendasar penelitian ini dibandingkan studi sebelumnya terletak pada penerapan model RFMT terhadap data pasien rawat inap serta evaluasi kinerja kedua algoritma tersebut secara bersamaan.

2. PENELITIAN TERDAHULU

Model *Recency*, *Frequency*, *Monetary* (RFM) yang diperkenalkan oleh Hughes pada tahun 1994 telah menjadi salah satu pendekatan standar dalam menganalisis perilaku pelanggan. Selama beberapa dekade terakhir, model ini secara luas digunakan dalam segmentasi berbasis perilaku untuk memahami nilai dan karakteristik pelanggan [3], [4], [13]. RFM menggambarkan perilaku konsumen dengan mengukur tiga dimensi utama, yaitu: waktu sejak transaksi terakhir dilakukan (*Recency*), frekuensi atau jumlah transaksi (*Frequency*), dan total nilai pembelian dalam periode tertentu (*Monetary*). Data RFM diperoleh dari riwayat transaksi pelanggan dan kemudian diklasifikasikan ke dalam kelompok diskrit melalui proses yang dikenal sebagai penilaian RFM (*RFM scoring*). Dalam praktiknya, terdapat dua pendekatan utama dalam menetapkan skor RFM: metode kuintil pelanggan dan metode kuintil perilaku [15]. Pada metode kuintil pelanggan, data pelanggan diurutkan berdasarkan nilai Frekuensi dan Moneter secara menurun, atau berdasarkan *Recency* secara menaik, lalu dibagi menjadi lima kelompok dengan jumlah anggota yang relatif sama. Skor diberikan dari 5 untuk kelompok dengan nilai tertinggi, hingga 1 untuk kelompok dengan nilai terendah, sebagaimana ditunjukkan dalam studi [4]. Sementara itu, metode kuintil perilaku mengelompokkan pelanggan berdasarkan intensitas perilaku yang ditunjukkan, yang dapat menghasilkan distribusi jumlah pelanggan yang tidak merata di tiap kuintil. Contohnya, [3] memberikan skor “1” untuk pelanggan yang hanya melakukan satu kali pembelian dan “5” bagi yang melakukan lebih dari enam kali. Studi lain oleh [13] menetapkan skor “1” untuk pengguna kartu kredit dengan nilai transaksi di bawah \$50 juta dan “5” untuk pengguna dengan transaksi di atas \$300 juta. Secara umum, pelanggan yang memiliki waktu pembelian paling baru, tingkat transaksi paling sering, dan pengeluaran paling tinggi akan memperoleh skor tertinggi pada masing-masing dimensi RFM.

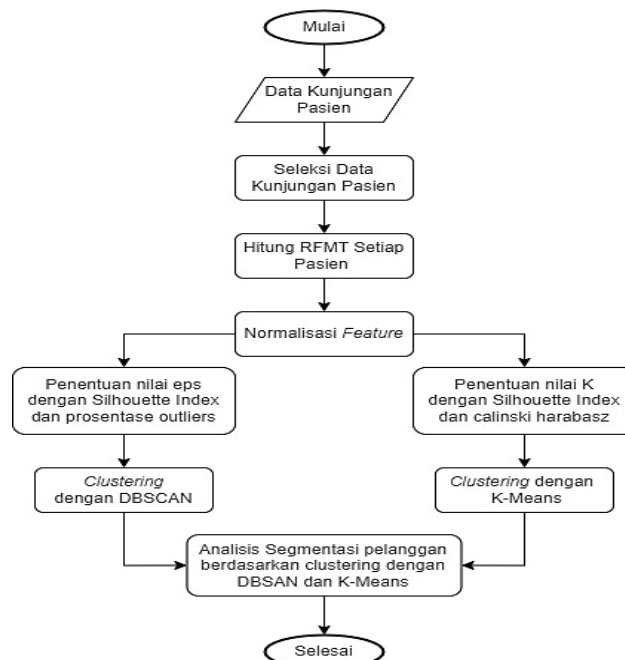
Setelah proses pemetaan data dilakukan, algoritma *clustering* dapat secara otomatis mengelompokkan data RFM ke dalam kluster yang sesuai. *Clustering* merupakan teknik dalam *machine learning* yang bertujuan membuat kelompok titik-titik data berdasarkan kemiripan fitur yaitu data yang memiliki karakter

sama diklasifikasikan menjadi kelompok yang sama, sementara data yang mempunyai karakteristik beda diletakkan pada kelompok berbeda [17]. Beberapa algoritma *clustering* yang umum digunakan meliputi K-Means [7], *Expectation-Maximization* (EM) [10], serta metode *hierarchical clustering* [11]. Dalam penerapannya pada model RFM, studi oleh [31] mengimplementasikan algoritma K-Means untuk mengidentifikasi pelanggan potensial di sektor ritel dan berhasil menentukan tiga kluster optimal. Sementara itu, [23] menggunakan kombinasi K-Means dan EM *clustering* untuk mengelompokkan pelanggan dari perusahaan ritel kecil dan menengah di Kuwait ke dalam empat kluster. Seiring perkembangan metode analisis perilaku pelanggan, beberapa penelitian telah mengembangkan variasi dari model RFM klasik. Misalnya, [24] memperluas model tersebut dengan menambahkan variabel “waktu sejak pembelian pertama” (T) dan “probabilitas pelanggan berhenti” (C). Selain itu, [25] dan [26] mengembangkan model LRFM dengan memperkenalkan dimensi *length* (L), yaitu selang waktu antara pembelian pertama dan terakhir yang dilakukan oleh pelanggan. Dalam konteks yang sama, [23] mengusulkan metrik baru berupa tingkat perubahan pembelian (C) untuk mencerminkan besarnya dan arah perubahan perilaku pelanggan dari waktu ke waktu. Dimensi *Interpurchase Time* (T), yang merepresentasikan interval waktu antar dua transaksi berturut-turut oleh pelanggan di toko atau platform yang sama, telah lama digunakan dalam analisis perilaku konsumen sejak era 1960-an [27], [28]. Penggunaan dimensi ini antara lain terlihat dalam penelitian [28] yang menghubungkan keteraturan pembelian dengan kecenderungan belanja daring, serta dalam studi [29] yang menganalisis pengaruh program loyalitas terhadap perilaku pembelian konsumen. Selanjutnya, [30] mengembangkan model *multi-category interpurchase time* untuk memprediksi kemungkinan pembelian pelanggan secara individual. Model ini juga dimanfaatkan dalam meningkatkan sistem rekomendasi produk dengan mempertimbangkan dinamika waktu antar pembelian lintas kategori produk.

