Perbandingan Metode Clustering K-Means, GMM, dan DBSCAN Berbasis Fitur RFM

Putu Nadya Putri Astina[a1], I Wayan Supriana[a2], I Made Satria Bimantara[a3]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana, Bali
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]nadyaputriast@gmail.com
[2]wayan.supriana@unud.ac.id
[3]satriabimantara@unud.ac.id

Abstract

In the digital banking era, understanding customer behavior has become essential for delivering relevant services and maintaining competitiveness. This study aims to develop and evaluate customer segmentation models by leveraging an extended RFM (Recency, Frequency, Monetary) model, incorporating both behavioral and demographic attributes. Preprocessing, feature engineering, handling outliers, and standardization were done on the data using a dataset of 100,000 bank transaction records from Kaggle. DBSCAN, Gaussian Mixture Model (GMM), and K-Means were the three clustering techniques that were employed and contrasted. The clustering performance was evaluated using the Silhouette Score, Calinski-Harabasz Index (CHI), and Davies-Bouldin Index (DBI). The output of DBSCAN was too noisy to be useful in the business world, despite having the highest scores on Silhouette: 0,667 and lowest score on DBI: 0,396. K-Means offered the most interpretable segmentation with five ideal clusters (Silhouette: 0,308; DBI: 0,957; CHI: 6191), identifying customer groups ranging from highly active to potentially inactive. The findings highlight the synergy between transactional features and clustering algorithms in generating actionable insights for banking strategy.

Keywords: *Customer Segmentation, DBSCAN, GMM, K-Means, RFM Model*

1. Pendahuluan

Di era digital dan revolusi 4.0 ini, telah mendorong adanya transformasi besar dalam sektor perbankan [1], terutama dengan meningkatnya adopsi teknologi dalam memenuhi kebutuhan nasabah yang kian kompleks. Transformasi ini tidak hanya mencakup layanan digital seperti *mobile banking* [2], tetapi juga perubahan strategi bisnis dari produk-sentris ke pelanggan-sentris [3]. Dengan persaingan yang semakin ketat dan ekspektasi nasabah yang terus meningkat, bank perlu memahami perilaku nasabah secara mendalam untuk dapat memberikan layanan yang relevan [4]. Oleh karena itu, dibutuhkan segmentasi nasabah untuk dapat mengidentifikasi kelompok nasabah yang memiliki karakteristik dan kebiasaan yang serupa [5].

Tiga variabel utama, tanggal terakhir transaksi (*recency*), banyaknya melakukan transaksi (*frequency*), dan jumlah nominal transaksi (*monetary*), digunakan dalam penelitian ini. Data historis transaksi pelanggan digunakan untuk penelitian ini [6]. Segmentasi pelanggan berdasarkan perilaku transaksional sebelumnya dapat dilakukan dengan sukses dengan metode RFM [7].

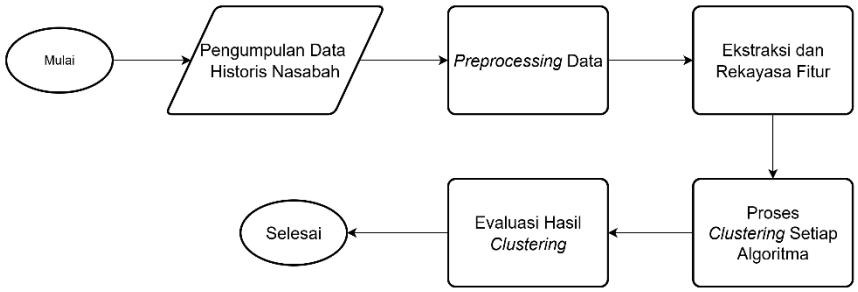
Sejumlah penelitian sebelumnya telah mengusulkan penerapan berbagai algoritma *clustering*, khususnya model *unsupervised learning*. Salah satu model yang sering digunakan adalah model RFM (*Recency*, *Frequency*, dan *Monetary*) karena kesederhanaannya dalam merepresentasikan perilaku transaksi nasabah. Penelitian [8] dan [9] telah memanfaatkan model RFM dalam kombinasi dengan algoritma K-Means dan DBSCAN untuk dapat membentuk *cluster* pelanggan yang dapat digunakan untuk strategi pemasaran yang baik.

Selanjutnya, pengembangan [10] dari menambahkan dimensi saldo (*balance*) sebagai indikator penting dalam melakukan nilai pelanggan, sehingga menjadi model RFM+B.

Selain menggabungkan metode yang ada, terdapat pendekatan komparatif [10] yang memaparkan bahwa pemilihan metode itu penting agar sesuai dengan struktur dan kompleksitas data. Selain itu, terdapat pendekatan *hybrid* [11] yang digunakan untuk meningkatkan akurasi segmentasi dan mengatasi keterbatasan masing-masing algoritma secara individual.

Namun, sebagian besar penelitian masih memiliki keterbatasan dalam mengintegrasikan atribut yang ada, serta minimnya evaluasi komparatif antar algoritma menggunakan metrik yang konsisten. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan segmentasi nasabah menggunakan model RFM yang diperluas dengan fitur perilaku dan demografis. Selain itu, penelitian ini juga akan menerapkan dan membandingkan performa algoritma K-Means, GMM, dan DBSCAN dengan mengevaluasi hasil segmentasinya menggunakan metrik seperti *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*.

