Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1346-1361
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparison of the DBSCAN and K-MEANS Algorithms in Segmenting Customers Using Public Transportation of Transjakarta Using the RFM Method

Perbandingan Algoritma DBSCAN dan K-MEANS dalam Segmentasi Pelanggan Pengguna Transportasi Publik Transjakarta Menggunakan Metode RFM

Aditiya Saputra[1*], Raka Yusuf[2]

[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer,
Universitas Mercu Buana, Indonesia

E-Mail: [1]AditiyaS1811@gmail.com, [2]raka@mercubuana.ac.id

Received Jun 6th 2024; Revised Jul 21th 2024; Accepted Jul 26th 2024
Corresponding Author: Aditiya Saputra

Abstract

Public transportation is very important in the life of individuals in an area. Transjakarta, introduced in 2004, is the longest Bus Rapid Transit (BRT) system in the world with 251.2 kilometers of routes, 14 corridors, and 287 stops throughout Jakarta. The system serves the public with 1,347 transportation units. As the number of users increases, issues such as crowding at stops and long queues arise, necessitating precise customer segmentation.This study employs the Recency, Frequency, Monetary (RFM) method for analyzing Transjakarta customer segmentation using DBSCAN and K-Means algorithms. The results show that DBSCAN requires longer processing times for certain clusters, while K-Means is faster in specific clusters. K-Means excels with a Silhouette Score of 0.714917 and a Davies-Bouldin Index of 0.365776, compared to DBSCAN's Silhouette Score of 0.699971 and Davies-Bouldin Index of 0.390784. K-Means is more effective in distinguishing customers based on frequency and monetary value, while DBSCAN can identify outliers with high interaction and monetary value. Overall, K-Means demonstrates better performance in segmenting Transjakarta customers.Based on these results, K-Means is more suitable for Transjakarta customer segmentation, which can help authorities design more efficient service strategies and improve customer satisfaction.

Keywords: DBSCAN, K-Means, Public Transportation, RFM, Segmentation

Abstrak

Transportasi umum sangat penting dalam kehidupan individu di suatu daerah. Transjakarta, diperkenalkan pada tahun 2004, adalah sistem Bus Rapid Transit (BRT) terpanjang di dunia dengan 251,2 kilometer jalur, 14 jalur, dan 287 halte di seluruh Jakarta. Sistem ini melayani masyarakat dengan 1.347 unit transportasi. Seiring peningkatan jumlah pengguna, masalah seperti kerumunan di halte dan antrian panjang muncul, sehingga diperlukan segmentasi pelanggan yang cermat.Penelitian ini menggunakan metode Recency, Frequency, Monetary (RFM) untuk analisis segmentasi pelanggan Transjakarta dengan algoritma DBSCAN dan K-Means. Hasil menunjukkan DBSCAN membutuhkan waktu pemrosesan lebih lama untuk klaster tertentu, sedangkan K-Means lebih cepat di klaster tertentu. K-Means unggul dengan Silhouette Score 0.714917 dan Davies-Bouldin Index 0.365776, dibandingkan DBSCAN dengan Silhouette Score 0.699971 dan Davies-Bouldin Index 0.390784. K-Means lebih efektif dalam membedakan pelanggan berdasarkan frekuensi dan nilai moneter, sementara DBSCAN dapat mengidentifikasi outlier dengan interaksi dan nilai moneter tinggi. Secara keseluruhan, K-Means menunjukkan performa yang lebih baik dalam segmentasi pelanggan Transjakarta. Berdasarkan hasil ini, K-Means lebih cocok digunakan untuk segmentasi pelanggan Transjakarta, yang dapat membantu pihak berwenang merancang strategi layanan yang lebih efisien dan meningkatkan kepuasan pelanggan.

Kata Kunci: DBSCAN, K-Means, RFM, Segmentasi, Transportasi Umum

1. PENDAHULUAN

Transportasi umum memainkan peran yang sangat penting dalam kehidupan individu di suatu daerah. Bukan hanya sebagai alternatif, tetapi juga salah satu kebutuhan yang paling penting yang mendukung berbagai kegiatan warga setiap harinya. Pemerintah mendorong Transjakarta pada tahun 2004 untuk memenuhi permintaan akan transportasi yang lebih baik dan mendukung olahraga masyarakat. Transjakarta merupakan kerangka kerja Bus Rapid Transit (BRT) dengan jalur terpanjang di dunia, yaitu 251,2 kilometer. Terdapat 14 jalur yang membentang di seluruh kota Jakarta, dengan 287 halte yang dapat diakses sepanjang waktu. Saat ini, kerangka kerja ini memiliki 1.347 unit transportasi yang terdiri dari dua kategori, yaitu transportasi tunggal dan transportasi ganda. Dalam memberikan pengaturan transportasi yang dapat diandalkan dan efektif bagi penduduk Jakarta, kolaborasi antara segmen publik dan swasta telah terbukti efektif dengan Transjakarta [1].

Namun permasalahan muncul seiring dengan kemajuan. Peningkatan jumlah pengguna atau pelanggan Transjakarta merupakan suatu hal yang besar. Semakin banyak orang yang bergantung pada layanan ini untuk bepergian, sehingga menyebabkan peningkatan penggunaan Transjakarta secara signifikan. Peningkatan pelanggan ini dapat menimbulkan banyak masalah, termasuk penimbunan. Dengan semakin banyaknya orang yang menggunakan Transjakarta, terjadi kerumunan di halte, antrian panjang dan lalu lintas padat. Hal ini dapat merepotkan pengguna dan bahkan mengganggu kinerja secara keseluruhan.

Oleh karena itu, segmentasi pelanggan yang cermat diperlukan untuk memahami perilaku, kebutuhan, dan preferensi berbagai kelompok pengguna di Transjakarta. Dengan memahami kelompok pengguna secara lebih spesifik, pihak berwenang dapat merancang strategi yang lebih efektif seperti optimasi rute dan jadwal, peningkatan retensi pelanggan Transjakarta dan pengembangan program loyalti, yang kemungkinan besar bisa dijadikan untuk mengatasi keterlambatan dan masalah lain yang timbul akibat pertumbuhan pesat pengguan Transjakarta.