Tujuan utama dari proses *clustering* dipandang dari sisi optimasi yaitu untuk memaksimalkan homogenitas di dalam masing-masing kluster (*intra-cluster similarity*) serta meningkatkan heterogenitas antar kluster (*inter-cluster dissimilarity*) [32]. Studi yang dilakukan oleh [33] membandingkan performa algoritma K-Means dan K-Medoids dalam pengkategorian data sebuah transaksi bongkar muat di Riau. Penelitian tersebut menunjukkan tidak terdapat perbedaan signifikan antara kedua algoritma dalam hal hasil pengelompokan data. Sementara itu, penelitian yang dilakukan oleh [34] menerapkan algoritma DBSCAN untuk melakukan klusterisasi terhadap kutipan teks Twitter terkait Pilkada Kota Pekanbaru. Dengan konfigurasi parameter *epsilon* (*Eps*) sebesar 0,1 dan *minimum points* (*MinPts*) sebanyak 10, diperoleh nilai *Silhouette Index* (SI) terbaik sebesar 0,413 dan berhasil membentuk 31 kluster. Penelitian lain oleh [34] juga mengeksplorasi kombinasi antara algoritma DBSCAN dan K-Medoids dalam mengelompokkan opini massa terkait bencana alam yang terjadi di Indonesia. Dalam studi tersebut, algoritma DBSCAN menunjukkan performa yang unggul dengan nilai validitas kluster tertinggi, ditunjukkan oleh nilai SI yaitu 0,9140 serta rerata durasi eksekusi selama 83,40 detik.

3. METODOLOGI PENELITIAN

Penelitian ini menggunakan metodologi segmentasi pelanggan seperti terlihat pada Gambar 1.



Gambar 1. Bagan Proses Segmentasi Pelanggan

3.1 Data Kunjungan Pasien

Dataset dalam penelitian ini diperoleh dari pengambilan data pasien rawat inap RS X di Sidoarjo sejak bulan Januari-Oktober 2022 dengan total 25.944 kunjungan dari 20.452 pasien. Atribut yang digunakan ditunjukkan pada Tabel 1.

Tabel 1. Atribut *Dataset* Pasien Rawat Inap

No	Atribut	Deskripsi
1	ID Pasien	Identitas unik pasien
2	ID Kunjungan	Identitas unik kunjungan rawat inap
3	Tanggal Kunjungan	Tanggal pasien masuk rawat inap
4	Kelas Kepesertaan	Jenis kepesertaan (BPJS, umum, asuransi)
5	Jenis Kelamin	Laki-laki / Perempuan
6	Tanggal Lahir	Data untuk menghitung usia
7	Usia	Umur pasien pada saat kunjungan
8	Pendidikan	Latar belakang pendidikan
9	Tagihan Layanan	Total biaya per kunjungan

Atribut ini dipilih karena mewakili karakteristik demografi, sosio-ekonomi, dan utilisasi layanan. Validasi pemilihan atribut dilakukan melalui diskusi dengan kepala unit layanan rawat inap dan staf SIM RS.

3.2 Perhitungan RFMT Setiap Pasien

Dalam penerapan segmentasi pelanggan menggunakan pendekatan RFMT, data yang tersedia dalam bentuk spreadsheet perlu dikonversi terlebih dahulu ke dalam format yang sesuai dengan dimensi RFMT. Langkah awal dilakukan dengan menggunakan atribut ID pasien sebagai dasar untuk menghitung setiap komponen dalam model RFMT. Nilai *Monetary* (M) diperoleh dengan menjumlahkan seluruh tagihan layanan dari pasien yang sama, sedangkan *Frequency* (F) dihitung berdasarkan jumlah kunjungan rawat inap yang dilakukan oleh masing-masing pasien. Nilai *Recency* (R) ditentukan dengan mengukur selang waktu antara kunjungan rawat inap terakhir hingga tanggal pengambilan data. Sementara itu, dimensi keempat yaitu *Interpurchase Time* (T), mencerminkan rata-rata interval waktu antar kunjungan pasien secara berurutan. Jika T_1 menyatakan tanggal kunjungan pertama dan T_n merupakan tanggal kunjungan terakhir, maka siklus belanja keseluruhan (L) dapat diperkirakan dalam satuan bulan sebagai selisih waktu antara T_n dan T_1 . Nilai T selanjutnya dapat dihitung melalui rumus sesuai Persamaan (1).

$$T = L / (F - 1) = (T_n - T_1) / (F - 1) \quad (1)$$

Untuk menghitung T, hanya pasien yang melakukan setidaknya dua pembelian (yaitu, $F \geq 2$) pada periode tertentu yang dipertimbangkan. Setelah *preprocessing* data, terdapat total kunjungan 25,944 dari 20,452 pasien pada *dataset*. Peneliti melaksanakan *Focus Group Discussion* (FGD) dengan manajemen rumah sakit untuk merumuskan pengkategorian nilai RFMT pada *dataset-dataset* pasien rawat inap. Hasil FGD seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Aturan Penilaian Diskritisasi RFMT