2. Metode Penelitian



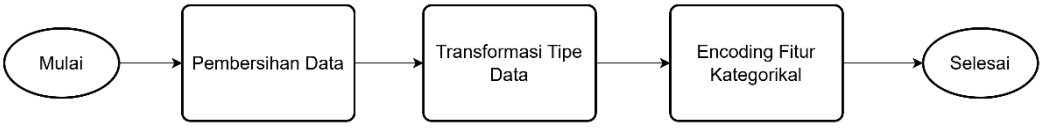
Gambar 1. Alur penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan tujuan penelitian untuk mengelompokkan nasabah bank berdasarkan pola transaksi dan atribut demografis. Secara garis besar, tahapan penelitian ini terdiri dari lima langkah utama seperti yang terdapat pada Gambar 1. Adapun penjelasan untuk setiap tahapannya adalah sebagai berikut:

2.1. Pengumpulan Data

Data yang digunakan merupakan data historis nasabah yang mencakup tanggal transaksi, nominal transaksi, kategori *merchant*, jenis transaksi, dan data demografis seperti usia, jenis kelamin, pendapatan, dan lokasi tempat tinggal. Data diakuisisi pada situs Kaggle. Seluruh data akan disimpan dalam satu tabel dengan setiap barisnya merepresentasikan satu nasabah.

2.2. *Preprocessing* Data

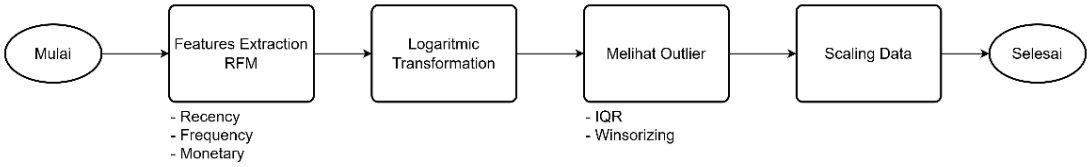


Gambar 2. Tahapan *Preprocessing* Data

Gambar 2 ini diperuntukkan bagi proses penyiapan dan pembersihan data agar analisis dapat berjalan dengan efektif. Adapun tahapannya di antaranya:

a. Aktivitas pembersihan data berperan dalam mengeliminasi duplikasi, data yang tidak akurat, maupun nilai yang tidak da.
b. Transformasi tipe data pada tanggal transaksi ke dalam format *datetime*.
c. *Encoding* fitur kategorikal, seperti gender dan lokasi akan dikodekan ke dalam bentuk numerik menggunakan *Label encoding* agar dapat diproses oleh algoritma *clustering*.

2.3. Ekstraksi dan Rekayasa Fitur



Gambar 3. Ekstraksi dan Rekayasa Fitur

Berdasarkan Gambar 3, usai tahapan *preprocessing*, tahap selanjutnya membangun fitur segmentasi melalui model RFM (*Recency*, *Frequency*, *Monetary*) [6]. *Recency* mengukur interval hari dari transaksi terakhir ke tanggal referensi, *Frequency* mencatat frekuensi transaksi tiap nasabah, dan *Monetary* menghitung akumulasi nilai transaksi [7]. Data dengan distribusi miring atau tidak normal diatasi melalui transformasi logaritma dengan *base* 10 atau *natural log*. *Outlier* ditangani menggunakan metode IQR dengan batas $Q_1 - 1.5 \times IQR$ sampai $Q_3 + 1.5 \times IQR$, lalu diterapkan *winsorizing* untuk mengganti nilai ekstrem. Tahap akhir adalah *scaling data* untuk menyamakan skala antar variabel agar tidak terjadi bias dalam analisis.

2.4. Proses *Clustering*

Setelah fitur numerik dibuat dan distandarisasi, tiga algoritma *clustering* digunakan untuk segmentasi pelanggan. K-Means, DBSCAN, dan *Gaussian Mixture Model* (GMM) dirancang untuk mempelajari berbagai struktur *cluster* data pelanggan dengan asumsi distribusi yang beragam.

a. K-Means

K-Means merupakan algoritma *clustering* yang umum digunakan karena kesederhanaan dan efisiensinya [9]. Data dibagi menjadi banyak kelompok dengan cara ini karena jarak kuadrat antara setiap titik data, *centroid*, dan pusat kelompok diminimalkan. Dimulai dengan inisialisasi centroid acak. Kemudian, menggunakan perhitungan jarak geometris, setiap poin data ditempatkan ke centroid terdekat. Persamaan (1) menunjukkan bahwa inertia adalah fungsi objektif yang diminimumkan.

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left|\left| x_j - \mu_i \right|\right|^2 \quad (1)$$

Di mana μ_i merepresentasikan *centroid* atau titik tengah dari *cluster* C_i. Setiap model *cluster* dibandingkan dengan beberapa metrik evaluasi, dan nilai k dihitung dari 2 hingga 10. Didasarkan pada Silhouette Score tertinggi, Davies-Bouldin Index terendah, dan Calinski-Harabasz Index tertinggi, nilai k terbaik dipilih.

b. Gaussian Mixture Model (GMM)

GMM memodelkan distribusi data sebagai campuran dari beberapa distribusi Gaussian multivariat dan memperhitungkan probabilitas keanggotaan tiap titik data terhadap *cluster* [12]. Berbeda dari algoritma K-Means, GMM tidak melakukan *hard assignment*, dalam artian memberikan probabilitas suatu titik data untuk semua *cluster*. Model ini cocok digunakan untuk data pelanggan seperti RFM karena sering kali distribusinya tidak homogen dan terdapat tumpang tindih antar segmen. Estimasi parameter dilakukan

dengan algoritma *Expetation-Maximization* (EM) dan *log-likelihood* total, seperti yang terlihat pada persamaan (2).

$$\log L = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \pi_i \times N(x_j | \mu_i, \sum_i) \right) \quad (2)$$

Parameter distribusi Gaussian $\mu_i, \sum_i$ dan bobot komponen π_i . Algoritma GMM mengevaluasi nilai k dalam rentang 2 sampai 10. Pemilihan komponen optimal dapat dilakukan menggunakan kriteria informasi Bayesian (BIC) dan Akaike (AIC) yang berfungsi mengukur tingkat kompleksitas model guna menghindari *overfitting*. Nilai k dengan BIC dan AIC terendah dipilih sebagai jumlah komponen optimal.BIC dapat dirumuskan seperti pada persamaan (3).