Dalam konteks penelitian ini, metode RFM (*Recency*, *Frequency*, *Monetary*) digunakan sebagai teknik analisis yang efektif untuk segmentasi pelanggan Transjakarta. RFM memungkinkan untuk mengelompokkan pelanggan berdasarkan seberapa baru (*Recency*) mereka menggunakan layanan, seberapa sering (*Frequency*) mereka melakukan transaksi, dan jumlah total uang (*Monetary*) yang mereka habiskan untuk transaksi [2]. K-Means merupakan algoritma clustering berbasis centroid yang membagi data ke dalam k cluster berdasarkan jarak terdekat ke centroid, sangat efektif untuk data yang terdistribusi secara merata. Di sisi lain, DBSCAN adalah algoritma clustering berbasis kepadatan yang membentuk cluster dari area dengan kepadatan tinggi, cocok untuk data dengan struktur yang tidak beraturan dan memiliki noise. Kedua algoritma ini memanfaatkan data RFM sebagai variabel input untuk mengidentifikasi kelompok pelanggan dengan karakteristik serupa, yang kemudian dapat ditargetkan dengan strategi pemasaran yang spesifik untuk meningkatkan kepuasan dan loyalitas pelanggan. [3].

Penelitian sebelumnya oleh Susanto (2022) [2], menunjukkan bahwa penggunaan K-Means dengan RFM berhasil mengidentifikasi pola pelanggan dengan membentuk tiga cluster, di mana cluster dengan akurasi tertinggi memiliki nilai silhouette coefficient sebesar 0.7578. Sementara itu, penelitian oleh Pata et al. (2023) [4], menunjukkan bahwa DBSCAN dengan RFM mampu membentuk dua cluster, dengan nilai silhouette coefficient tertinggi sebesar 0.6134. Kelebihan dari penelitian ini adalah kemampuan K-Means dalam mengelompokkan data yang homogen dan efisien dalam menangani volume data besar, serta kemampuan DBSCAN dalam menangani data dengan distribusi yang tidak beraturan dan mengidentifikasi noise, memberikan pemahaman yang lebih dalam tentang pola perilaku pelanggan.

Algoritma K-Means merupakan metode clustering berbasis jarak yang digunakan dalam analisis data untuk mempartisi kumpulan data menjadi beberapa cluster atau kelompok yang berbeda. Metode K-Means mempunyai beberapa kelemahan yang signifikan. Kelebihannya adalah kemudahan implementasi, memungkinkan proses yang cepat dan efisien, membandingkan kecepatan relatif pada komputasinya dengan metode metode lainnya, metode ini cocok untuk kumpulan data besar [5]. DBSCAN (*Density-based spatial clustering ofapplications with noisealgorithm)* menyediakan metode yang sangat sederhana untuk mengumpulkan kumpulan data, DBSCAN dapat mengidentifikasi wilayah dengan kepadatan tinggi dan menggabungkanya dengan wilayah kepadatan rendah berdasarkan karakteristik kepadatan yang berbeda setiap cluster. Metode ini melakukan analisi yang efisien berdasarkan nilai ambang batas, memungkinkan kumpulan data dianalisis dengan presisi tinggi. Bekerja dengan baik ketika diterapkan pada berbagai format dataset dengan data spasial sebagai dasar dan cukup efisien ketika menangani data yang sangat jarang [6].

Dengan kelebihan pada algoritma K-Means dan DBSCAN penulis ingin melakukan implementasi kedua algoritma tersebut dengan clustering untuk segmentasi pelanggan menggunakan metode RFM, serta menentukan algoritma mana yang lebih cocok untuk segmentasi pelanggan studi kasus transportasi publik Transjakarta.

2. TEORI PENDUKUNG

2.1 Segmentasi Pelanggan

Segmentasi adalah proses membagi pelanggan menjadi beberapa kelompok berdasarkan tingkat loyalitasnya, dengan tujuan untuk mengembangkan strategi pemasaran yang lebih efektif. Dalam konteks

pemasaran, segmentasi memegang peranan penting dalam pemasaran relasional yang bertujuan untuk meningkatkan kualitas hubungan dengan pelanggan, menjadikan mereka lebih menarik, dan lebih memahami kebutuhan mereka.

Metode segmentasi pelanggan membantu mengidentifikasi berbagai jenis pelanggan, sehingga memungkinkan pengambilan keputusan yang lebih baik melalui pemahaman yang lebih baik tentang pelanggan. Hasil dari segmentasi pelanggan ini dapat digunakan sebagai panduan untuk mengembangkan strategi pemasaran dan cross-selling produk baru yang disesuaikan dengan masing-masing kelompok serta untuk mengembangkan produk yang lebih cocok untuk kelompok pelanggan yang bernilai lebih tinggi [3].

2.2 Algoritma Clustering

Clustering merupakan suatu teknik dalam bidang data mining yang bertujuan untuk mengelompokkan data ke dalam cluster atau kelompok berdasarkan kesamaan karakteristik. Proses clustering melibatkan pengelompokan data berdasarkan jarak terdekat dengan objek lain dalam kumpulan data dan data tersebut dikelompokkan secara acak. Ada banyak metode pengelompokan berbeda yang dapat diterapkan pada kumpulan data besar [5]. Pada penelitian ini metode clustering yang digunakan untuk mengelompokkan pelanggan adalah DBSCAN dan K-Means.

1. **DBSCAN**

   Algoritma DBSCAN merupakan salah satu algoritma non-parametrik unsupervised learning yang artinya tidak bergantung pada asumsi-asumsi tertentu pada saat melakukan clustering data. Pada prinsipnya DBSCAN dapat membentuk cluster dengan bentuk acak atau tidak terbatas dan dengan mudah mengatasi situasi noise atau outlier di dalam cluster. Algoritma ini mengidentifikasi wilayah dengan kepadatan tinggi sebagai cluster, menggunakan dua parameter yang harus ditentukan secara cermat. Parameter tersebut adalah radius batasan, dilambangkan dengan ε, dan jumlah minimum objek (minObj) yang diperlukan untuk menentukan apakah suatu wilayah bertipe blok, dilambangkan dengan MinObj [7].Persamaan umum yang digunakan dalam DBSCAN antara lain :

$$\epsilon - neighborhood = \{ y \in D \mid dist(x, y) \leq \epsilon \} \quad (1)$$

   Di mana :

   ϵ : adalah radius lingkungan.
   D : adalah dataset.
   dist(x, y) : adalah jarak antara data x dan y.

