



## *Customer Segmentation with Clustering Algorithm Based on Recency, Frequency, and Monetary (RFM) Attributes*

### **Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency, Frequency dan Monetary (RFM)**

Muhamad Fikri Fadhillah<sup>1</sup>, Aldo Lovely Arief Suyoso<sup>2</sup>, Ira Puspitasari<sup>3</sup>

<sup>1,2,3</sup>Sekolah Pascasarjana, Universitas Airlangga, Indonesia

E-Mail: <sup>1</sup>muhamad.fikri.fadhillah-2023@pasca.unair.ac.id,  
<sup>2</sup>aldo.lovely.arief-2023@pasca.unair.ac.id, <sup>3</sup>ira-p@fst.unair.ac.id

Received Jun 25th 2024; Revised Aug 21th 2024; Accepted Sep 22th 2024; Available Online Nov 24th 2024

Corresponding Author: Muhamad Fikri Fadhillah

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

#### **Abstract**

*Free Trade Agreement which have been signed with developed countries has caused changes in consumer character in Indonesia plus the effects of the Covid-19 pandemic, almost all domestic commodities have experienced a decline in sales and business profits every year [2]. Companies must change other strategies in attracting customers using different customer characteristic data, one of which is sales transaction data [3]. The data will later be segmented using the clustering method. Clustering is a data analysis technique that aims to group objects into groups or clusters based on similar characteristics or features. Data analysis with clustering is needed to identify patterns and extract information from data sets with large variations and quantities, such as identifying customer segmentation. This research uses the Recency, Frequency, and Monetary (RFM) model as the main attribute and k-means, agglomerative, and DBSCAN clustering techniques. Evaluation based on the silhouette score, Davis-Bouldin index, and Calinski-Harabasz index shows the best results in forming three clusters using the k-means clustering technique for customer segmentation in this study with a silhouette score index of 0.364, Davis-Bouldin of 0.93 and Calinski-Harabasz amounting to 1303.6 has produces 3 clusters of loyal customers, adequate customers and churn or lost customers.*

*Keyword: Agglomerative, Clustering, Customer Segmentation, DBSCAN, K-means, RFM*

#### **Abstrak**

Perdagangan bebas yang disepakati dengan negara maju menimbulkan perubahan karakter konsumen di Indonesia ditambah efek pandemi Covid-19 semua komoditi dalam negeri hampir mengalami penurunan penjualan dan laba usaha tiap tahunnya[2]. Perusahaan harus mengubah strategi lain dalam menarik pelanggan menggunakan data karakteristik pelanggan yang berbeda-beda salah satunya data transaksi penjualan [3]. Data tersebut nantinya akan disegmentasi dengan metode clustering. Clustering adalah teknik analisis data yang bertujuan mengelompokkan objek-objek ke dalam grup atau klaster berdasarkan kesamaan karakteristik atau fiturnya. Analisis data dengan clustering diperlukan untuk mengidentifikasi pola dan mengekstrak informasi dari kumpulan data dengan variasi dan jumlah besar, seperti pada identifikasi segmentasi pelanggan. Hasil clustering ini akan memudahkan dalam merumuskan startegi pemasaran berorientasi pelanggan. Penelitian ini menggunakan model Recency, Frequency, and Monetary (RFM) sebagai atribut utama dan teknik clustering K-Means, Agglomerative, dan DBSCAN. Evaluasi berdasarkan Silhouette Score, Davis-Bouldin Index, dan Calinski-Harabasz Index menunjukkan hasil terbaik pada pembentukan tiga cluster dengan algoritma k-means untuk segmentasi pelanggan dalam penelitian ini dengan indeks silhouette score sebesar 0.364, davis-bouldin sebesar 0.93 dan calinski-harabasz sebesar 1303.6 ini menghasilkan 3 klaster pelanggan loyal customer, adequate customer dan churn atau lost customer.

Kata Kunci: Agglomerative, Clustering, DBSCAN, K-means, RFM, Segmentasi Pelanggan

#### **1. PENDAHULUAN**

PT. SID merupakan perusahaan yang bergerak di bidang distributor material bahan bangunan dengan beberapa cabang di Indonesia [1]. Model bisnis perusahaan ini adalah *business to business* (B2B) yang mendistribusikan produknya kepada sesama pelaku bisnis [4]. PT SID berdiri pada tahun 1989 dan telah

menjadi pelaku utama pada industri distribusi semen dan non semen. Dengan ribuan pelanggan yang tersebar di seluruh Indonesia, perusahaan berkomitmen dalam menyediakan produk dan pelayanan berkualitas tinggi kepada para pelanggannya. Dalam era perdagangan bebas seperti saat ini, terjadi persaingan antar produsen untuk menghasilkan barang yang bermutu, efisiensi dan efektifitas tinggi karena tindakannya selalu didasarkan pada prinsip ekonomi [5]. Oleh karenanya, strategi pemasaran yang tepat menjadi hal yang sangat penting agar perusahaan dapat mempertahankan pangsa pasar dan menjaga pelanggan yang ada serta menarik pelanggan baru [6].