$$BIC = -2 \times \log L + k \times \log n \quad (3)$$

c. Density-Based Spatial Clustering of Application with Noise (DBSCAN)

DBSCAN merupakan sebuah algoritma yang mengelompokkan titik data berdasarkan kerapatan lokal, tanpa perlu menentukan banyaknya *cluster* di awal [11]. Suatu titik dianggap sebagai inti *cluster* jika memiliki jumlah tetangga minimal atau *min sample* dalam radius ε tertentu. Aturan *cluster* DBSCAN dapat dilihat pada persamaan (3).

$$|N_\varepsilon(p)| \geq \min\_samples \quad (3)$$

Dengan $N_\varepsilon(p)$ adalah himpunan titik dalam jarak ε dari titik p, *grid search* dilakukan terhadap nilai ε (misal dari 0.3 hingga 1.5) dan $\min\_samples$ (misal dari 3 hingga 18). *Silhouette Score*, DBI, dan CHI dihitung untuk setiap kombinasi, tetapi hanya untuk hasil *cluster* yang memiliki lebih dari satu *cluster* dan tidak terpengaruh oleh *noise*.

2.5. Evaluasi Hasil *Clustering*

Tiga metrik digunakan untuk menilai kualitas hasil segmentasi masing-masing algoritma. Metrik yang digunakan adalah Silhouette *score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Setiap metrik evaluasi diuraikan sebagai berikut:

a. Silhouette Score

Metrik ini digunakan untuk mengukur seberapa dekat suatu titik dengan *cluster*-nya dibandingkan dengan *cluster* terdekat lainnya [6], [9]. Nilainya berada antara -1 hingga 1. Adapun persamaan untuk metrik ini dapat dilihat pada persamaan (4).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

Dengan $a(i)$ adalah rata-rata jarak antara titik i dan semua titik dalam *cluster*-nya dan $b(i)$ adalah rata-rata jarak antara titik i dan semua titik pada *cluster* terdekat.

b. Davies-Bouldin Index

Metrik ini menghitung rata-rata dari rasio jarak dalam *cluster* terhadap jarak antar pusat *cluster* yang berbeda [6], [11]. Formulanya dapat dilihat pada persamaan (5).

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max j \neq i \left(\frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)} \right) \quad (5)$$

Di mana σ_i adalah rata-rata jarak titik-titik dalam *cluster* i ke pusat *cluster* μ_i, dan $d(\mu_i, \mu_j)$ merupakan jarak antar pusat *cluster*. Semakin rendah nilai DBI menandakan *cluster* yang lebih terpisah dan padat.

c. Calinski-Harabasz Index (CHI)

Calinski-Harabasz Index, disebut pula *Variance Ratio Criterion*, digunakan untuk mengukur rasio antara variansi antar *cluster* dan dalam *cluster* [13]. Perhitungannya dapat dilihat pada (6).

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{n-k}{k-1} \quad (6)$$

Dengan $TR(B_k)$ adalah jumlah kuadrat antar *cluster*, $Tr(W_k)$ adalah jumlah kuadrat dalam *cluster*, n adalah banyak data, dan k adalah banyak *cluster*. Nilai CHI yang semakin tinggi mengindikasikan segmentasi yang lebih baik.

3. Hasil dan Pembahasan

Penelitian ini bertujuan untuk membandingkan hasil segmentasi nasabah bank mengguna tiga metode, yaitu K-Means, GMM, dan DBSCAN yang diimplementasikan dalam bahasa pemrograman Python. Penelitian ini juga bertujuan untuk mengevaluasi hasil segmentasi menggunakan tiga metrik evaluasi, di antaranya Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index (CHI).

3.1. *Preprocessing* Data

Data yang digunakan pada penelitian ini merupakan data sekunder yang berasal dari situs Kaggle. *Dataset* yang digunakan merupakan bank_transaction_dataset oleh Toby Bushenyula. Tahapan *preprocessing* dimulai dengan menghapus data yang duplikat, melakukan transformasi tanggal transaksi ke dalam format *datetime*, dan melakukan *label encoding* pada kolom Customer Gender dan Customer Location. Pada *label encoding*, pengkodean dilakukan menggunakan urutan alfabet, sebagai contoh pada Customer Gender “female” dikodekan dengan “0”, “male” dengan “1”, dan “other” dengan “2”. Banyaknya data pada tahap ini adalah 100.000 data atau *record*. Hasil *preprocessing* data untuk 5 data pertama disajikan pada tabel 1.

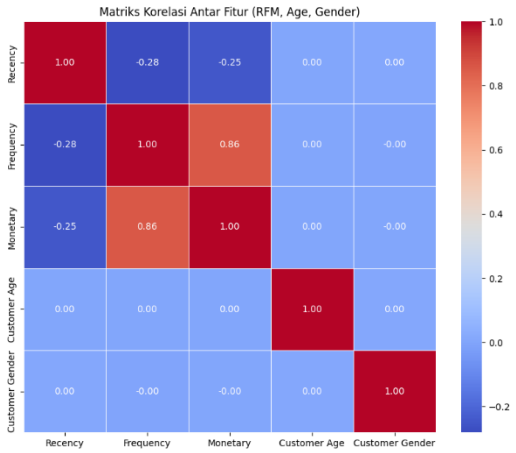
Tabel 1. Hasil *Preprocessing* Data

Trans ID	Cust ID	Date	Amount	Categ.	Type	Age	Gender	Income	Loc
TXN1000000	CUST1144	2024-02-05	1154.17	Clothing	Refund	30	0	54800.0	33703
TXN1000001	CUST2063	2024-01-07	4232.30	Restaurant	Purchase	21	1	39949.0	15054
TXN1000002	CUST2130	2024-03-10	1630.74	Gas Station	Refund	69	2	122072.0	32520
TXN1000003	CUST7597	2024-05-02	335.72	Electronics	Purchase	57	2	90891.0	25969
TXN1000004	CUST1234	2024-07-04	1410.51	Gas Station	Refund	40	1	150545.0	38009