2. **K-Means**

   K-Means merupakan algoritma di bidang unsupervised data mining yang menggunakan centroid sebagai pusat setiap kelompok data (cluster). K-Means merupakan algoritma yang relatif sederhana, dimana tingkat kemiripan antar data mempunyai pengaruh yang besar dalam proses clustering menggunakan K-Means. Algoritma clustering K-Means sering digunakan karena kesederhanaannya dan kemampuannya untuk melakukan konvergensi dengan cepat. Namun perlu diperhatikan bahwa nilai K (jumlah cluster yang diinginkan) harus ditentukan terlebih dahulu, dan pemilihan nilai K ini akan langsung mempengaruhi hasil konvergensi. Algoritma K-Means dapat dijelaskan dengan mempertimbangkan representasinya sebagai sebuah cluster, yang nilai pusat atau meannya merupakan titik sentral dari setiap kelompok data [8]. Persamaan umum yang digunakan dalam K-Means antara lain:

$$J = \sum_{i}^{k} = 1 \ \sum_{j}^{n} = 1 \ || \ x_j^{(i)} - \mu_i \ ||^2 \quad (2)$$

   Di mana :

   J : adalah fungsi objektif yang ingin diminimalkan.
   k : adalah jumlah cluster.
   n : adalah jumlah data.
   $x_j^{(i)}$: adalah data ke-j yang termasuk dalam cluster ke-i.
   μ_i : adalah centroid dari cluster ke-i.
   $|| \ x_j^{(i)} - \mu_i \ ||^2$: adalah jarak Euclidean antara data $x_j^{(i)}$ dan centroid μ_i.

2.3 Metode RFM

Model RFM diketahui mengklasifikasikan pelanggan ke dalam kelompok berbeda dengan menganalisis data historis transaksi mereka. Hal ini melibatkan faktor-faktor seperti frekuensi pembelian terakhir pelanggan, frekuensi aktivitas pembelian mereka, dan nilai moneter dari pengeluaran mereka. Namun, model RFM tradisional tidak menggunakan atribut konsumen tambahan dan hanya mempertimbangkan variabel terkait

transaksi seperti recency pembelian, frekuensi pembelian, dan nilai moneter. Oleh karena itu, banyak penelitian telah dilakukan untuk memasukkan variabel tambahan ke dalam model RFM konvensional dan menggunakan teknik pembelajaran mesin untuk meningkatkan efisiensi segmentasi pelanggan [9].

2.4 Mobilitas Penumpang

Teori mobilitas penumpang mencakup konsep bahwa pergerakan dan aksesibilitas yang efisien dalam sistem transportasi memainkan peran krusial dalam mempengaruhi pengalaman dan keputusan mobilitas individu. Faktor-faktor seperti kecepatan, ketersediaan moda transportasi, dan kualitas layanan memengaruhi keputusan pemilihan transportasi oleh penumpang. Teori ini juga mengaitkan mobilitas dengan aspek sosial dan ekonomi, di mana aksesibilitas yang baik dapat meningkatkan konektivitas antarwilayah, meningkatkan kesempatan pekerjaan, dan memperkuat integrasi sosial [10]. Teori mobilitas penumpang juga mempertimbangkan peran teknologi dalam mengubah pola mobilitas, seperti kemajuan dalam transportasi berbasis teknologi digital yang dapat memfasilitasi pengalaman perjalanan yang lebih efisien dan terkoneksi. Dengan memahami teori mobilitas penumpang, pengembangan kebijakan transportasi yang lebih berkelanjutan dan sistem mobilitas yang inklusif dapat dihasilkan, meningkatkan efisiensi dan kualitas hidup di masyarakat.

2.5 Industri Transportasi Publik

Pemerintah terus melakukan upaya peningkatan kualitas pelayanan angkutan umum dengan tujuan mendorong masyarakat untuk memilih angkutan umum sebagai sarana transportasi utama mereka [11]. Namun kenyataannya, belum semua orang sepenuhnya beralih menggunakan transportasi umum. Masih banyak pengguna kendaraan pribadi contohnya kendaraan roda dua.

2.6 Analisis Data

Analisis data adalah sumber pengetahuan berharga yang dapat mengungkapkan bagaimana sebuah bisnis mengambil keputusan. Namun, setiap bisnis menghadapi tantangan unik, seperti proses pengumpulan data, visualisasi, dan analisis data pelanggan, yang dianggap sebagai aset bisnis paling berharga. Dengan tujuan membuka peluang baru, menarik lebih banyak pelanggan, meningkatkan keuntungan, dan meningkatkan kualitas pengambilan keputusan, setiap bisnis memulai analisis data. Dalam konteks umum, perusahaan sering kali memiliki data pelanggan dalam jumlah besar, dan penerapan Ilmu Data dapat membantu mereka mendapatkan umpan balik pelanggan yang nyata untuk pengembangan produk dan desain strategi pemasaran dengan lebih efisien [12].