Kategori	Kuintil	R(Bulan)	F	M (Juta)	T (Hari)
Sangat Rendah	1	≥ 9	1	< 1	0
Rendah	2	7 – 8	2	1 – 3	≥ 90
Sedang	3	5 – 6	3	4 – 6	60 – 89
Tinggi	4	3 – 4	4	7 – 8	30 – 59
Sangat Tinggi	5	≤ 2	> 4	≥ 9	1 – 29

Tabel 2 menggambarkan pembagian pasien ke dalam lima kategori berdasarkan nilai *Recency*, *Frequency*, *Monetary*, dan *Interpurchase Time*. Semakin tinggi kategori, semakin tinggi pula loyalitas dan nilai ekonomis pasien bagi rumah sakit. Kategori pasien sangat rendah jika kunjungan terakhir pasien lebih dari 9 bulan, frekuensi kunjungan hanya 1 kali dalam satu periode, dimana nilai tagihan pasien kurang dari Rp 1.000.000,00 dan jarak antar kunjungan bernilai 0 atau lebih besar dari 9 bulan. Kategori pasien rendah jika kunjungan terakhir pasien antara 7 – 9 bulan, frekuensi kunjungan 2 kali dalam satu periode, dimana nilai tagihan pasien antara Rp 1.000.000,00 – Rp 3.000.000,00 dan jarak antar kunjungan bernilai lebih dari 90 hari. Kategori pasien sedang jika kunjungan terakhir pasien antara 5 – 6 bulan, frekuensi kunjungan 3 kali dalam satu periode, dimana nilai tagihan pasien antara Rp 4.000.000,00 – Rp 6.000.000,00 dan jarak antar kunjungan bernilai 60 – 89 hari. Kategori pasien tinggi jika kunjungan terakhir pasien antara 3 – 4 bulan, frekuensi kunjungan 4 kali dalam satu periode, dimana nilai tagihan pasien antara Rp 7.000.000,00 – Rp 8.000.000,00 dan jarak antar kunjungan bernilai 30 – 59 hari. Kategori pasien sangat tinggi jika kunjungan

terakhir pasien kurang dari 2 bulan, frekuensi kunjungan lebih dari 4 kali dalam satu periode, dimana nilai tagihan pasien lebih dari Rp 9.000.000,00 dan jarak antar kunjungan bernilai kurang dari 30 hari.

Tabel 3. Contoh Nilai RFMT Pasien

Px_Id	Recency (R)	Frequency (F)	Monetary (M)	Interpurchase Time (T)	R_Rank	F_Rank	M_Rank	T_Rank
P001	1	20	57.15	10	5	5	5	5
P002	5	15	246.31	9	3	5	5	5
P003	0	14	161.17	21	5	5	5	5
P003	0	13	101.51	19	5	5	5	5
P004	0	13	63.35	21	5	5	5	5

Tabel 3 menyajikan contoh hasil perhitungan dimensi RFMT pada beberapa pasien rawat inap. Variabel *Recency* (R) menunjukkan jarak waktu sejak kunjungan terakhir pasien, *Frequency* (F) merepresentasikan jumlah total kunjungan rawat inap, sedangkan *Monetary* (M) menggambarkan total biaya layanan yang dikeluarkan pasien selama periode penelitian. Variabel *Interpurchase Time* (T) dihitung sebagai rata-rata selang waktu antar kunjungan rawat inap secara berurutan. Selain nilai asli, setiap dimensi RFMT kemudian ditransformasikan ke dalam skor kuintil (*R_Rank*, *F_Rank*, *M_Rank*, *T_Rank*) dengan rentang 1–5. Skor 1 menunjukkan kategori sangat rendah, sedangkan skor 5 menunjukkan kategori sangat tinggi. Sebagai contoh, pasien dengan *px_id* P001 memiliki nilai *Frequency* = 15 dan *Monetary* = 246,3 juta, sehingga mendapatkan skor maksimal (5) pada dimensi F dan M. Hal ini menunjukkan pasien tersebut memiliki tingkat kunjungan tinggi dan biaya rawat inap yang besar.

Penyajian skor diskritisasi ini memudahkan proses *clustering*, karena algoritma dapat mengelompokkan pasien berdasarkan tingkat kesamaan pola kunjungan dan kontribusi finansial secara lebih terukur. Dengan demikian, tabel ini memperlihatkan bagaimana data mentah dikonversi ke dalam format RFMT terstandarisasi sebelum dianalisis menggunakan algoritma DBSCAN dan K-Means.

3.3 Normalisasi Fitur

Dalam rangka menganalisis hasil klusterisasi, setiap parameter dalam model RFMT yaitu *Recency* (R), *Frequency* (F), *Monetary* (M), dan *Interpurchase Time* (T) diklasifikasikan ke dalam lima tingkatan kategori, yaitu: Sangat Rendah, Rendah, Sedang, Tinggi, dan Sangat Tinggi. Data yang diperoleh terlebih dahulu melalui tahap *data cleaning* dan *normalisasi* untuk memastikan kualitas dan konsistensinya sebelum proses analisis lebih lanjut. Setelah data dipersiapkan, algoritma DBSCAN dan K-Means diimplementasikan melalui serangkaian eksperimen. Pada pengujian dengan DBSCAN, digunakan kombinasi parameter *Eps* dan *MinPts* yang bervariasi. Pada algoritma K-Means dilakukan beberapa percobaan dengan nilai *K* yang berbeda untuk menentukan seberapa besar jumlah klaster yang optimal. Validitas hasil klusterisasi masing-masing eksperimen dievaluasi menggunakan SI sebagai indikator utama untuk menilai kualitas pembentukan klaster.

3.4 Clustering Dengan Algoritma DBSCAN

Algoritma DBSCAN bekerja dengan cara mengidentifikasi titik-titik inti (*core points*) yang berada pada area dengan kepadatan data tinggi, lalu memperluas klaster dari titik-titik tersebut secara bertahap. Pembentukan klaster dalam DBSCAN bergantung pada dua parameter utama, yaitu *MinPts* dan *epsilon* (ϵ). *MinPts* menentukan jumlah minimum titik yang harus berada dalam radius ϵ agar suatu titik bisa dianggap titik inti (*core point*). Parameter ini juga berperan dalam menentukan tingkat toleransi terhadap keberadaan *noise* dalam data [16].