3.2. Ekstraksi dan Rekayasa Fitur

Tabel 1 menampilkan informasi tanggal, jumlah, kategori, jenis transaksi, usia, gender, pendapatan, dan lokasi pelanggan. Namun, analisis hanya akan memanfaatkan tiga fitur utama yaitu *monetary*, *frequency*, dan *recency*. Kolom yang digunakan adalah kolom identitas pelanggan, tanggal, dan jumlah *income*. Fitur *recency* berasal dari kapan terakhir seorang *customer* ID melakukan transaksi. Fitur *frequency* diperoleh dengan menghitung kemunculan setiap ID pelanggan dalam *dataset*, yang mencerminkan total transaksi yang dilakukan masing-masing pelanggan. Analisis korelasi dilakukan terhadap lima variabel: *recency*, *frequency*, *monetary*, usia, dan gender. Berdasarkan Gambar 1, tidak ditemukan korelasi yang signifikan antara fitur usia dan gender dengan fitur RFM.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 1. Matriks Korelasi Antar Fitur

Selanjutnya, dengan hanya mengambil fitur RFM saja, didapatkan 9.000 data *customer*. Tabel 2 menggambarkan hasil dari ekstraksi fitur. Setelah mendapatkan fitur yang akan dianalisis, tahapan berikutnya adalah transformasi logaritmik mengurangi kemiringan data. Hal ini dapat membuat distribusi data lebih simetris atau normal. Seperti yang ditunjukkan pada Gambar 2, Gambar 3, dan Gambar 4, di mana hasil transformasi masih menunjukkan kemiringan negatif meskipun distribusi data sudah hampir normal, dengan nilai *skewness* pada fitur *monetary* terbesar. Selain melihat *skewness*, diobservasi pula *outlier* untuk masing-masing fitur. Untuk melihat *outlier* dapat menggunakan *boxplot*.

Berdasarkan gambar 5, fitur *Recency* tidak menunjukkan *outlier*, tetapi *frequency* dan *monetary* masing-masing memiliki *outlier* 3,58% (322 data) dan 2,12% (191 data). Untuk mengatasinya, akan dilakukan *winsorizing* dengan mengganti nilai ekstrem di bawah persentil ke-4 dan di atas persentil ke-96 menggunakan nilai persentil tersebut. Terakhir, Customer ID akan dihapus karena tidak relevan untuk segmentasi.

Setelah *outlier* tertangani, langkah selanjutnya adalah *scaling* menggunakan StandardScaler. Tahapan ini penting agar tidak ada fitur yang mendominasi, memastikan semua berada pada skala yang sama. Dengan begitu, model K-Means yang sensitif terhadap skala fitur bisa bekerja tanpa bias. Hasil *scaling* ini disajikan pada Tabel 3.

Tabel 2. Ekstraksi Fitur RFM

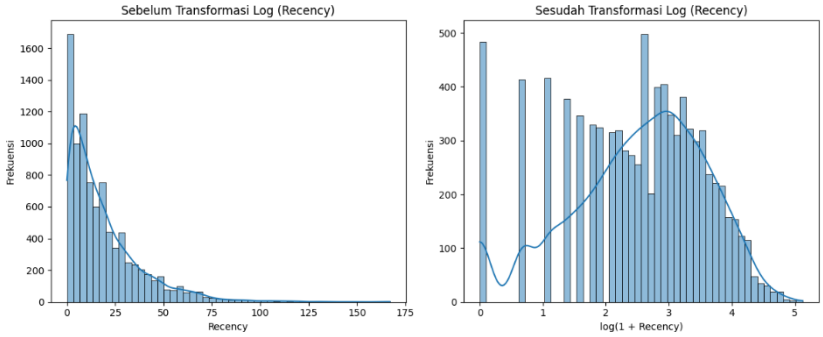
Cust ID	Recency	Frequency	Monetary
CUST1144	31	7	21352.33
CUST2063	3	11	27857.36
CUST2130	13	11	22481.19
CUST7597	20	16	36508.36
CUST1234	13	12	30892.78

Tabel 3. Hasil *Scaling* Fitur

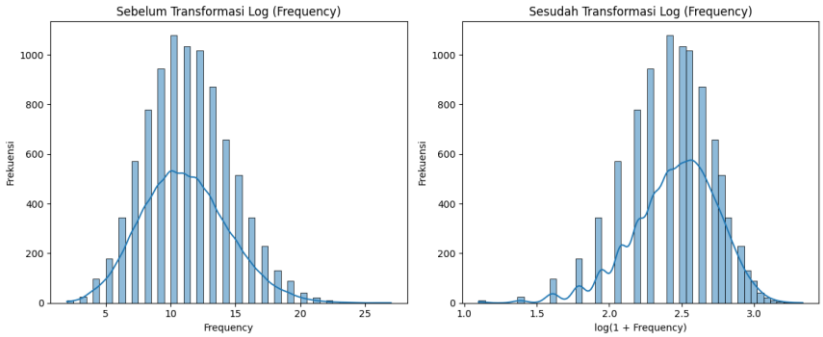
Recency	Frequency	Monetary
0.908365	-1.454919	-0.600981
-1.019012	0.097029	0.172076
0.142139	0.097029	-0.451222

p-ISSN: 2986-3929
e-ISSN: 3032-1948

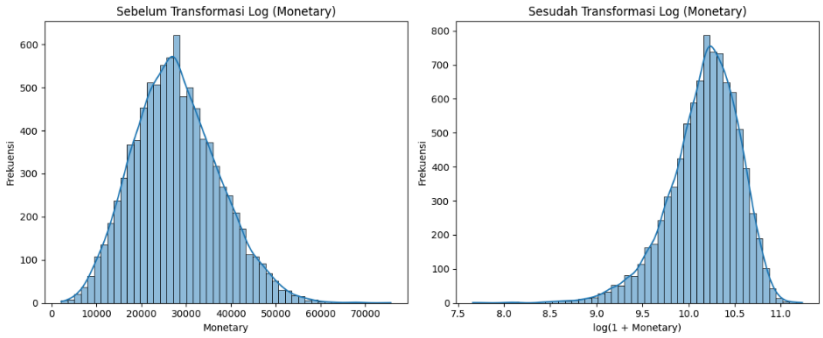
Recency	Frequency	Monetary
0.517954	1.430200	0.958244
0.142139	0.403399	0.472727