Saat ini berbagai perusahaan dalam pengembangan strategi pemasaran tidak hanya mengutamakan *product oriented*, tetapi juga mulai memperhatikan sisi *customer oriented* [7]. Alasannya adalah karena *customer* atau pelanggan juga sebagai salah satu aset perusahaan memiliki peranan penting dalam pengembangan bisnis dan sumber keuntungan perusahaan. Salah satu strategi orientasi pelanggan yang banyak dilakukan oleh perusahaan-perusahaan adalah melakukan segmentasi pelanggan. Segmentasi pelanggan adalah strategi dalam membagi pelanggan ke beberapa kelompok yang memiliki karakteristik atau perilaku yang sama [8]. Tujuannya dapat mengelompokkan pelanggan berdasarkan kesamaan kebutuhan dan karakteristik bisnisnya, sehingga perusahaan dapat memberikan penawaran atau layanan lebih maksimal dan tepat sasaran sesuai dengan aspek-aspek tersebut [2][9].

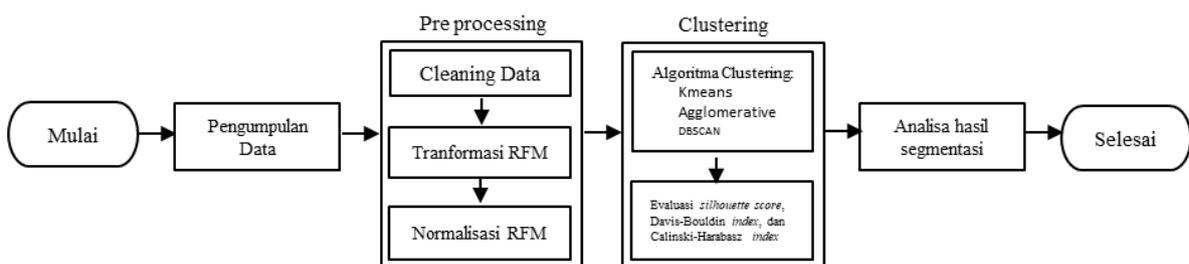
PT. SID yang berencana ingin melakukan perubahan strategi dengan segmentasi pelanggan memiliki ratusan ribu data penjualan riwayat transaksi apada *database* penjualan perusahaan yang dapat dimanfaatkan dalam membuat Model RFM. Model Recency (R), Frequency (F) dan Monetary (M) merupakan metode yang efektif dan populer saat ini untuk melakukan segmentasi berdasarkan data riwayat transaksi [10]. Model ini mengelompokkan pelanggan berdasarkan tiga karakter yaitu *Recency* adalah interval waktu sejak transaksi terakhir dari pelanggan, *Frequency* adalah jumlah pembelian yang dilakukan pelanggan, dan *Monetary* adalah jumlah uang yang dibelanjakan oleh pelanggan dalam bertransaksi dengan perusahaan [3][11].

Untuk dapat melakukan segmentasi pelanggan berdasarkan model RFM ini, dapat menggunakan teknik data mining yaitu *clustering*. *Clustering* merupakan teknik yang digunakan mengelompokkan data menjadi beberapa kelompok sehingga tiap data pada kelompok yang sama pasti memiliki kemiripan karakteristik yang tinggi, sedangkan data yang memiliki karakteristik berbeda akan ditempatkan pada kelompok yang lain [4][12]. Algoritma clustering yang digunakan dalam penelitian ini adalah k-means, *agglomerative* dan DBSCAN. Terdapat beberapa penelitian sebelumnya yang memiliki pendekatan yang sama. Penelitian dari Aulia Dewi Savitri dkk (2018) melakukan segmentasi pelanggan dengan model RFM dan algoritma k-means menghasilkan 3 segmen pelanggan dengan menunjukkan bahwa peringkat pertama merupakan kelompok pelanggan yang profitable dan seterusnya [13]. Penelitian lainnya oleh Alfi Fadliana dkk (2015) melakukan segmentasi pelanggan dengan algoritma *agglomerative* menghasilkan 4 segmen pelanggan klinik yaitu kondisi sehat, sedang, berat dan sangat berat [14]. Penelitian oleh Moch Irfan Chanafi (2019) menggunakan algoritma DBSCAN dan segmentasi pelanggan menghasilkan 5 kluster [15]. Dan penelitian algoritma k-means dengan modifikasi berdasarkan atribut RFM juga dilakukan oleh Darmawan dkk (2021) untuk segmentasi pelanggan nasabah bank [16] dan Hidayat & Rismawati (2020) untuk segmentasi mahasiswa di sebuah universitas [17].

Penelitian ini melakukan segmentasi pelanggan pada PT. SID dengan model RFM dan tiga algoritma clustering berdasarkan karakteristik dan perilakunya untuk mengidentifikasi pelanggan yang loyal dan menguntungkan perusahaan. Perbedaan penelitian ini dengan penelitian sebelumnya adalah pada subyek penelitian perusahaan B2B yang bergerak di bidang distribusi atau berlaku sebagai distributor bukan sebagai *retail*. Sehingga sampai saat ini, masih belum ada penelitian yang melakukan penelitian dengan data, tempat penelitian (perusahaan), dan percobaan yang sama. Hasil dari penelitian ini dapat bermanfaat bagi proses bisnis PT SID untuk pengembangan strategi pemasaran yang berorientasi pelanggan serta dapat memberikan penawaran produk, pelayanan lebih maksimal dan tepat sasaran.