Langkah-langkah pelaksanaan algoritma DBSCAN dapat dijabarkan sebagai berikut:

1. Menentukan nilai parameter *MinPts* dan *Eps* yang akan digunakan dalam proses klusterisasi.
2. Memilih secara acak satu titik data sebagai titik awal (p).
3. Menghitung jarak antara titik p dan titik-titik lainnya yang mungkin terjangkau dalam radius *Eps*. Penghitungan dilakukan menggunakan rumus jarak *Euclidean* sebagaimana dinyatakan pada Persamaan (2).

$$dij = \sqrt{\sum (x_{ia} - x_{ja})^2} \quad (2)$$

Dimana x_{ia} adalah variabel ke- a dari obyek i ($i=1, \dots, n$; $a=1, \dots, p$) dan dij adalah nilai euclidean distance.

4. Sebuah klaster terbentuk apabila jumlah titik dalam radius *Eps* dari titik p melebihi nilai *MinPts*, dimana titik p diklasifikasikan menjadi titik inti (*core point*).
5. Tahap ke-3 dan ke-4 diulangi untuk seluruh titik dalam *dataset*. Apabila titik p menjadi titik batas (*border point*) dan ditemukan tidak ada titik lain *density reachable* dari p, maka disimpulkan proses dapat dilanjutkan dengan titik data lainnya.

3.5 Clustering Dengan Algoritma K-Means

K-Means dapat dijadikan pilihan teknik *clustering* untuk pengelompokan data. Cara kerja algoritma ini mengelompokkan data yang memiliki atribut atau fitur serupa pada satu kluster sama. Jika ada data berbeda akan dimasukkan dalam kluster lainnya [20]. K-Means merupakan algoritma *clustering* paling sederhana sekaligus paling banyak digunakan dalam praktik mengelompokkan data [20]. Jarak ke-I (x_i) pusat cluster ke-k (c_k), yang disebut (d_{ik}), bisa dihitung menggunakan rumus jarak *Euclidean* sebagaimana ditunjukkan dalam Persamaan (3).

$$d_{ik} = \sqrt{\sum_{j=1}^m (c_{ij} - x_{ik})^2} \quad (3)$$

3.6 Silhouette Index (SI)

Metode SI dapat digunakan sebagai salah satu pendekatan validasi kluster yang termasuk dalam kategori *internal validation* [21]. Validasi berdasarkan kriteria internal dapat mengevaluasi kualitas klusterisasi dengan mengkomparasi jarak rerata suatu objek terhadap objek lain pada kluster sama, serta terhadap objek pada kluster berbeda [22]. Nilai SI dihitung menggunakan rumus Persamaan (4), dimana persamaan ini merepresentasikan keseimbangan antara koherensi internal kluster dan pemisahan antar kluster.

$$SI = \frac{1}{n} \sum_{i=1}^n (b(i) - a(i) \max \{a(i), b(i)\}) \quad (4)$$

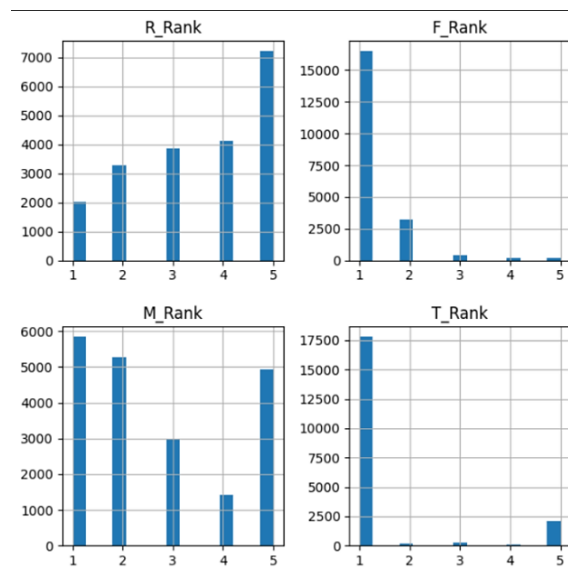
Dimana $a(i)$ merupakan jarak rerata sampel i ke sampel lainnya dalam satu *cluster*, dan $b(i)$ mewakili jarak minimal sampel dari sampel i ke cluster lain.

3.7 Calinski Harabasz Index (CHI)

Indeks validitas *Calinski-Harabasz* (CH) digunakan untuk mengevaluasi kualitas klusterisasi dengan membandingkan tingkat pemisahan antar kluster (*between-cluster dispersion* atau SSB) terhadap tingkat kepadatan dalam kluster (*within-cluster dispersion* atau SSW). Nilai perbandingan tersebut kemudian dikalibrasi dengan faktor normalisasi yang diperoleh dari selisih antara jumlah total data dengan jumlah kluster yang ada, dibagi jumlah kluster dikurangi satu. Semakin tinggi nilai CHI yang dihasilkan, semakin baik struktur kluster yang terbentuk, sehingga nilai maksimum CHI mengindikasikan jumlah kluster yang optimal [18].

4. HASIL DAN DISKUSI

Pemahaman terhadap karakteristik masing-masing kluster yang terbentuk melalui proses klusterisasi memungkinkan pihak manajemen rumah sakit untuk merancang strategi pemasaran yang lebih terarah sesuai dengan segmen pasien tertentu. Pendekatan ini berkontribusi pada peningkatan efektivitas upaya retensi pasien. Visualisasi distribusi nilai RFMT dalam skala diskrit disajikan pada Gambar 2.