Gambar 2. Transformasi Logaritmik Fitur *Recency*

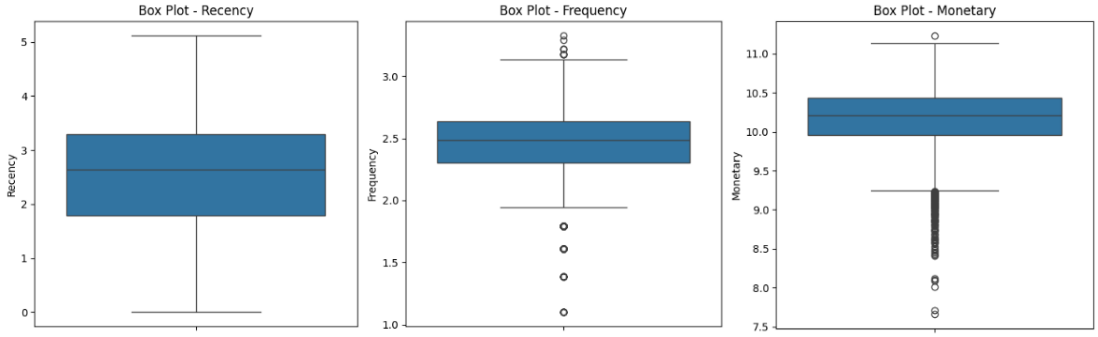


Gambar 3. Transformasi Logaritmik Fitur *Frequency*



Gambar 4. Transformasi Logaritmik Fitur *Monetary*

p-ISSN: 2986-3929
e-ISSN: 3032-1948



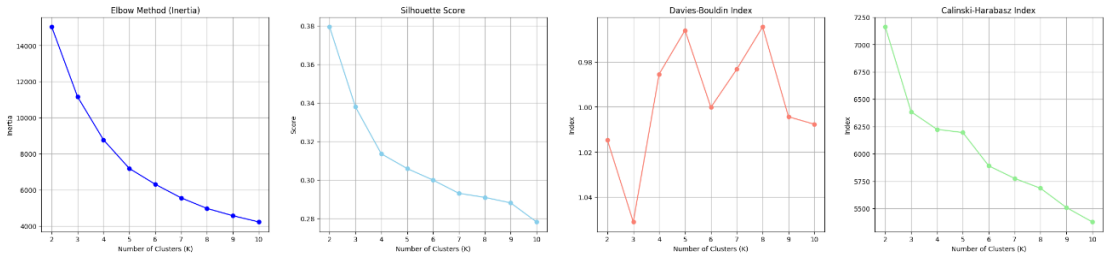
Gambar 5. Hasil *Outlier* Menggunakan *Boxplot*

3.3. *Clustering*

Tahapan ini merupakan tahapan untuk melakukan segmentasi hingga mengevaluasi hasil segmentasi. Seperti pada penjelasan bagian sebelumnya, terdapat tiga metode yang akan digunakan, yaitu K-Means, GMM, dan DBSCAN.

a. K-Means

Pada metode K-Means, untuk memilih nilai k (banyaknya *cluster*) optimal, dilakukan dengan mengevaluasi beberapa metrik, seperti inertia (*elbow method*), *silhouette score*, *Calinski Harabasz score*, dan *Davies Bouldin score*. Inertia mengukur total jarak kuadrat antar data terhadap pusat *cluster*-nya, dan digunakan dalam *elbow method* untuk melihat titik "siku" sebagai indikasi k optimal. *Silhouette Score* mengukur seberapa baik pemisahan antar *cluster* (semakin tinggi semakin baik), sedangkan *Calinski-Harabasz Index* menilai rasio antara variasi antar dan dalam *cluster* (semakin tinggi semakin baik). Sementara itu, *Davies-Bouldin Index* mengukur kemiripan antar *cluster* (semakin rendah semakin baik). Evaluasi dilakukan dengan variasi k=2 hingga k=10.



Gambar 6. Hasil Evaluasi K-Means

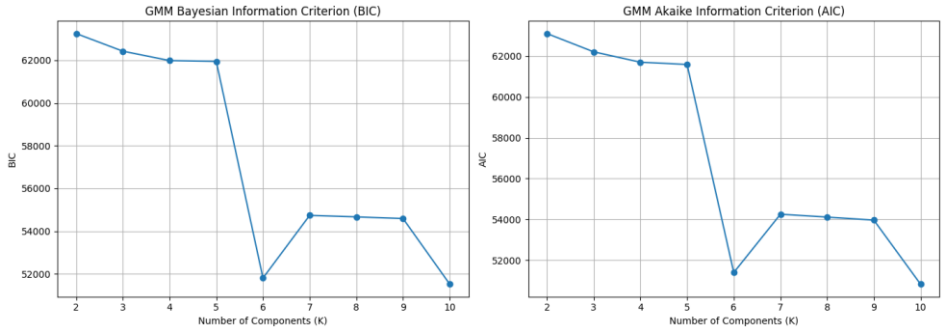
Nilai $k = 5$ ditetapkan sebagai parameter terbaik, seperti yang ditunjukkan dalam Gambar 6. Keputusan ini didasarkan pada penemuan titik "*elbow*" paling signifikan pada grafik inertia, serta pencapaian kedua terendah pada indeks *Davies-Bouldin*, dengan nilai 0.966. Walaupun *Silhouette Score* dan *Calinski-Harabasz Index* mendapatkan nilai tertinggi pada $k = 2$, pembentukan hanya dua *cluster* tidak mencukupi untuk menunjukkan keragaman karakteristik perilaku konsumen. Untuk keperluan analisis bisnis, segmentasi pelanggan yang lebih informatif dan praktis dapat dilakukan dengan penetapan $k = 5$. Selain itu, ini berhasil mengimbangi kualitas clustering dan tingkat kompleksitas model dengan sangat baik. Untuk $k = 5$, hasil algoritma K-Means didistribusikan dalam Gambar 10.