## 2. METODOLOGI PENELITIAN

Penelitian ini bertujuan melakukan segmentasi pelanggan PT. SID untuk mengetahui jenis kategori pelanggan yang bertransaksi pada perusahaan menggunakan model RFM dan algoritma *clustering*. Seluruh proses dan tahapan penelitian sampai tujuan digambarkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

## 2.1 Pengumpulan Data

Proses pengumpulan diambil dari sistem database penjualan PT. SID di Gresik. Data yang digunakan adalah rekapan transaksi penjualan perusahaan selama 2 tahun terakhir mulai dari 1 Januari 2022 hingga 31 Desember 2023. Data tersebut memiliki total 118.315 transaksi dari 1.570 pelanggan yang melakukan transaksi dengan perusahaan.

## 2.2 Preprocessing Data

Tahap *preprocessing* terdiri dari tiga tahap yaitu *cleaning* untuk membersihkan data dari nilai yang tidak relevan, nilai yang hilang, dan lainnya. Tahap selanjutnya adalah transformasi data transaksi pelanggan menjadi variabel yang digunakan untuk membuat model RFM. Model ini memiliki tiga atribut yaitu *Recency* (R), *Frequency* (F) dan *Monetary* (M) [18]. *Recency* adalah interval waktu pelanggan sejak pembelian terakhir dengan jangka waktu tertentu, *Frekuensi* adalah jumlah pembelian yang dilakukan pelanggan dalam periode tertentu, dan *Moneter* adalah jumlah uang yang pelanggan keluarkan kepada perusahaan dalam periode tertentu [19]. Atribut RFM yang sudah jadi harus di normalisasi untuk mencegah adanya data yang memiliki nilai terlalu besar dibanding dengan nilai yang lain yang akan dapat mengakibatkan proses training tidak sesuai keinginan [16]. Menggunakan fitur *Standard Scaler* merupakan metode *preprocessing* di mana metode tersebut akan melakukan standarisasi fitur dengan menghapus rata-rata dan menskalakan unit varian. Proses ini akan dilakukan pada setiap fitur pada sampel [20]. *Preprocessing* ini dilakukan untuk mencegah adanya data yang memiliki nilai terlalu besar dibanding dengan nilai yang lain yang akan dapat mengakibatkan proses training tidak sesuai keinginan [21]. Rumus dari *Standard Scaler* ditunjukkan pada Persamaan (1), di mana  $X$  adalah rata-rata nilai sampel dan  $\sigma$  adalah standar deviasi [22].

$$Z = \frac{x_i - X}{\sigma} \quad (1)$$

## 2.3 Analisis Recency, Frequency, Monetary (RFM)

Analisis *Recency*, *Frequency*, *Monetary* (RFM) merupakan proses analisis perilaku pelanggan. Dalam menentukan segmentasi pelanggan, digunakan model RFM berdasarkan tiga variabel yaitu *recency* terakhir melakukan transaksi, *frequency* dari transaksi, dan *monetary* dari jumlah transaksi setiap pelanggan. Menurut Tsipsis dan Chorianopoulos (2009), analisis RFM terdiri *Recency*, *Frequency*, *Monetary*. *Recency* merupakan variabel untuk mengukur nilai pelanggan berdasarkan rentang waktu (tanggal, bulan, tahun) transaksi terakhir pelanggan sampai saat ini [23]. Semakin kecil rentang waktu maka nilai *recency* semakin besar. *Frequency* merupakan variabel untuk mengukur nilai pelanggan berdasarkan jumlah transaksi yang dilakukan pelanggan dalam satu periode. Semakin banyak jumlah transaksi yang dilakukan maka nilai  $f$  semakin besar. *Monetary* merupakan variabel untuk mengukur nilai pelanggan berdasarkan jumlah besaran uang yang dikeluarkan pelanggan dalam satu periode. Semakin banyak jumlah besaran uang yang dikeluarkan pelanggan maka nilai  $M$  semakin besar.

## 2.4 Clustering

*Clustering* adalah sebuah teknik *data mining* unsupervised learning untuk mengelompokkan suatu data yang memiliki kesamaan antara satu data dengan data yang lainnya [24]. Untuk menunjang keberhasilan proses analisis, maka diperlukan juga penerapan suatu algoritma. Ada tiga algoritma yang akan dibandingkan dalam penelitian ini. Pertama algoritma K-Means yaitu teknik pengelompokan data non-hirarki yang memisahkan data ke dalam cluster, mengelompokkan data dengan fitur yang sama bersama-sama dan mengelompokkan data dengan karakteristik yang berbeda ke dalam kelompok yang berbeda [12]. Kedua, *agglomerative hierarchical* yaitu mengelompokkan data yang mirip dalam hirarki yang sama dan yang tidak mirip di hirarki yang agak jauh dimana metode ini sering digunakan dibedakan menurut cara perhitungan tingkat kemiripan antar data [25]. Ketiga, DBSCAN adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data [26].