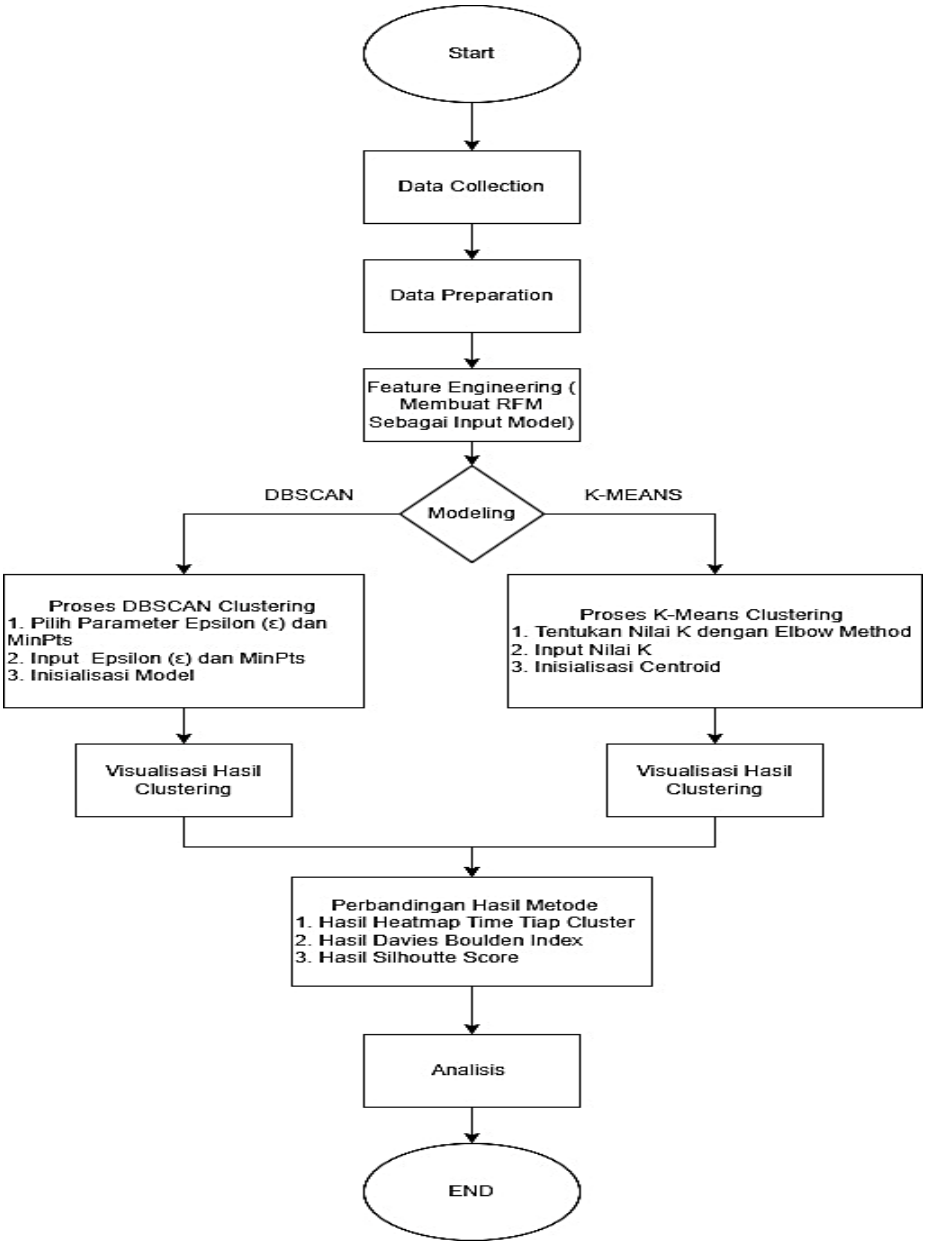
2.7 Pemrograman Komputer

Pada penelitian ini teori pendukung pemrograman komputer digunakan untuk clustering yaitu bahasa pemrograman Python dimana telah menjadi pilihan utama dalam dunia analisis data karena berbagai alasan. Python dikenal mudah dipelajari dan dapat digunakan oleh berbagai usia. Manfaat Python tidak terbatas pada kemudahan penggunaan. Bahasa ini juga dilengkapi dengan banyak perpustakaan yang melayani berbagai tujuan, memungkinkan pengguna untuk menggunakannya terlepas dari sistem operasi yang mereka gunakan.

Python bersifat open source, artinya kode sumbernya dapat dilihat dan dimodifikasi secara bebas. Beberapa perpustakaan Python penting termasuk NumPy, Pandas, Matplotlib, dan Scikit-learn. Masing-masing perpustakaan ini memainkan peran tertentu dalam analisis data, pemodelan statistik, visualisasi data, dan pembelajaran mesin. Kekuatan Python tidak berhenti di situ, ini juga sangat kompatibel dan dapat dengan mudah diintegrasikan dengan banyak teknologi lainnya, termasuk database, alat big data, kerangka web, dan banyak lagi. Hal ini memungkinkan pengguna untuk mengakses dan mengelola data dari berbagai sumber yang relevan dalam konteks analisis data [13].

3. METODE PENELITIAN

Pendekatan penelitian pada studi kasus ini adalah kuantitatif. Pendekatan kuantitatif digunakan untuk mengumpulkan dan menganalisis data secara numerik dengan tujuan memahami perbandingan antara algoritma DBSCAN dan K-Means dalam proses segmentasi pelanggan pengguna transportasi publik menggunakan metode RFM. Dengan menggunakan metode kuantitatif, penelitian ini akan mengandalkan data angka dan statistik untuk menyusun pemahaman yang lebih objektif dan mendalam tentang efektivitas kedua algoritma tersebut dalam konteks yang spesifik, yaitu studi kasus pada pengguna transportasi publik di Transjakarta. Pendekatan ini memungkinkan peneliti untuk mengukur dan membandingkan hasil secara sistematis, memberikan dasar yang kuat untuk menarik kesimpulan, dan mendukung generalisasi temuan terkait algoritma clustering dan segmentasi pelanggan dalam konteks transportasi publik. Berikut adalah gambaran umum tahapan penelitian yang diambil dalam melaksanakan penelitian ini menggunakan diagram alir yang ditunjukkan pada gambar 1.



Gambar 1. Tahapan Penelitian

Keterangan diagram alir penelitian:

1. Data Collection
   Tahapan ini melibatkan pengumpulan data dari sumber eksternal, yaitu Kaggle, yang berisi Transjakarta General Transit Feed Specification. Pengumpulan data dilakukan sesuai dengan prosedur penelitian yang akan dijalankan guna menghimpun informasi [14].

2. Data Preparation
   Pada tahapan ini mempersiapkan data yang akan digunakan dalam analisis clustering, adapun pada penlitian ini proses data preparation meliputi, pembersihan data dan normalisasi data[15].

3. Feature Engineering
   Feature Engineering adalah proses pembuatan fitur baru yang lebih relevan dan bermakna dari data mentah yang ada[16]. Pada penelitian ini, membuat RFM (Recency, Frequency, Monetary) dan dilakukan Normalisasi data sebagai input untuk model:
   a. Recency: Seberapa baru pelanggan terakhir kali melakukan transaksi.
   b. Frequency: Seberapa sering pelanggan melakukan transaksi dalam periode waktu tertentu.
   c. Monetary: Total nilai uang yang dihabiskan oleh pelanggan dalam periode waktu tertentu.

d. Normalisasi data RFM : Min-Max Scaling dilakukan yaitu teknik normalisasi yang mengubah nilai-nilai fitur sehingga berada dalam rentang antara 0 dan 1 [17]. Proses ini dilakukan dengan menggunakan rumus berikut:

$$Xscaled = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

Di mana :

X : adalah nilai asli dari fitur
X_{min} : adalah nilai minimum dari fitur tersebut
X_{max} : adalah nilai maksimum dari fitur tersebut

4. Modelling

Modelling Algoritma terbagi menjadi dua, yaitu algoritma DBSCAN dan K-Means,yang memiliki proses sebagai berikut:

a. Proses DBSCAN Clustering:

Langkah awal adalah memilih parameter Epsilon (ε) dan MinPts dengan menentukan nilai untuk parameter epsilon (ε), yang merupakan radius lingkungan, serta MinPts yang merupakan jumlah minimum titik dalam radius ε agar titik tersebut dapat dianggap sebagai titik inti. Kemudian, masukkan nilai parameter yang telah dipilih ke dalam model DBSCAN dan inisialisasi model dengan parameter yang telah ditentukan [6].

b. Proses K-Means Clustering

Langkah awal adalah menentukan nilai K menggunakan Elbow Method, gunakan elbow method untuk menentukan jumlah cluster (K) yang optimal dengan memplotkan nilai inersia terhadap nilai K yang berbeda dan menemukan titik siku. Selanjutnya, masukkan nilai K yang optimal ke dalam model K-Means, pilih secara acak titik-titik pusat awal, dan jalankan algoritma K-Means untuk mengelompokkan data. Setelah menjalankan model K-Means dan membentuk klaster, visualisasikan hasil pengelompokan.

5. Perbandingan Hasil Metode

Setelah hasil pengelompokan dari kedua metode tersedia, bandingkan hasilnya dengan menggunakan beberapa metrik yaitu dengan, memvisualisasikan waktu setiap klaster sebagai Heatmap untuk memahami distribusi waktu di dalam klaster, menghitung indeks Davies-Bolden untuk menilai kualitas pengelompokan dengan nilai yang lebih kecil yang mengindikasikan klaster yang lebih baik [5], dan menghitung Silhoutte Score untuk menilai seberapa baik titik-titik data dikelompokkan di dalam klaster yang sama dan seberapa baik perbedaannya dengan klaster lain, dengan nilai mulai dari -1 (klaster buruk) hingga 1 (klaster baik) [3].

6. Analisis

Langkah terakhir adalah analisa hasil, disini akan menjelaskan hasil dari perbandingan kedua algoritma, maupun hasil analisa tiap cluster ataupun segmentasi tiap pelanggan yang dihasilkan.

4. HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini adalah data transaksi pelanggan Transjakarta dari Transjakarta General Transit Feed Specification pada bulan April 2023, terdiri dari 189.501 baris dan 22 kolom. Adapun pengolahan dataset nya sebagai berikut:

4.1 Data Cleaning

Pada proses ini dilakukan dengan cara df.dropna untuk menghapus semua baris dalam DataFrame yang memiliki satu atau lebih nilai null (missing values),sehingga data yang tersisa dan akan digunakan menjadi sebanyak 153.963 baris dan 22 kolom. Adapun attribute kolom ditunjukkan pada tabel 1.

Tabel 1. Kolom Dataset

Atribute	Keterangan
TransID	Kode unik pada setiap transaksi
PayCardID	Nomor identifikasi unik untuk kartu pembayaran yang digunakan dalam transaksi.
PayCardBank	Nama bank penerbit kartu pembayaran.
PayCardName	Nama pemegang kartu pembayaran.
PayCardSex	Jenis kelamin pemegang kartu pembayaran.
PayCardBirthDate	Tanggal lahir pemegang kartu pembayaran.
CorridorID	Nomor identifikasi unik untuk koridor (jalur) Transjakarta.

Atribute	Keterangan
CorridorName	Nama koridor (jalur) Transjakarta yang digunakan.
Direction	Arah perjalanan pada koridor yang digunakan.
TapInStops	Kode atau ID halte tempat penumpang melakukan tap-in (masuk).
TapInStopsName	Nama halte tempat penumpang melakukan tap-in (masuk).
TapInStopsLat	Koordinat lintang (latitude) dari halte tempat penumpang melakukan tap-in.
TapInStopsLon	Koordinat bujur (longitude) dari halte tempat penumpang melakukan tap-in.
StopStartSeq	Urutan atau nomor seq halte awal dalam perjalanan.
TapInTime	Waktu saat penumpang melakukan tap-in (masuk).
TapOutStops	Kode atau ID halte tempat penumpang melakukan tap-out (keluar).
TapOutStopsName	Nama halte tempat penumpang melakukan tap-out (keluar).
TapOutStopsLat	Koordinat lintang (latitude) dari halte tempat penumpang melakukan tap-out.
TapOutStopsLon	Koordinat bujur (longitude) dari halte tempat penumpang melakukan tap-out.
StopsEndSeq	Urutan atau nomor seq halte akhir dalam perjalanan.
TapOutTime	Waktu saat penumpang melakukan tap-out (keluar).
PayAmount	Jumlah biaya yang dibayarkan untuk perjalanan tersebut.

4.2 Feature Engineering

Feature Engineering adalah proses mengubah data mentah menjadi format yang lebih ringkas dan akurat untuk pemilihan fitur yang akan digunakan dalam model machine learning [18]. Adapun kolom yang akan digunakan untuk pembuatan fitur input seperti tabel 2.

Tabel 2. Kolom yang Digunakan Untuk Feature Engineering

	PayCard Bank	PayCardName	tapInTime	tapOutTime	PayAmount	TapInHour	tapInHour	tapDay
0	dki	Dr.Janet Nashruddin.M.Ak	2023-04-03 06:53:02	2023-04-03 07:13:28	3500	6	7	Senin
1	dki	Dr.Janet Nashruddin.M.Ak	2023-04-04 06:15:51	2023-04-04 06:55:34	3500	6	6	Selasa
2	dki	Dr.Janet Nashruddin.M.Ak	2023-04-04 17:20:19	2023-04-04 18:15:12	3500	17	18	Selasa
3	dki	Dr.Janet Nashruddin.M.Ak	2023-04-05 06:11:12	2023-04-05 06:57:01	3500	6	6	Rabu
4	dki	Dr.Janet Nashruddin.M.Ak	2023-04-05 17:10:40	2023-04-05 18:54:44	3500	17	18	Rabu
...	...	...	...	...	...	...	...	...
153958	Emoney	R.A Malika Samosir S.Kom	2023-04-02 13:00:49	2023-04-02 14:54:30	3500	13	14	Minggu
153959	dki	Amalia Gunarto	2023-04-04 17:10:40	2023-04-04 18:38:31	3500	17	18	Selasa
153960	brizzi	R.A Lintang Wibisono	2023-04-05 10:37:19	2023-04-05 12:20:28	3500	10	12	Rabu
153961	online	Yunita Sitompul	2023-04-14 13:34:20	2023-04-14 16:05:12	3500	13	16	Jumat
153962	dki	Omar Rahayu	2023-04-29 09:14:32	2023-04-29 11:53:39	3500	9	11	Sabtu

153963 rows × 9 columns

Pada penelitian ini Fitur yang dibuat yaitu Recency, Freqeuncy dan Monetery sebagai input untuk model machine learning, adapun proses pembuatan inputnya sebagai berikut:

1. Recency
   Recency diukur dengan menghitung jumlah hari dari transaksi terakhir pelanggan hingga tanggal terakhir dalam dataset. Pertama, ambil tanggal terakhir dari kolom tapOutTime menggunakan max(). Lalu, hitung selisih hari antara tanggal maksimum dan tanggal transaksi terakhir, dan simpan hasilnya dalam kolom baru bernama Recency. Ini memperbarui dataframe clv untuk merefleksikan seberapa baru transaksi terakhir pelanggan.