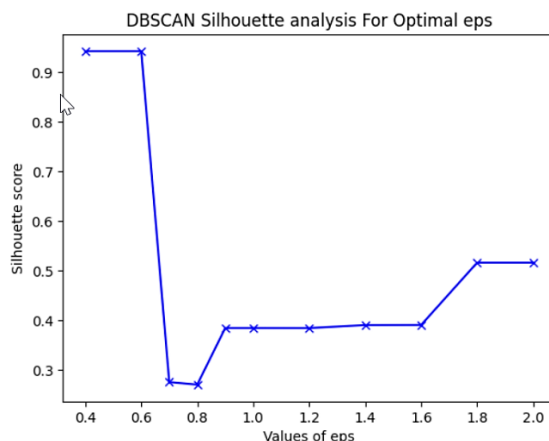
Gambar 2. Distribusi RFMT

Gambar 2 menunjukkan distribusi pasien rawat inap dimana skor R tertinggi sejumlah 7.197 pasien yang masuk dalam kategori 5 (sangat tinggi). Skor F tertinggi sejumlah 16.476 pasien yang masuk dalam

kategori 1 (sangat rendah). Skor M tertinggi sejumlah 5.847 pasien yang masuk dalam kategori 1 (sangat rendah). Skor T tertinggi sejumlah 17.795 pasien yang masuk dalam kategori 1 (sangat rendah).

4.1 Klasterisasi menggunakan Algoritma DBSCAN

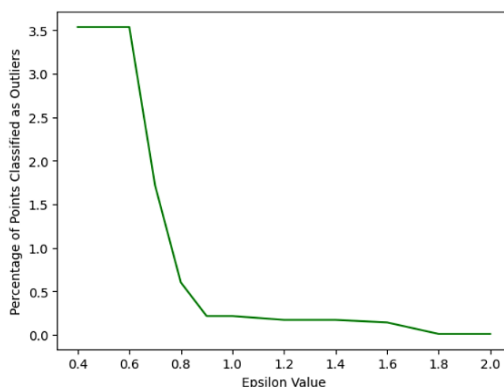
Pemanfaatan algoritma DBSCAN dalam klasterisasi dilakukan melalui sebelas percobaan, masing-masing menggunakan kombinasi nilai *Eps* dan *MinPts* yang berbeda. Rentang nilai *Eps* 0,4 – 2. Nilai *MinPts* ditetapkan sebesar 24. Selanjutnya, dilakukan evaluasi terhadap hasil klasterisasi melalui pengukuran validitas klaster menggunakan SI. Hasil perhitungan nilai SI untuk masing-masing percobaan DBSCAN disajikan pada Gambar 3.



Gambar 3. Diagram *Silhouette Index* untuk Pencarian Nilai Optimal *Eps*

Gambar 3 menunjukkan variasi nilai SI terhadap perubahan parameter *Eps* pada algoritma DBSCAN dengan *MinPts*=24. Terlihat bahwa pada nilai *Eps* kecil (sekitar 0,4–0,6), DBSCAN menghasilkan skor SI yang cukup tinggi (>0,9). Namun demikian, skor ini tidak mencerminkan klasterisasi yang valid karena sebagian besar data justru dianggap sebagai *noise* sehingga jumlah klaster yang terbentuk sangat sedikit. Dengan kata lain, meskipun SI tinggi, kualitas klasterisasi tidak representatif karena hanya segelintir data yang dikelompokkan.

Ketika nilai *Eps* diperbesar (0,8–1,4), skor SI mengalami penurunan signifikan dan cenderung stabil pada kisaran rendah (0,3–0,4). Kondisi ini menunjukkan bahwa dengan rentang *Eps* tersebut, DBSCAN berhasil mengelompokkan lebih banyak data ke dalam klaster, tetapi pemisahan antar klaster tidak terlalu baik. Nilai SI terbaik yang lebih realistis diperoleh pada *Eps*=0,9 dengan *MinPts*=24, yaitu sebesar 0,384. Konfigurasi ini menghasilkan klasterisasi dengan jumlah *noise* yang relatif lebih sedikit dibandingkan nilai *Eps* terlalu kecil.



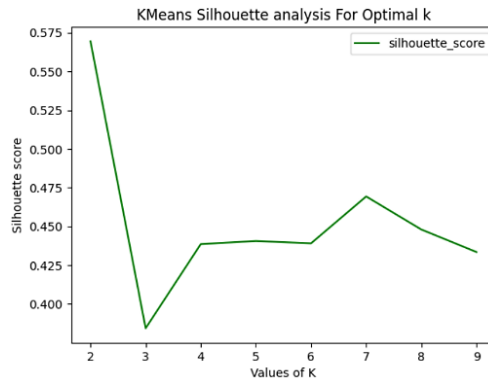
Gambar 4. Diagram Nilai *Error* Disetiap *Eps*

Gambar 4 menunjukkan hubungan antara parameter *Eps* dengan persentase data yang diklasifikasikan sebagai *outlier* oleh algoritma DBSCAN. Terlihat bahwa pada nilai *Eps* yang sangat kecil (sekitar 0,4–0,6), persentase data yang dianggap *outlier* relatif tinggi, mencapai lebih dari 3,5%. Hal ini wajar karena dengan radius pencarian yang sempit, banyak titik data tidak memiliki cukup tetangga (*MinPts*) sehingga tidak dapat membentuk klaster dan akhirnya dikategorikan sebagai *noise*.

Seiring bertambahnya nilai *Eps*, persentase *outlier* menurun drastis. Pada nilai *Eps*=0,9 jumlah *outlier* mendekati nol, sehingga hampir seluruh data berhasil masuk ke dalam kluster. Kondisi ini menunjukkan bahwa *Eps*=0,9 dengan *MinPts*=24 merupakan salah satu konfigurasi yang cukup optimal bagi DBSCAN, karena mampu mengurangi jumlah *noise* secara signifikan.