Untuk mengetahui algoritma mana yang memberikan hasil clustering hasil terbaik, dapat dilakukan dengan mengukur validitas cluster [27]. Ada tiga metrik yang digunakan sebagai pertimbangan nilai validitas *clustering*. Pertama, *Silhouette Score index* dengan dua variabel utama, yang pertama yaitu (a), merupakan jarak rata – rata antara datum ke- $i$  dengan semua data didalam sebuah cluster yang sama. Lalu terdapat (b), variabel  $b(i)$  merupakan nilai rata – rata jarak terkecil terhadap data ke- $i$  terhadap semua titik dari cluster yang lain Persamaan dari *silhouette index* dapat didefinisikan seperti dibawah ini:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

Rata – rata lebar nilai *silhouette* yang terbesar menunjukkan bahwa itu merupakan jumlah cluster yang terbaik [28]. Metrik kedua, *Davies Bouldin Index* (DBI) dimana Nilai DBI yang kecil menunjukkan hasil

clustering pada data tersebut terkelompok dengan baik dan semakin homogen [29] Berikut adalah persamaan untuk menghitung nilai DBI:

$$DBI = \frac{1}{k} \sum_{i=1}^k \max_{i=j} (R_{i,j}) \quad (3)$$

Matriks ketiga adalah Calinski-Harabasz Index yang juga dikenal sebagai Variance Ratio Criterion, dapat digunakan untuk evaluasi model clustering. Calinski-Harabasz Index adalah rasio jumlah dispersi antar kluster dan dispersi dalam kluster untuk semua kluster (dispersi didefinisikan sebagai jumlah kuadrat jarak) [30]. Untuk sekumpulan data X berukuran n yang telah dikelompokkan menjadi k kluster, Calinski-Harabasz Index dinyatakan dalam persamaan berikut :

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n-k}{k-1} \quad (4)$$

## 2.5 Algoritma Clustering

Tiga Algoritma yang akan dibandingkan dalam penelitian ini. Pertama algoritma K-Means. Kmeans termasuk salah satu algoritma clustering yang paling sering digunakan. Algoritma Kmeans adalah satu tipe clustering dengan konsep membagi n objek atau data menjadi K cluster. Output dari metode Kmeans ini bergantung pada jumlah K dan posisi titik centroid yang ditentukan di awal[28]. Langkah-langkah algoritma untuk melakukan Kmeans clustering adalah sebagai berikut [31] :

1. Menentukan nilai k sebagai jumlah cluster yang dibentuk
2. Menentukan nilai awal centroid atau titik pusat cluster. Nilai centroid ditentukan secara acak.
3. Menghitung jarak antara titik centroid dengan titik setiap objek, dapat dilakukan dengan menggunakan Euclidean Distance dengan rumus sebagai berikut:

$$De = \sqrt{(xi - si)^2 + (yt - tt)^2} \quad (5)$$

Dimana: De merupakan Euclidean Distance; i merupakan banyaknya data; (x,y) merupakan koordinat data; dan (s,t) merupakan koordinat centroid

4. Mengelompokkan data hingga terbentuk cluster dengan titik centroid dari tiap cluster merupakan titik centroid yang terdekat. Penentuan anggota cluster adalah dengan memperhitungkan jarak minimum objek.
5. Memperbarui nilai centroid setiap cluster.
6. Mengulangi langkah c hingga akhir sampai nilai titik centroid tidak berubah lagi.

Kedua, algoritma *agglomerative hierarchical* atau hierarki. *Clustering* dengan pendekatan hirarki akan mengelompokkan data yang mirip dalam hirarki yang sama dan yang tidak mirip di hirarki yang agak jauh. *Agglomerative* melakukan *clustering* dari N cluster menjadi satu kesatuan *cluster*, dimana N adalah jumlah data, sebaliknya *divisive* melakukan proses *clustering* dari satu *cluster* menjadi N *cluster* [32]. Beberapa metode hierarchical clustering yang sering digunakan dibedakan menurut cara perhitungan tingkat kemiripan. Ada yang menggunakan *Single Linkage*, *Complete Linkage*, *Average Linkage*, *Average Group Linkage* dan lain-lain. *Hierarchical clustering* dapat digambarkan melalui dendogram. Dendogram disusun dengan membuat *similarity matrix* yang memuat tingkat kemiripan antar data yang dikelompokkan. Tingkat kemiripan bisa dihitung dengan berbagai macam cara seperti dengan Euclidean *Distance* pada rumus (6).

$$D_{(x_2, x_1)} = \sqrt{\sum_{j=1}^d |x_{2j} - x_{1j}|^2} \quad (6)$$

Dimana: D merupakan Nilai Euclidean Distance; dan x merupakan Variabel ke-j dari obyek j (j=1, ..., n)

Ketiga, *Density-Based Spatial Clustering of Application with Noise* (DBSCAN) merupakan metode pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan yang dimaksud dalam DBSCAN adalah banyaknya data atau *minimum points* (MinPts) yang berada dalam radius Eps ( $\epsilon$ ) dari setiap data [33]. Konsep kepadatan seperti ini menghasilkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan (*noise*).

$$x \in N \text{ Eps } (y) \wedge |N \text{ Eps } (y)| \geq \text{MinPts} \quad (7)$$

DBSCAN akan melakukan clustering sesuai dengan parameter masukannya yaitu epsilon dan minpts. Jumlah cluster yang dihasilkan DBSCAN akan bergantung pada kedua parameter tersebut.