2. Frequency
   Frequency untuk setiap pelanggan dalam DataFrame dihitung berdasarkan jumlah transaksi yang dilakukan dengan kartu pembayaran (payCardName). Pertama, gunakan value_counts() untuk menghitung frekuensi setiap payCardName, lalu konversi hasilnya menjadi DataFrame baru dengan kolom payCardName dan Frequency.

3. Monetery
   Nilai moneter untuk setiap pelanggan dihitung berdasarkan total uang yang dibelanjakan dengan kartu pembayaran tertentu (payCardName). Pertama, data dikelompokkan berdasarkan payCardName menggunakan groupby(), lalu total payAmount untuk setiap grup dihitung dengan sum(). Hasilnya diubah menjadi DataFrame baru dengan kolom payCardName dan total payAmount.

4. Normalisasi Data RFM
   Normalisasi beberapa kolom dalam dataframe data dilakukan menggunakan teknik Min-Max Scaling. Pertama, tentukan kolom yang akan diskalakan dan simpan dalam daftar columns_to_scale, yang mencakup 'Recency', 'Frequency', dan 'Value'.Adapun kolom setelah dilakukan pembuatan fitur input dan normalisasi ditunjukkan pada tabel 3.

Tabel 3. Dataframe yang Berisi Kolom Input RFM

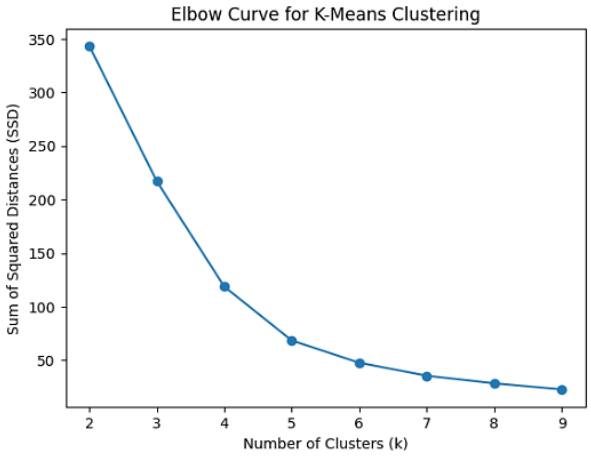
	payCardName	Recency	Frequency	Value
0	Dr. Janet Nashiruddin, M.Ak	0.931034	0.492958	0.161538
1	Balamantri Rahayu	0.931034	0.507042	0.000000
2	Dian Mustofa	0.931034	0.478873	0.897436
3	Dasa Prakarsa, S.I.Kom	0.931034	0.422535	0.000000
4	Elvina Hasanah	0.931034	0.464789	0.000000
...	...	...	...	...

9449Rows × 4 columns

4.3 Pembentukan Model

Pada tahapan ini, model penelitian dibuat menggunakan Algoritma K-Means dan DBSCAN:

1. K-Means



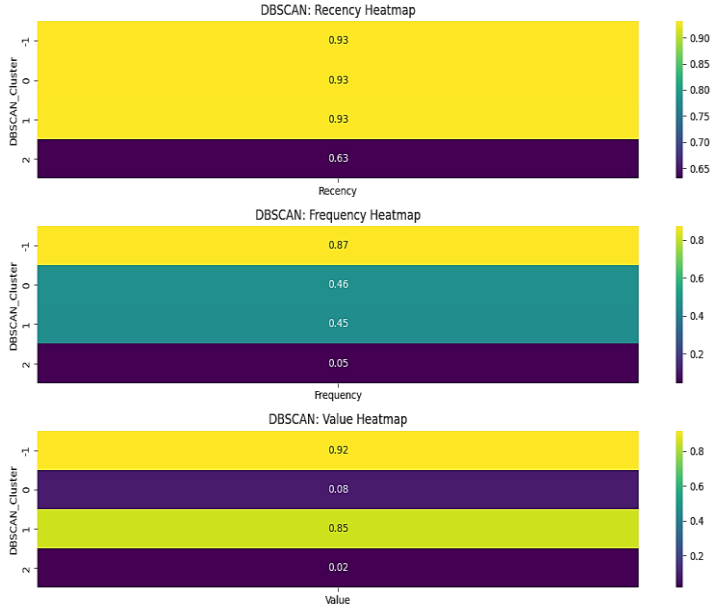
Gambar 2. Curve Elbow Method K-Means

Pada pembuatan algoritma K-Means, langkah awal adalah menggunakan elbow method untuk menentukan jumlah cluster. Berdasarkan grafik elbow method diatas, jumlah cluster terbaik adalah K=4. Selanjutnya, inisialisasi n_cluster atau K dilakukan dengan nilai 4, menggunakan data input yang berisi Recency, Frequency, dan Value atau Monetary.

2. DBSCAN
   Pada pembuatan algoritma DBSCAN, nilai klaster atau k ditentukan dengan mencari parameter terbaik (eps dan min_samples) berdasarkan silhouette score. Langkah berikutnya adalah mencari parameter terbaik menggunakan silhouette score. Nilai eps (jarak maksimum antara dua sampel untuk dianggap bagian dari satu sama lain) dan min_samples (jumlah minimum sampel dalam radius eps agar titik dianggap inti) ditentukan. Sebuah loop bersarang mencoba setiap kombinasi nilai eps dan min_samples. Untuk setiap kombinasi, DBSCAN diterapkan pada data, dan jumlah cluster yang dihasilkan (tidak termasuk noise) dihitung. Jika lebih dari satu cluster terbentuk, silhouette score dihitung untuk mengevaluasi kualitas clustering. Parameter dengan silhouette score tertinggi dicatat sebagai yang terbaik. Setelah iterasi, parameter terbaik yang ditemukan adalah eps 0.2 dan min_samples 5.