4.2 Klasterisasi menggunakan Algoritma K-Means

Pemanfaatan algoritma K-Means dalam klasterisasi dengan variasi jumlah kluster sebanyak 2-10 merujuk pada pendekatan yang digunakan dalam penelitian sebelumnya [33]. Selanjutnya dilakukan evaluasi validitas kluster menggunakan metrik SI. Nilai SI yang dihasilkan dari klasterisasi terlihat pada Gambar 5.

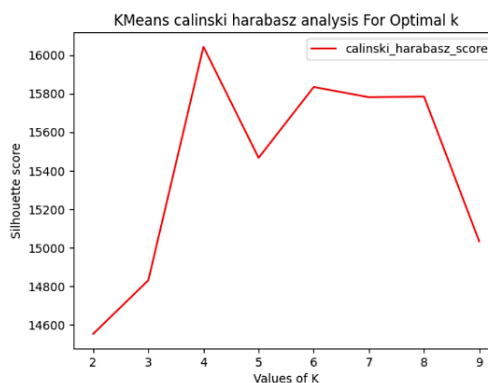


Gambar 5. *Silhouette Index* untuk Pencarian Nilai K Paling Optimal

Gambar 5 menunjukkan nilai SI pada berbagai jumlah kluster ($k=2$ sampai $k=10$) yang dihasilkan oleh algoritma K-Means. Terlihat bahwa nilai SI berfluktuasi pada setiap variasi jumlah kluster, namun puncak tertinggi diperoleh pada $k=7$ dengan SI sebesar 0,399. Nilai ini mengindikasikan bahwa pemisahan antar kluster cukup jelas dan homogenitas dalam kluster relatif baik pada jumlah kluster tersebut.

Pada jumlah kluster yang lebih kecil ($k=2-3$), nilai SI cenderung rendah ($<0,2$) yang menandakan struktur kluster kurang representatif karena data dipaksa masuk ke dalam kelompok yang terlalu sedikit. Sementara pada jumlah kluster yang lebih besar ($k=8-10$), nilai SI kembali menurun, menunjukkan bahwa penambahan kluster justru tidak meningkatkan kualitas pemisahan, melainkan membuat klasterisasi menjadi kurang stabil.

Dengan demikian, $k=7$ dipandang sebagai jumlah kluster optimal untuk *dataset* pasien rawat inap berbasis RFMT. Hal ini menunjukkan bahwa pasien dapat dipetakan ke dalam tujuh segmen berbeda dengan karakteristik perilaku yang unik, misalnya kelompok pasien dengan kunjungan sering dan biaya tinggi, hingga kelompok pasien dengan kunjungan jarang dan biaya rendah. Evaluasi validitas kluster juga dilakukan menggunakan CHI. Nilai CHI yang diperoleh dari klasterisasi data memanfaatkan algoritma K-Means ditampilkan pada diagram pada Gambar 6..



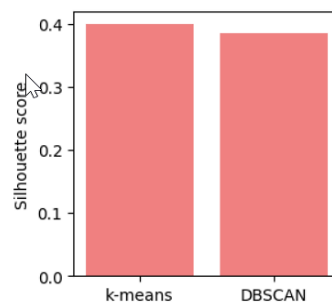
Gambar 6. Diagram *Calinski Harabasz* untuk Pencarian Nilai K Paling Optimal

Gambar 6 menunjukkan nilai CHI untuk berbagai jumlah kluster ($k=2$ sampai $k=10$) menggunakan algoritma K-Means. Indeks CH mengukur rasio dispersi antar-kluster terhadap dispersi intra-kluster, sehingga semakin tinggi nilai CHI menunjukkan struktur kluster yang lebih baik, dengan kluster yang kompak sekaligus terpisah jelas. Hasil analisis menunjukkan bahwa nilai CHI tertinggi diperoleh pada $k=4$, yang menandakan bahwa secara matematis jumlah kluster 4 menghasilkan pemisahan antar kelompok yang paling

jelas. Namun, nilai CHI pada $k=6-7$ juga relatif tinggi dan stabil, sehingga jumlah kluster di rentang tersebut masih dapat dianggap valid. Perbedaan ini perlu dicermati dengan membandingkan hasil CHI dan SI (Gambar 6 sebelumnya). Pada SI, jumlah kluster optimal ditemukan pada $k=7$ dengan $SI=0,399$, sedangkan pada indeks CH jumlah optimal berada pada $k=4$ dengan nilai tertinggi 16.000+. Hal ini mengindikasikan adanya trade-off antara homogenitas dalam kluster dan pemisahan antar kluster.

4.3 Perbandingan Algoritma DBSCAN dan K-Means

Komparasi antara dua algoritma yaitu DBSCAN dan K-Means dilakukan melalui identifikasi jumlah kluster optimal berdasarkan nilai tertinggi SI. Visualisasi komparasi SI paling tinggi yang dicapai oleh masing-masing algoritma disajikan pada Gambar 7.



Gambar 7. Diagram Perbandingan *Silhouette Index* Algoritma DBSCAN dan K-Means

Evaluasi kualitas klusterisasi dua algoritma yaitu DBSCAN dan K-Means dinilai berdasarkan metrik SI. Hasilnya menunjukkan algoritma DBSCAN memiliki nilai SI tertinggi sebesar 0,384 diperoleh dengan parameter *Eps* sebesar 0,9 dan *MinPts* sebanyak 24. Sementara itu, pada algoritma K-Means, nilai SI tertinggi tercapai pada saat jumlah kluster (k) adalah 7, dengan nilai sebesar 0,399. Berdasarkan hasil penelitian ini menunjukkan bahwa algoritma K-Means kualitas klusterisasinya lebih baik daripada DBSCAN, dimana konfigurasi optimal diperoleh pada $k=7$.