Langkah-langkah dalam algoritma DBSCAN sebagai berikut [34]:

1. Tentukan titik point p awal secara acak
2. Ambil semua point yang density reachable terhadap titik p
3. Jika p adalah core point maka cluster terbentuk
4. Jika p adalah border point, tidak ada yang merupakan hubungan density-reachable dari p dan DBSCAN akan menguji point selanjutnya.
5. Lakukan proses sampai semua point telah selesai diproses

### 3. HASIL DAN ANALISIS

Pengumpulan data dilakukan dengan penarikan data sistem informasi penjualan PT SID region Jawa Timur. Data yang digunakan adalah riwayat detail transaksi penjualan pelanggan mulai dari 1 Januari 2021 - 31 Desember 2023 dengan total 118.314 transaksi. Terdapat beberapa variabel yang tersedia pada data, namun tidak semua variabel tersebut akan digunakan saat melakukan segmentasi pelanggan nantinya. Dataset yang akan digunakan pada penelitian ini dapat dilihat pada Tabel .

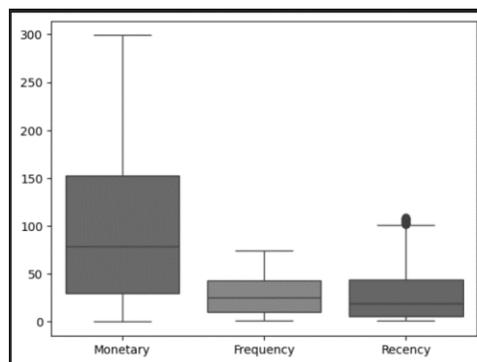
**Tabel 1.** Data Transaksi Pelanggan

Tanggal	Nama Cabang	Kode Customer	Billing	Kode Barang	...	Billing	Actual (Rp)
12/30/2023	SUMENEP	2125876	8500752575	1005176	...	8500752575	9,729,729
12/30/2023	SUMENEP	2125876	8500752576	1005405	...	8500752576	5,631,756
12/30/2023	SUMENEP	2125022	8500752916	1005407	...	8500752916	7,509,009
12/30/2023	SUMENEP	2121920	8500752571	1005405	...	8500752571	5,631,757
...	...	...	...	...	...	...	...
01/03/2022	LEGUNDI	2112708	8500753076	1005405	...	8500753076	9,671,171

Selanjutnya pada tahap preprocessing data yang terdiri dari *cleaning*, transformasi dan normalisasi. Saat melakukan *cleaning* data, nilai di tiap variabel yang tidak relevan akan dihapus seperti nilai yang kosong, outlier dan hanya akan mempertahankan nilai dari variabel yang penting saja. variabel yang akan dipertahankan yaitu Tanggal, Kode Customer, Billing, dan Actual Rp. Variable tersebut dipertahankan karena diperlukan untuk melakukan transformasi ke dalam model variable RFM sebelum ke proses clustering. Setelah data sudah ditransformasi ke dalam bentuk model RFM, dapat dicek dengan boxplot untuk melihat apakah data hasil transformasi RFM masih terdapat outlier. hasilnya seperti pada Gambar 2 tidak terdapat outlier sehingga dataset transformasi RFM akan dinormalisasi dengan standart scaler. Hasil dari proses *cleaning*, transformasi, dan normalisasi data dapat dilihat berturut-turut pada Tabel 2, Tabel 3 dan Tabel 4.

**Tabel 2.** *Cleaning* Data

No	Tanggal	Kode Customer	Billing	Actual (Rp)	Id_Pelanggan	Tanggal	Netto
1	12/30/2023	2125876	8500752575	9,729,729	1045/53I	11/01/2022	523243.25
2	12/30/2023	2125876	8500752576	5,631,756	1045/53I	11/01/2022	2318918.94
3	12/30/2023	2125022	8500752916	7,509,009	1045/53I	11/01/2022	324324.33
4	12/30/2023	2121920	8500752571	5,631,757	1045/53I	11/01/2022	973873.88
...	...	...	...	...	...	...	...
118.314	01/03/2022	2112708	8500753076	9,671,171	1045/42N	30/12/2022	973873.88



**Gambar 2.** Boxplot dari variable RFM

**Tabel 3.** Transformasi Data

No	Id_Pelanggan	Recency	Frequency	Monetary
1	2100023	60	15	94
2	2102602	38	44	232
3	2102603	38	58	220
4	2102604	31	67	269
...	...	...	...	...
1570	2500013	5	18	86

**Tabel 4.** Normalisasi Data

No	Id_Pelanggan	Recency	Frequency	Monetary
1	2100023	1,114	-0,203	0,370
2	2102602	0,709	0,797	1,145
3	2102603	0,709	1,054	1,099
4	2102604	0,528	1,188	1,272
...	...	...	...	...
1570	2500013	4	1	86

Selanjutnya adalah melakukan clustering menggunakan algoritma kmeans, *agglomerative* & DBSCAN dimana masing-masing model algoritma memberikan jumlah kluster yang berbeda-beda. Pada model algoritma kmeans menghasilkan 3 kluster, algoritma *agglomerative* menghasilkan 3 kluster dan algoritma DBSCAN menghasilkan 2 kluster. Dengan rincian jumlah kluster dan isinya seperti terlihat pada tabel 5.