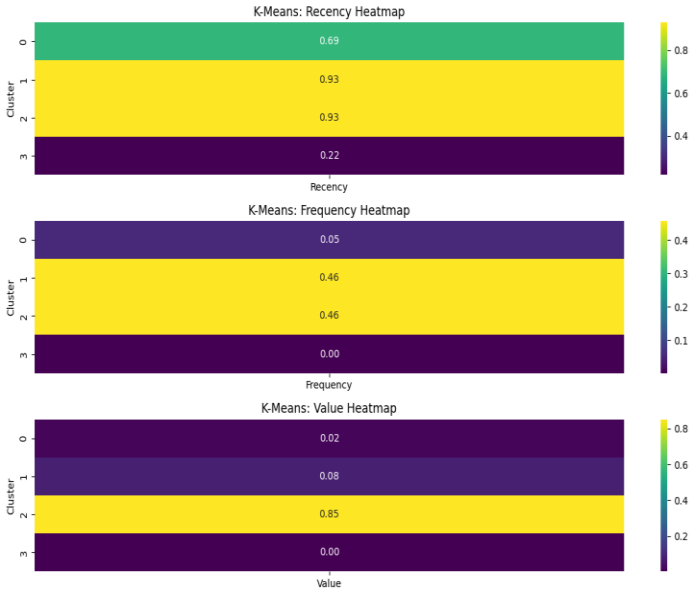
4.4 Perbandingan Hasil Metode

1. Heatmap Time
   a. DBSCAN
      Gambar 3 terdiri dari tiga heatmap yang menggambarkan distribusi nilai rata-rata dari fitur Recency, Frequency, dan Value untuk setiap cluster yang dihasilkan oleh algoritma DBSCAN. Setiap heatmap memiliki sumbu vertikal yang menunjukkan cluster yang teridentifikasi oleh DBSCAN (diberi label -1, 0, 1, dan 2) dan sumbu horizontal yang menunjukkan skala dari fitur yang relevan. Warna dalam heatmap mengindikasikan nilai rata-rata dari masing-masing fitur untuk setiap cluster, dengan skala warna dari ungu (nilai rendah) ke kuning (nilai tinggi).



Gambar 3. Heatmap Time DBSCAN

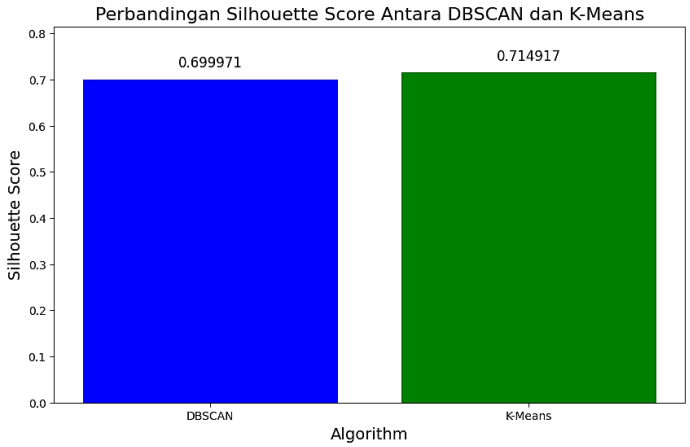
   b. K-Means
      Sedangkan gambar 4 menampilkan tiga Heatmap yang menunjukkan distribusi fitur Recency, Frequency, dan Value untuk setiap klaster yang dihasilkan oleh algoritme K-Means. Setiap Heatmap memiliki sumbu vertikal yang menunjukkan cluster yang diidentifikasi oleh algoritme K-Means (diberi nomor 0, 1, 2, dan 3) dan sumbu horizontal yang menunjukkan metrik fitur yang sesuai. Warna pada Heatmap menunjukkan nilai rata-rata setiap atribut untuk setiap cluster, dan skala warna berkisar dari ungu (nilai rendah) hingga kuning (nilai tinggi).



Gambar 4. Heatmap Time K-Means

2. Silhoutte Score

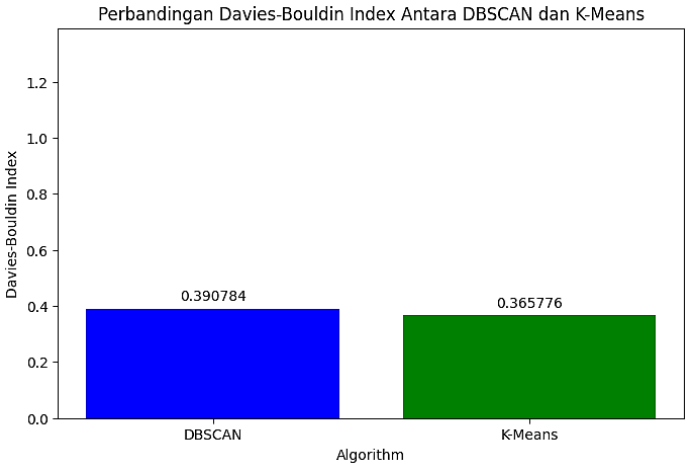
Gambar 5 merupakan perbandingan silhouette score antara dua algoritma clustering DBSCAN dan K-Means. Silhouette score adalah metrik evaluasi yang mengukur sejauh mana objek dalam suatu klaster berbeda dari klaster lainnya. Nilai silhouette score berkisar antara -1 hingga 1, dengan nilai positif menunjukkan bahwa objek lebih cocok dalam klasternya daripada dalam klaster lainnya. DBSCAN memiliki silhouette score sekitar 0.699971 (ditunjukkan oleh batang biru). K-Means memiliki silhouette score yang sedikit lebih tinggi, yaitu sekitar 0.714917 (ditunjukkan oleh batang hijau).