b. *Gaussian Mixture Model* (GMM)

Gaussian Mixture Model (GMM), yang didasarkan pada probabilitas, menganggap bahwa data berasal dari campuran beberapa distribusi normal Gaussian. GMM memungkinkan setiap data memiliki probabilitas keanggotaan terhadap setiap *cluster* (*soft assignment*),

p-ISSN: 2986-3929
e-ISSN: 3032-1948

berbeda dengan K-Means, yang membagi data secara keras (*hard assignment*). Untuk menentukan jumlah komponen (*cluster*) optimal pada GMM, digunakan dua metrik seleksi model yaitu *Bayesian Information Criterion* (BIC) dan *Akaike Information Criterion* (AIC). Keduanya mengevaluasi *trade-off* antara kualitas pemodelan dan kompleksitas model. Semakin baik model menjelaskan data dengan tingkat kompleksitas yang rendah. Karena BIC memberikan penalti yang lebih besar terhadap model yang terlalu kompleks, ia cenderung lebih konservatif.



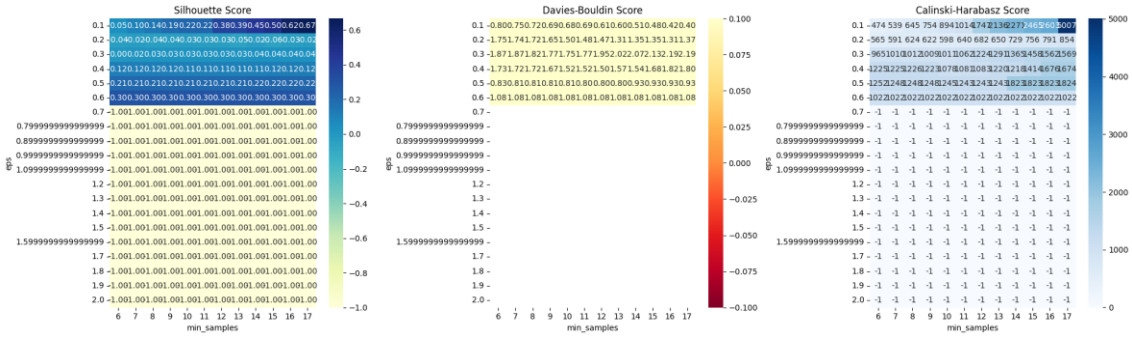
Gambar 7. Hasil Evaluasi GMM

Berdasarkan Gambar 7, hasil evaluasi GMM terhadap jumlah komponen dari 2 hingga 10 menunjukkan bahwa nilai BIC dan AIC sama-sama mencapai titik minimum pada $k = 10$, yang mengindikasikan bahwa model dengan 10 komponen memberikan keseimbangan terbaik antara kecocokan terhadap data dan kompleksitas model. Penurunan tajam nilai BIC dan AIC dari $k = 5$ ke $k = 6$ menandakan peningkatan signifikan dalam kualitas pemodelan, sementara kenaikan dari $k = 6$ ke $k = 7$, tetapi konsisten hingga $k = 9$, dan dilanjutkan dengan penurunan dari $k = 9$ ke $k = 10$. Ini menguatkan bahwa data memang lebih kompleks dan tidak cukup dijelaskan oleh sedikit *cluster*. Selain itu, dengan pendekatan probabilistik yang dimiliki GMM, menyebabkan hasil segmentasi lebih fleksibel karena setiap pelanggan tidak secara mutlak dikategorikan ke satu *cluster*, tetapi memiliki derajat keanggotaan terhadap beberapa *cluster*.

c. DBSCAN

DBSCAN, sebuah teknik *clustering* berbasis densitas, menciptakan *cluster* berdasarkan jumlah data yang relevan. DBSCAN memiliki dua parameter yang berbeda, yaitu eps (*epsilon*) yang menentukan radius tetangga maksimum, dan min_samples yang menyatakan jumlah minimum titik dalam radius eps agar suatu titik dianggap sebagai "*core point*". Pemilihan min_samples dari 6 hingga 17 didasarkan pada prinsip umum $\min\_samples \geq dimensi + 1$ yaitu $3 + 1$, serta mempertimbangkan ukuran data. Berbeda dengan K-Means atau GMM, DBSCAN tidak mengharuskan jumlah *cluster* ditentukan di awal dan memiliki keunggulan dalam mendeteksi *cluster* dengan bentuk arbitrer serta mengelompokkan *outlier* sebagai *noise* (bukan bagian dari *cluster* mana pun). Untuk mengevaluasi kualitas cluster, digunakan tiga metrik: *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*. DBSCAN tidak menghasilkan *cluster* untuk semua *set* parameter. Akibatnya, metrik tersebut diberikan nilai -1 yang berarti tidak valid.