**Tabel 5.** Jumlah Kluster dan isinya dari setiap model algoritma *clustering*

Jumlah Kluster	Kmeans	Agglomerative	DBSCAN
1	694	493	1478
2	383	742	34
3	494	336	-

Hasil model *clustering* kemudian divalidasi menggunakan *silhouette score*, DBI dan Calinski-Harabasz untuk mengetahui algoritma *clustering* yang cocok. Semakin besar nilai *silhouette score* dan Calinski-Harabasz maka semakin baik hasil clustering dan sebaliknya semakin kecil nilai DBI maka hasil clustering juga semakin baik. Table matriks nilai tiap model algoritma pada dapat dilihat pada Tabel 6.

**Tabel 6.** Matriks validasi tiap algoritma *clustering*

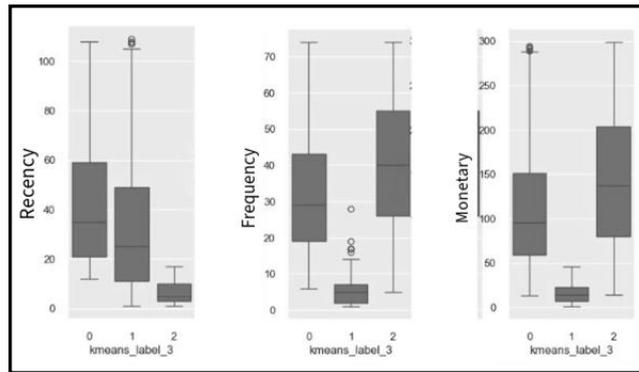
Matriks validasi	<i>silhouette score</i>	DBI	Calinski-Harabasz
Kmeans	0,364	0,93	1303,6
Agglomerative	0,354	0,914	1150,8
DBSCAN	0,407	1,232	212,63

Dari tabel matriks validasi diketahui algoritma kmeans memiliki nilai validasi terbaik karena parameter *silhouette score* tertinggi kedua sebesar 0.364, DBI terendah kedua sebesar 0.93 dan Calinski-Harabasz tertinggi pertama sebesar 1303.6 dengan demikian secara urutan ranking kumulatif algoritma kmeans yang paling baik untuk dibuat model *clustering* pada penelitian ini.

Tahapan selanjutnya adalah melakukan segmentasi disertai dengan analisis karakter pelanggan pada setiap segmennya berdasarkan hasil clustering algoritma kmeans. Setiap pelanggan akan diberi label kluster dengan kode 0, 1, 2 seperti ditunjukkan pada Tabel 8 dan karakteristik setiap labelnya berdasarkan nilai variable RFM ditunjukkan pada boxplot di Gambar 3.

**Tabel 7.** Data pelanggan dengan label segmentasi

Id_Pelanggan	Recency	Frequency	Monetary	Kmeans Label
2100023	60	15	94	0
2102602	38	44	232	0
2102603	38	58	220	0
2102604	31	67	269	0
...	...	...	...	...
2500013	5	18	86	2



**Gambar 3.** Boxplot dari setiap cluster berdasarkan variable RFM

Dari Gambar 3 dapat diketahui analisa karakter setiap klaster antara lain:

1. Cluster 0 terdiri dari pelanggan dengan rentang nilai *recency* tinggi, *frequency* dan *monetary* yang semuanya tinggi. Dengan skor tertinggi di semua kategori, maka klaster ini adalah *loyal customer*, pelanggan setia yang bersedia mengeluarkan banyak uang dan kemungkinan besar akan segera melakukan pembelian lagi. Pelanggan ini cenderung merespons program loyalitas. Mereka akan tertarik dengan produk yang baru diluncurkan. Karena pelanggan ini loyal terhadap merek dan produk, menawarkan diskon kepada mereka menjadi kurang masuk akal secara bisnis. Sebaliknya, Nilai *Lifetime Customer* mereka dapat ditingkatkan dengan merekomendasikan item dan produk mahal berdasarkan pembelian sebelumnya dengan pelayanan prioritas terbaik tentunya.
2. Cluster 1 mencakup pelanggan dengan rentang nilai *frequency* dan *monetary* yang rendah namun rentang nilai *recency* yang tinggi. Pelanggan ini lebih diprediksi cenderung merespons rekomendasi produk berdasarkan pembelian sebelumnya serta insentif yang terkait dengan ambang batas pengeluaran seperti pengiriman pembelian gratis atau hadiah gratis jika transaksi berada di atas nilai pesanan rata-rata merek. Atas respon.
3. Cluster 2 mencakup pelanggan dengan rentang nilai *frequency* dan *monetary* yang baik, namun rentang nilai *recency* yang rendah. Ini diprediksi adalah pelanggan yang sebelumnya berada di cluster lebih tinggi tetapi sekarang skornya lebih rendah. Inilah pelanggan yang berisiko mengalami kerugian atau beralih ke kompetitor dari PT SID. Oleh karenanya klaster ini harus ditargetkan dengan informasi benefit seperti harga diskon, peluncuran produk terbaru, dan penawaran eksklusif.