4.4 Diskusi Penelitian

Hasil penelitian menunjukkan bahwa algoritma K-Means lebih unggul dibandingkan DBSCAN dalam melakukan segmentasi pasien rawat inap berbasis model RFMT. K-Means mampu menghasilkan nilai SI tertinggi sebesar 0,399 pada jumlah kluster optimal $k=7$, sedangkan DBSCAN hanya mencapai SI 0,384 pada parameter *Eps*=0,9 dengan *MinPts*=24. Perbedaan ini menegaskan bahwa K-Means lebih sesuai untuk data dengan distribusi variabel yang relatif seragam, seperti pola kunjungan dan biaya rawat inap pasien.

Jika dibandingkan dengan penelitian terdahulu yang menerapkan model RFM pada sektor ritel dan perbankan, penelitian ini memperluas konteks penerapan dengan menambahkan variabel *Interpurchase Time* (T), sehingga menghasilkan model RFMT yang lebih komprehensif untuk menganalisis perilaku pasien. Hasil penelitian ini sejalan dengan studi yang menyatakan bahwa penambahan variabel temporal dapat meningkatkan akurasi segmentasi pelanggan.

Secara praktis, temuan ini memiliki implikasi penting bagi manajemen rumah sakit. Segmentasi pasien ke dalam tujuh kelompok memungkinkan penyusunan strategi pemasaran yang lebih tepat sasaran, misalnya:

1. Pasien bernilai tinggi (frekuensi tinggi, biaya besar, interval kunjungan pendek) dapat menjadi target program retensi khusus seperti layanan prioritas dan membership.
2. Pasien bernilai rendah dapat menjadi target program promosi atau edukasi untuk meningkatkan frekuensi kunjungan.

Keterbatasan penelitian ini terletak pada penggunaan *dataset* tunggal dari satu rumah sakit, sehingga generalisasi hasil masih terbatas. Selain itu, evaluasi kluster hanya menggunakan dua metrik validitas internal (SI dan *Calinski-Harabasz*), belum melibatkan validasi eksternal dengan data lain.

Penelitian lanjutan dapat mengembangkan model dengan menambahkan variabel klinis (misalnya diagnosis, lama rawat inap, atau komorbiditas) agar segmentasi lebih kaya makna. Selain itu, penggunaan metode *clustering* lain seperti *Gaussian Mixture Model* atau *Hierarchical Clustering* dapat diuji untuk melihat performa dibandingkan K-Means.

5. KESIMPULAN

Penelitian ini membuktikan bahwa penerapan model RFMT dalam segmentasi pasien rawat inap memberikan informasi yang lebih komprehensif dibandingkan model RFM tradisional. Temuan baru yang

diperoleh adalah bahwa algoritma K-Means dengan konfigurasi $k=7$ mampu menghasilkan segmentasi pasien yang lebih valid dibandingkan DBSCAN, dengan nilai SI tertinggi sebesar 0,399. Implikasi utama dari penelitian ini adalah manajemen rumah sakit dapat memanfaatkan hasil segmentasi untuk merumuskan strategi retensi pasien yang lebih sistematis dan tepat sasaran. Pasien dengan nilai tinggi dapat difokuskan pada program loyalitas, sementara pasien bernilai rendah dapat menjadi sasaran strategi promosi.

Penelitian mendatang disarankan untuk menguji model RFMT pada konteks rumah sakit yang berbeda, menambahkan variabel klinis sebagai faktor segmentasi, serta membandingkan dengan algoritma *clustering* lain. Dengan demikian, penelitian ini membuka peluang pengembangan strategi pemasaran berbasis data yang lebih kuat dalam bidang layanan kesehatan.