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 8. Hasil Evaluasi DBSCAN

Hasil evaluasi pada Gambar 8 menunjukkan nilai tertinggi *Silhouette Score* (0.666) dan nilai terendah *Davies-Bouldin Score* (0.396), serta nilai *Calinski-Harabasz Score* yang sangat tinggi (5006.59), yang secara teoritis menunjukkan struktur *cluster* yang sangat baik. Namun, terlalu banyak *cluster* (21 *cluster*) dan terlalu banyak *noise* (8218 dari total data). Seperti yang ditunjukkan oleh Gambar 9, DBSCAN menyaring terlalu banyak data sebagai *outlier* pada parameter tertentu, meskipun pemisahan antar *cluster* sangat mudah. Hal mungkin tidak ideal dalam situasi bisnis seperti segmentasi pelanggan. Oleh karena itu, meskipun metrik menunjukkan hasil yang sangat baik, hasil *clustering* masih perlu diperiksa. Ini terutama benar dalam situasi di mana terlalu banyak suara untuk ditindaklanjuti secara strategis.

3.4. Evaluasi *Clustering*

Setelah mendapatkan hasil *clustering* pada setiap metode, tahapan berikutnya adalah melakukan pemilihan model terbaik menggunakan metrik evaluasi. Metrik evaluasi yang digunakan meliputi *Silhouette Score*, *Calinski-Harabasz Index* dan *Davies-Bouldin Index*. Hasil evaluasi untuk masing-masing metode tersaji pada Tabel 4.

Tabel 4. Perbandingan Hasil Metrik Evaluasi

Metode	Banyak *cluster*	Skor Silhoutte	Skor Calinski-Harabasz	Skor Davies-Bouldin
K-Means	5	0.308416	6191.029871	0.956978
GMM	10	0.207887	3707.304588	1.333408
DBSCAN	21	0.666477	5006.593880	0.395659

DBSCAN memang menunjukkan skor *Silhouette* tertinggi dan *Davies-Bouldin* terendah (nilai terbaik), tetapi tidak dipilih sebagai model. Hal ini disebabkan karena DBSCAN menghasilkan banyak *noise* (data tak terkelompok), meskipun skor *Silhouette*-nya tertinggi dan *Davies-Bouldin* terendah. Sebaliknya, K-Means yang menduduki peringkat kedua untuk skor *Silhouette* dan *Davies-Bouldin*, dengan *cluster* yang terkelompok rapi, dianggap metode terbaik untuk evaluasi *cluster* pada data ini (berdasarkan Gambar 10), terlepas dari skor *Calinski-Harabasz*-nya yang paling tinggi. *Centroid* untuk setiap *cluster* pada K-Means disajikan pada Tabel 5.

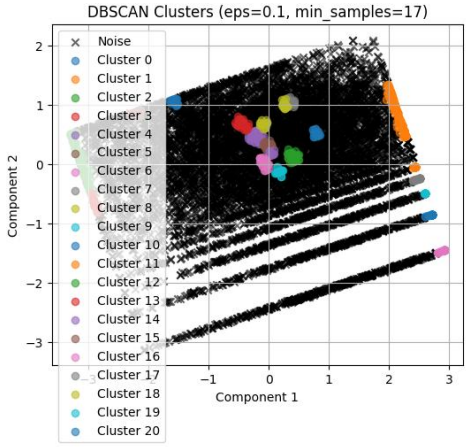
Tabel 5. *Centroid* pada Setiap *Cluster*

Cluster	Recency	Frequency	Monetary
0	21.279882	10.295130	25536.333371
1	1.808052	13.490122	34340.945533
2	15.320682	14.146096	36573.450222

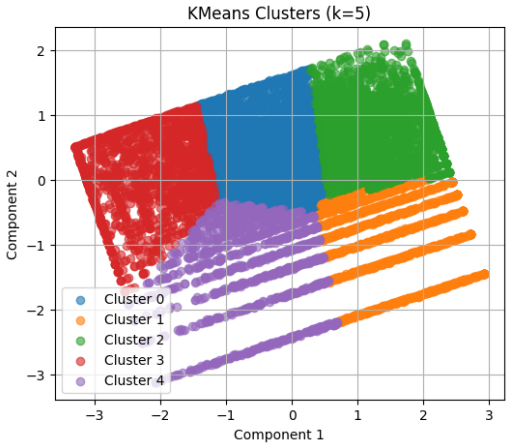
p-ISSN: 2986-3929
e-ISSN: 3032-1948

Cluster	Recency	Frequency	Monetary
3	27.537651	6.981287	15864.246402
4	2.694525	8.859342	20791.749400

Cluster 2 dan 1 sama-sama dihuni pelanggan bernilai tinggi yang sering bertransaksi dan mengeluarkan uang di atas $34.000. Bedanya, *Cluster* 1 sangat aktif (*recency* 1.8 hari), sedangkan *Cluster* 2 sedikit lebih pasif (*recency* 15 hari), ini bisa jadi indikasi awal penurunan aktivitas. *Cluster* 4 tampaknya mencakup pelanggan baru yang aktif (*recency* rendah) dengan pengeluaran sedang, menunjukkan peluang untuk berkembang. Sementara itu, *Cluster* 3 terdiri dari pelanggan dengan frekuensi dan pengeluaran rendah yang kemungkinan sudah pasif. Terakhir, *Cluster* 0, meskipun *recency*-nya tinggi, masih menunjukkan nilai transaksi menengah, menjadikannya kandidat untuk reaktivasi.



Gambar 9. Hasil *Clustering* DBSCAN



Gambar 10. Hasil Clustering K-Means (k=5)

4. Kesimpulan

Penelitian ini berhasil mengembangkan dan membandingkan segmentasi nasabah bank menggunakan model RFM dengan tiga algoritma *clustering*, yaitu K-Means, GMM, dan DBSCAN. Evaluasi performa menggunakan tiga metrik menunjukkan bahwa DBSCAN menghasilkan skor *Silhouette* tertinggi sebesar 0,666, skor Davies-Bouldin terendah sebesar 0,396, dan Calinski-Harabasz sebesar 5006,59. Namun, metode ini membentuk 21 *cluster* dengan 8.218 data (sekitar 91,3% dari total 9.000 nasabah) terklasifikasi sebagai *noise*, sehingga kurang relevan untuk aplikasi bisnis.