#### 4. KESIMPULAN

Penelitian Segmentasi pelanggan menggunakan model variabel RFM dan algoritma clustering. Dari data riwayat penjualan PT SID region Jawa Timur dari 1 Januari 2021 – 31 Desember 2023 yang terdiri dari 118.314 transaksi yang memuat 1.570 pelanggan. Untuk menentukan algoritma cluster yang tepat berdasarkan nilai dari 3 matriks validasi. Hasilnya menunjukkan algoritma kmeans dengan urutan terbaik dengan nilai *silhouette score* 0.364, DBI 0.93 dan Calinski-Harabasz 1303.6. Data pelanggan kemudian hasil clustering dengan algoritma kmeans menghasilkan 3 klaster pelanggan. Terdiri dari cluster 0 berisi 694 pelanggan dikategorikan sebagai loyal customer, cluster 1 yang terdiri dari 383 pelanggan dikategorikan sebagai adequate customer dan klaster 2 dengan 494 pelanggan dikategorikan sebagai churn atau lost customer Sebagai pelanggan yang loyal, perusahaan perlu memberikan apresiasi untuk mempertahankan hubungan baik dengan pelanggan seperti memberikan diskon, ataupun penawaran khusus. Kemudian untuk klaster dibawahnya bisa diberi rekomendasi produk dan program pemasaran. Sedangkan segmen terbawah *lost customer*, perusahaan perlu mengambil langkah yang tepat untuk mencoba memulihkan hubungan dengan pelanggan dan menganalisis faktor dan penyebab pelanggan pada klaster ini berhenti bertransaksi atau beralih ke kompetitor.

#### REFERENSI

- [1] R. D. Handoyo, L. Sugiharti, and M. A. E. Padilla, "Trade Creation And Trade Diversion Effects: The Case Of The Asean Plus Six Free Trade Area," *Buletin Ekonomi Moneter dan Perbankan*, vol. 24, no. 1, pp. 93–118, Apr. 2021, doi: 10.21098/bemp.v24i1.1163.
- [2] A. Siswati, "Dampak Pandemi Covid-19 pada Kinerja Keuangan (Studi kasus pada Perusahaan Teknologi yang Listing di BEI)," 2021. [Online]. Available: <http://jurnal.unw.ac.id/index.php/jibaku/index>

- [3] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1785–1792, May 2022, doi: 10.1016/j.jksuci.2019.12.011.
- [4] N. R. Maulina, I. Surjandari, and A. M. M. Rus, "Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering," in *2019 16th International Conference on Service Systems and Service Management (ICSSSM)*, IEEE, Jul. 2019, pp. 1–6. doi: 10.1109/ICSSSM.2019.8887739.
- [5] P. Bebas, O. : Syawitri, P. Bawon, E. R. Palandeng, and F. Baftim, "Dampak Perdagangan Bebas Pada Era Globalisasi Di Indonesia Dalam UU NO. 44 TAHUN 2007 Tentang Kawasan."
- [6] S. S. Prasetyo, M. Mustafid, and A. R. Hakim, "Penerapan Fuzzy C-Means Kluster Untuk Segmentasi Pelanggan E-Commerce Dengan Metode Recency Frequency Monetary (RFM)," *Jurnal Gaussian*, vol. 9, no. 4, pp. 421–433, Dec. 2020, doi: 10.14710/j.gauss.v9i4.29445.
- [7] P. Paduloh, M. Widyantoro, and J. Supratman, "Segmentasi Pelanggan Distributor Daging Sapi Menggunakan Pendekatan Recency, Frequency, Monetary (RFM) dan Fuzzy C-Means Clustering," *Jurnal Optimalisasi*, vol. 8, no. 1, p. 43, Apr. 2022, doi: 10.35308/jopt.v8i1.5181.
- [8] R. Nugraha, "RANCANG BANGUN SISTEM CRM (CUSTOMER RELATIONSHIP MANAGEMENT) BERBASIS WEB DENGAN POLA MVC," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 1, pp. 70–85, Mar. 2022, doi: 10.35957/jatisi.v9i1.1388.
- [9] E. Ernawati, S. S. K. Baharin, and F. Kasmin, "A review of data mining methods in RFM-based customer segmentation," *J Phys Conf Ser*, vol. 1869, no. 1, p. 012085, Apr. 2021, doi: 10.1088/1742-6596/1869/1/012085.
- [10] O. Dogan, B. Oztaysi, and A. Isik, "Fuzzy RFM Analysis in Car Rental Sector," *International Journal of Technology and Engineering Studies*, vol. 7, no. 2, 2021, doi: 10.20469/ijtes.7.10002-2.
- [11] Y. Parikh and E. Abdelfattah, "Clustering Algorithms and RFM Analysis Performed on Retail Transactions," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, Oct. 2020, pp. 0506–0511. doi: 10.1109/UEMCON51285.2020.9298123.
- [12] S. Monalisa and F. Kurnia, "Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, p. 110, Feb. 2019, doi: 10.12928/telkomnika.v17i1.9394.
- [13] A. D. Savitri, F. A. Bachtiar, and N. Y. Setyawan, "Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus: Belle Crown Malang)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 9, pp. 2957–2966, Feb. 2018, [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2489>
- [14] J. Matematika Fakultas Sains Dan, "Penerapan Metode Agglomerative Hierarchical Clustering Untuk Klasifikasi Kabupaten/Kota Di Provinsi Jawa Timur Berdasarkan Kualitas Pelayanan Keluarga Berencana SKRIPSI OLEH ALFI FADLIANA NIM. 10610077," 2015.
- [15] M. I. Chanafi, D. P. Hapsari, R. K. Hapsari, and T. Indriyani, "Implementasi Algoritma Clustering Untuk Pengelompokan Pelanggan Retail Berdasarkan Skor Recency, Frequency, Dan Monetary."
- [16] arief soma darmawan, D. Sugianti, and A. Syaifudin, "Segementasi Nasabah Tabungan Pada BMT XXX dengan Metode Fuzzy C Means dan Model RFM," *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 10, no. 2, pp. 74–78, Jun. 2021, doi: 10.30591/smartcomp.v10i2.2355.
- [17] S. Hidayat, R. Rismayati, M. Tajuddin, and N. L. P. Merawati, "Segmentation of university customers loyalty based on RFM analysis using fuzzy c-means clustering," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 133–139, Apr. 2020, doi: 10.14710/jtsiskom.8.2.2020.133-139.
- [18] V. Hermawati and R. Sulaiman, "Penentuan Segmentasi Pelanggan E-Commerce Menggunakan Fuzzy C-Means Dan Model Fuzzy Rfm," *MATHunesa: Jurnal Ilmiah Matematika*, vol. 9, no. 1, pp. 76–88, Jan. 2021, doi: 10.26740/mathunesa.v9n1.p76-88.
- [19] A. Wicaksono, F. A. Bachtiar, and N. Y. Setiawan, "Segmentasi Pelanggan Menggunakan Fuzzy C-Means Clustering berdasarkan RFM Model pada E-Commerce (Studi Kasus: E-Commerce XYZ)," 2021. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [20] A. Dursun and M. Caber, "Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis," *Tour Manag Perspect*, vol. 18, pp. 153–160, Apr. 2016, doi: 10.1016/j.tmp.2016.03.001.
- [21] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl Soft Comput*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.
- [22] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 5, no. 1, pp. 401–413, Dec. 2023, doi: 10.37385/jaets.v5i1.3080.
- [23] "Segmentation Applications in Telecommunications," in *Data Mining Techniques in CRM*, Wiley, 2010, pp. 291–332. doi: 10.1002/9780470685815.ch7.