REFERENSI

- [1] Aditama, tjandra yoga. (2004). Manajemen administrasi rumah sakit. Edisi 2. Jakarta: ui press
- [2] C. Benabdellah, a. Benghabrit, and i. Bouhaddou, "a survey of clustering algorithms for an industrial context," *procedia comput. Sci.*, vol. 148, pp. 291–302, 2019, doi: 10.1016/j.procs.2019.01.022.
- [3] Chang, hc, tsai, hp, 2011. Analisis rfm grup sebagai kerangka kerja baru untuk menemukan perilaku konsumsi pelanggan yang lebih baik. *Sistem pakar. Aplikasi* 38 (12), 14499–14513. <https://doi.org/10.1016/j.eswa.2011.05.034>.
- [4] Christy, aj, umamakeswari, a., priyatharsini, l., neyaa, a., 2018. Peringkat rfm—an pendekatan yang efektif untuk segmentasi pelanggan. *Universitas j. King saud, komp. & info. Sains*. <https://doi.org/10.1016/j.jksuci.2018.09.004>
- [5] Elrod, c., stanley, s., cudney, e., fisher, c., 2015. Studi empiris memanfaatkan qfd untuk mengembangkan strategi pemasaran internasional. <https://doi.org/10.3390/su70810756>.
- [6] Haider, t. And shakib, s. (2017) a study on the influences of advertisement on consumer buying behavior. *Business studies journal*, 9.
- [7] Garbade, mj, 2018. Memahami k-means clustering pada machine learning. <https://towardsdatascience.com/understanding-k-means-clustering-in-machinelearning-6a6e67336aa1>.
- [8] Guo, j., gao, z., liu, n., wu, y., 2018. Merekomendasikan produk dengan pertimbangan waktu dan harga antar kategori multi kategori. *Generasi masa depan. Komputer. Sistem*. 78, 451–461. <https://doi.org/10.1016/j.future.2017.02.031>.
- [9] Hood, n., clarke, g., clarke, m., 2016. Segmentasi pasar toserba uk yang berkembang untuk perencanaan lokasi ritel. *Int. Pdt. Distribusi ritel. Konsumsi. Res.* 26 (2), 113–136. <https://doi.org/10.1080/09593969.2015.1086403>.
- [10] Hui, j., 2019. Machine learning-expectation-maximization algorithm (em). https://medium.com/@jonathan_hui/machine-learning-expectation-maximization-algorithmem-2e954cb76959.
- [11] Malav, k. Kadam, and p. Kamat, "prediction of heart disease using kmeans and artificial neural network as hybrid approach to improve accuracy," *int. J. Eng. Technol.*, vol. 9, no. 4, pp. 3081–3085, 2017.
- [12] W. Jing, c. Zhao, and c. Jiang, "an improvement method of dbscan algorithm on cloud computing" *procedia comput. Sci.*, vol. 147, pp. 596–604, 2019, doi: 10.1016/j.procs.2019.01.208.
- [13] Maryani, i., riana, d., 2017. Clustering dan profiling pelanggan menggunakan rfm untuk rekomendasi manajemen hubungan pelanggan. Dalam: 2017 5th international conference on cyber and it service management (citism), hlm. 1–6. <https://doi.org/10.1109/citism.2017.8089258>.
- [14] M. Metisen and h. L. Sari, "analisis clustering menggunakan metode k-means dalam pengelompokan penjualan produk pada swalayan fadhila" *j. Media infotama*, vol. 11, no. 2, pp. 110–118. Issn: 1858 – 2680, 2015.
- [15] Miglautsch, jr, 2000. Pemikiran tentang penilaian rfm. *J. Tanda basis data. Pelanggan manajer strategi*. 8 (1), 67–72. <https://doi.org/10.1057/palgrave.jdm.3240019>.
- [16] M. Pietrzykowski, "comparison of mini-models based on various clustering algorithms," *procedia comput. Sci.*, vol. 176, pp. 3563–3570, 2020, doi: 10.1016/j.procs.2020.09.030.
- [17] Seif, g., 2018. Algoritma pengelompokan data yang perlu diketahui para ilmuwan. <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>.
- [18] Baarsch, j., & celebi, m. E., 2012. Investigation of internal validity measures for k-means clustering. *International multiconference of engineers and computer scientists 1* (hal. 14-16). La: louisiana board of regents.
- [19] Rahmawati and m. Faisal, "analisis cluster untuk pengelompokan desa berdasarkan indikator penyakit diare," *saintifik*, vol. 5, no. 1, pp. 75–80, 2019.
- [20] I. D. Iskandar, m. W. Pertiwi, m. Kusmira, and i. Mirulloh, "komparasi algoritma clustering data media online," *j. Ikraith inform.*, vol. 2, no. 4, pp. 1–8, 2018.
- [21] M. A. Nahdliyah, t. Widiarihi, and a.prahtutama, "metode k-medoids clustering dengan validasi silhouette index dan cindex," vol. 8, no. 2, pp. 161–170, 2019.

- [22] X. Wang and y. Xu, “an improved index for clustering validation based on silhouette index and calinski-harabasz index,” *iop conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, 2019, doi: 10.1088/1757-899x/569/5/052024
- [23] Yoseph, f., & heikkilä, m. (2018). Segmenting retail customers with an enhanced rfm and a hybrid regression/clustering method. In 2018 international conference on machine learning and data engineering (icmlde) (pp. 109–116). Ieee. <https://doi.org/10.1109/icmlde.2018.00029>
- [24] Yeh, i.c., yang, k.j., ting, t.m., 2009. Knowledge discovery on rfm model using bernoulli sequence. *Expert syst. Appl.* 36 (3), 5866–5871. <https://doi.org/10.1016/j.eswa.2008.07.018>.
- [25] Alvandi, m., fazli, s., abdoli, f.s., 2012. K-mean clustering method for analysis customer lifetime value with lrfm relationship model in banking services. *Int. Res. Judgement* 3 (11), 2294–2302.
- [26] Wei, j.t., lin, s.y., weng, c.c., wu, h.h., 2012. A case study of applying lrfm model in market segmentation of a children’s dental clinic. *Expert syst. Appl.* 39 (5), 5529–5533. <https://doi.org/10.1016/j.eswa.2011.11.066>.
- [27] Morrison, d.g., 1966. Interpurchase time and brand loyalty. *J. Mar. Res.* 3 (3), 289–291. <https://doi.org/10.1177/002224376600300309>.
- [28] Vakratsas, d., bass, f.m., 2002. The relationship between purchase regularity and propensity to accelerate. *J. Retailing* 78 (2), 119–129. [https://doi.org/10.1016/s0022-4359\(02\)00068-4](https://doi.org/10.1016/s0022-4359(02)00068-4).
- [29] Meyer-waarden, l., 2008. The influence of loyalty programme membership on customer purchase behaviour. *Eur.j.mark.* 42 (1/2), 87–114. <https://doi.org/10.1108/03090560810840925>.
- [30] Guo, r.s., 2009. A multi-category inter-purchase time model based on hierarchical bayesian theory. *Expert syst. Appl.* 36 (3), 6301–6308. <https://doi.org/10.1016/j.eswa.2008.08.059>.
- [31] Anitha, p., patil, m.m., 2019. Rfm model for customer purchase behavior using k-means algorithm. *J. King saud univ., comp. & info. Sci.* <https://doi.org/10.1016/j.jksuci.2019.12.011>.
- [32] A. C. Benabdellah, a. Benghabrit, and i. Bouhaddou, “a survey of clustering algorithms for an industrial context,” *procedia comput. Sci.*, vol. 148, pp. 291–302, 2019, doi:10.1016/j.procs.2019.01.022.
- [33] I. Kamila, u. Khairunnisa, and m.mustakim, “perbandingan algoritma kmeans dan k-medoids untuk pengelompokan data transaksi bongkar muat di provinsi riau,” *j. Ilm. Rekayasa dan manaj. Sist. Inf.*, vol. 5, no. 1, p. 119, 2019, doi: 10.24014/rmsi.v5i1.7381.
- [34] Mustakim et al., “DBSCAN algorithm: twitter text clustering of trend topic pilkada pekanbaru,” *j. Phys. Conf. Ser.*, vol. 1363, no. 1, 2019, doi: 10.1088/1742-6596/1363/1/012001