Sebaliknya, K-Means dengan 5 *cluster* menghasilkan skor *Silhouette* 0,308, Davies-Bouldin 0,957, dan Calinski-Harabasz 6191,03. Meskipun lebih rendah dibanding DBSCAN, metode ini dapat memberikan segmentasi yang lebih stabil. Hasilnya membentuk lima segmen pelanggan: (1) nasabah sangat aktif dengan pengeluaran tinggi (*recency* 1,8 hari dan *monetary* sekitar $34.340), (2) nasabah bernilai tinggi tetapi aktivitasnya mulai menurun (*recency* 15 hari dan *monetary* sekitar $36.573), (3) nasabah pasif dengan frekuensi dan pengeluaran rendah (*monetary* sekitar $15.864), (4) nasabah baru dengan potensi berkembang (*recency* 2,7 hari, *monetary* sekitar $20.791), dan (5) nasabah dengan transaksi menengah yang menjadi kandidat reaktivasi (*monetary* sekitar $25.536).

Dengan demikian, meskipun DBSCAN unggul secara metrik, K-Means dipilih sebagai metode terbaik karena menghasilkan lima *cluster* yang dapat ditindaklanjuti untuk strategi bisnis. Segmentasi ini dapat digunakan bank untuk menyusun strategi pemasaran terarah, seperti mempertahankan nasabah bernilai tinggi melalui loyalitas, mengaktifkan kembali nasabah pasif

dengan promosi khusus, serta mengembangkan nasabah baru dengan program edukasi dan loyalitas.

Daftar Pustaka

[1] M. A. Suharbi and H. Margono, “Kebutuhan transformasi bank digital Indonesia di era revolusi industri 4.0,” *Fair Value J. Ilm. Akunt. dan Keuang.*, vol. 4, no. 10, pp. 4749–4759, May 2022, doi: 10.32670/FAIRVALUE.V4I10.1758.

[2] A. Wardhana, “Pengaruh Kualitas Layanan Mobile Banking (M-Banking) Terhadap Kepuasan Nasabah di Indonesia,” *DeReMa (Development Research of Management) Jurnal Manajemen*, vol. 10, no. 2. pp. 273–284, 2015. doi: 10.19166/derema.v10i2.164.

[3] P. A. Lestari, M. I. Fasa, U. Islam, N. Raden, I. Lampung, and B. Lampung, “Transformasi Digital Banking : Manfaat Dan Risiko Transaksi Online Modern (Internet Banking Dan Mobile Banking) Transformasi Digital Banking : Manfaat Dan Risiko Transaksi Online Modern (Internet Banking Dan Mobile Banking),” vol. 3, no. 4, 2025.

[4] S. Winasis and S. Riyanto, “Transformasi Digital di Industri Perbankan Indonesia : Impak pada Stress Kerja Karyawan,” *IQTISHADIA J. Ekon. Perbank. Syariah*, vol. 7, no. 1, pp. 55–64, Jul. 2020, doi: 10.19105/IQTISHADIA.V7I1.3162.

[5] V. Bunga Tiara, A. M. Siregar, D. S. K. Kusumaningrum, and T. Rohana, “Bank Customer Segmentation Model Using Machine Learning,” *J. Nas. Pendidik. Tek. Inform.*, vol. 13, no. 1, pp. 66–79, 2024, doi: 10.23887/janapati.v13i1.75233.

[6] B. E. Adiana, I. Soesanti, and A. E. Permanasari, “Analisis Segmentasi Pelanggan Menggunakan Kombinasi Rfm Model Dan Teknik Clustering,” *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, Apr. 2018, doi: 10.21460/JUTEI.2018.21.76.

[7] Y. Yonata, H. M. #2, and C. Viona, “Analisis Clustering Pelanggan Berdasarkan Data Transaksi Penjualan Menggunakan Metode Recency, Frequency, Monetary (RFM) (Studi Kasus: CV XYZ),” *J. Telemat.*, vol. 16, no. 2, pp. 77–84, Jan. 2021, doi: 10.61769/TELEMATIKA.V16I2.379.

[8] M. Aliyev, E. Ahmadov, H. Gadirli, A. Mammadova, and E. Alasgarov, “Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning,” 2020, [Online]. Available: http://arxiv.org/abs/2008.08662

[9] R. Shirole, L. Salokhe, and S. Jadhav, “Customer Segmentation using RFM Model and K-Means Clustering,” *Int. J. Sci. Res. Sci. Technol.*, pp. 591–597, Jan. 2021, doi: 10.32628/IJSRST2183118.

[10] U. Firdaus and D. N. Utama, “Development Of Bank’s Customer Segmentation Model Based On Rfm+B Approach,” *ICIC Express Lett. Part B*, vol. 12, no. 1, pp. 17–26, 2021, doi: 10.24507/icicelb.12.01.17.

[11] X. Yan, Y. Li, F. Nie, and R. Li, “Bank Customer Segmentation and Marketing Strategies Based on Improved DBSCAN Algorithm,” *Appl. Sci. 2025, Vol. 15, Page 3138*, vol. 15, no. 6, p. 3138, Mar. 2025, doi: 10.3390/APP15063138.

[12] J. M. John, O. Shobayo, and B. Ogunleye, “An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market,” 2023, doi: 10.3390/analytics2040042.

[13] M. F. Fadhillah, A. Lovely, A. Suyoso, and I. Puspitasari, “Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency, Frequency dan Monetary (RFM),” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 1, pp. 48–56, Nov. 2025, doi: 10.57152/MALCOM.V5I1.1491.