- 
- [24] V. Herlinda and D. Darwis, "Analisis Clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means," *Darwis, Dartono*, vol. 2, no. 2, pp. 94–99, 2021, [Online]. Available: <http://jim.teknokrat.ac.id/index.php/JTSI>
- [25] W. Widyawati, W. L. Y. Saptomo, and Y. R. W. Utami, "Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan," *Jurnal Ilmiah SINUS*, vol. 18, no. 1, p. 75, Jan. 2020, doi: 10.30646/sinus.v18i1.448.
- [26] I. D. Id, "Modifikasi DBSCAN (Density-Based Spatial Clustering with Noise) Pada Objek 3 Dimensi," *Jurnal Komputer Terapan*, vol. 3, no. 1, 2017.
- [27] A.-Y. Al-Yasir, M. Afdal, Z. Zarnelly, and A. Marsal, "Analisis Loyalitas Pelanggan Business To Business Berdasarkan Model RFM Menggunakan Algoritma Fuzzy C-Means," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 359–365, Jan. 2024, doi: 10.57152/malcom.v4i1.1163.
- [28] T. Kodinariya and P. Makwana, "Review on Determining of Cluster in K-means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90–95, Jan. 2013.
- [29] Y. Arie Wijaya, D. Achmad Kurniady, E. Setyanto, W. Sanur Tarihoran, D. Rusmana, and R. Rahim, "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities," *TEM Journal*, pp. 1099–1103, Aug. 2021, doi: 10.18421/TEM103-13.
- [30] S. Asteriadis, K. Karpouzis, N. Shaker, and G. N. Yannakakis, "Towards Detecting Clusters of Players using Visual and Gameplay Behavioral Cues," *Procedia Comput Sci*, vol. 15, pp. 140–147, 2012, doi: 10.1016/j.procs.2012.10.065.
- [31] Tan, Pang-Ning, M. Steinbach, M. Adeyeye Oshin, V. Kumar, and Vipin, *Introduction to Data Mining*. 2005.
- [32] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. Wiley, 2011. doi: 10.1002/9780470977811.
- [33] W.-T. Wang, Y.-L. Wu, C.-Y. Tang, and M.-K. Hor, "Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data," in *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, IEEE, Jul. 2015, pp. 445–451. doi: 10.1109/ICMLC.2015.7340962.
- [34] W. Rohalidyawati, R. Rahmawati, and M. Mustafid, "Segmentasi Pelanggan E-Money Dengan Menggunakan Algoritma Dbscan (Density Based Spatial Clustering Applications With Noise) Di Provinsi Dki Jakarta," *Jurnal Gaussian*, vol. 9, no. 2, pp. 162–169, May 2020, doi: 10.14710/j.gauss.v9i2.27818.