Gambar 5. Grafik Silhoutte Score

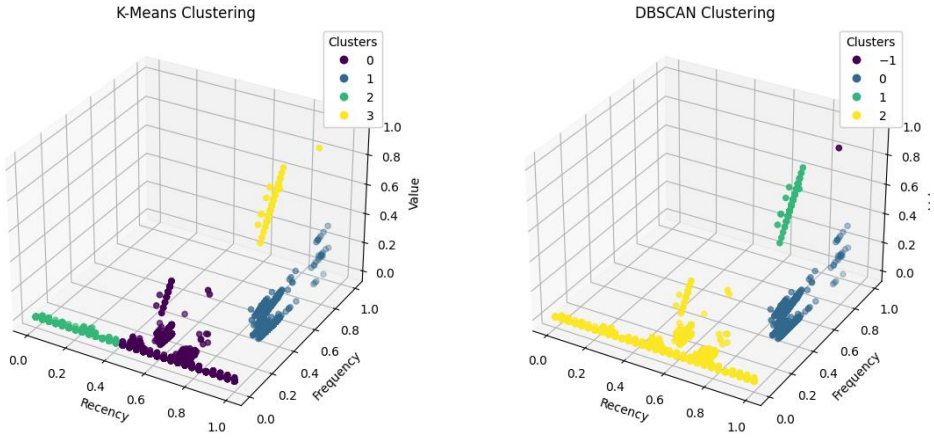
3. Davies Boulden Index

Gambar 6 merupakan perbandingan Davies-Bouldin Index antara dua algoritma clustering DBSCAN dan K-Means. Davies-Bouldin Index adalah metrik evaluasi yang mengukur sejauh mana objek dalam suatu klaster berbeda dari klaster lainnya. Nilai Davies-Bouldin Index berkisar antara 0 hingga lebih tinggi, dengan nilai lebih rendah menunjukkan clustering yang lebih baik. DBSCAN memiliki Davies-Bouldin Index sekitar 0.390784 (ditunjukkan oleh batang biru). K-Means memiliki Davies-Bouldin Index yang sedikit lebih rendah, yaitu sekitar 0.365776 (ditunjukkan oleh batang hijau).



Gambar 6. Grafik Davies Bolden Index

4. Perbandingan Hasil Visualisasi clustering
   Gambar 7 merupakan plot yang mewakili titik data yang dikelompokkan menggunakan algoritma clustering yang berbeda. Di sebelah kiri adalah plot yang berjudul "K-Means Clustering," dan di sebelah kanan adalah plot yang berjudul "DBSCAN Clustering." Kedua plot memiliki sumbu yang diberi label 'Recency' (kebaruan), 'Frequency' (frekuensi), dan 'Monetary' (nilai uang dalam skala logaritmik).



Gambar 7. Grafik 3D Visualisasi Clustering

5. Perbandingan Hasil Segmentasi berdasarkan Recency Frequency dan Monetery di setiap Cluster dengan Algoritma DBSCAN dan K-Means

Tabel 4. Hasil Segmentasi Recency

Segmentasi Pelanggan Berdasarkan Cluster Recency							
Cluster	DBSCAN			Cluster	K-Means		
	Mean	Min	Max		Mean	Min	Max
-1	0.93	0.93	0.93	0	0.68	0.45	1.00
0	0.93	0.86	0.93	1	0.93	0.86	0.93
1	0.93	0.90	0.93	2	0.21	0.00	0.41
2	0.63	0.00	1.00	3	0.93	0.90	0.93

Tabel 5. Hasil Segmentasi Frequency

Segmentasi Pelanggan Berdasarkan Cluster Frequency							
Cluster	DBSCAN			Cluster	K-Means		
	Mean	Min	Max		Mean	Min	Max
-1	0.87	0.87	0.87	0	0.68	0.45	1.00
0	0.46	0.32	1.00	1	0.93	0.86	0.93
1	0.45	0.34	0.54	2	0.21	0.00	0.41
2	0.04	0.00	0.30	3	0.93	0.90	0.93

Tabel 6. Hasil Segmentasi Monetery

Segmentasi Pelanggan Berdasarkan Cluster Monetery							
Cluster	DBSCAN			Cluster	K-Means		
	Mean	Min	Max		Mean	Min	Max
-1	0.92	0.92	0.92	0	0.02	0.00	0.36
0	0.08	0.00	0.32	1	0.08	0.00	0.32
1	0.85	0.64	1.00	2	0.004	0.00	0.03
2	0.02	0.00	0.36	3	0.85	0.64	1.00

Pada tabel 4,5 dan 6 merupakan perbandingan hasil segmentasi pelanggan berdasarkan Recency, Frequency, dan Monetary, mean, min, dan max memiliki peran penting dalam memahami distribusi data dalam setiap cluster. Mean (rata-rata) menunjukkan nilai rata-rata dari semua data dalam cluster tersebut, memberikan gambaran umum tentang perilaku pelanggan di cluster tersebut. Min (minimum) menunjukkan nilai terkecil, yang mengindikasikan batas bawah dari perilaku pelanggan dalam cluster, seperti jumlah transaksi terkecil atau pengeluaran terkecil. Max (maksimum) menunjukkan nilai terbesar, mengindikasikan batas atas dari perilaku pelanggan, seperti jumlah transaksi terbesar atau pengeluaran terbesar. Dengan memahami mean, min, dan max dari setiap cluster yang dihasilkan oleh algoritma DBSCAN dan K-Means, kita dapat memperoleh wawasan mendalam tentang pola perilaku pelanggan, termasuk variasi dalam penggunaan layanan dan pengeluaran mereka.

4.5 Analisis

Pada Bagian ini merupakan hasil analisis dari perbandingan algoritma DBSCAN dan K-Means untuk segmentasi pelanggan transjakarta dengan metode RFM sebagai input, dengan menghasilkan tiga perbandingan, yang pertama adalah Heatmap time, bisa dilihat pada tabel 7.

Tabel 7. Heatmap Time Tiap Cluster

Heatmap Time Tiap Cluster Satuan Detik							
Cluster	DBSCAN			Cluster	K-Means		
	Frequency	Recency	Monetery		Frequency	Recency	Monetery
-1	0.93	0.87	0.92	0	0.68	0.05	0.02
0	0.93	0.46	0.08	1	0.93	0.46	0.08
1	0.93	0.45	0.85	2	0.20	0.01	0.01
2	0.63	0.05	0.02	3	0.93	0.46	0.85
Rata-rata	0.855	0.457	0.467	Rata-Rata	0.685	0.245	0.24

Tabel 7 merupakan Heatmap Time dari masing-masing algoritma yaitu DBSCAN dan K-Means, dimana ini merupakan waktu yang dibutuhkan untuk proses di setiap cluster, dari data di atas, terlihat bahwa waktu pemrosesan untuk Cluster -1 (Ouliers) pada DBSCAN membutuhkan waktu yang cukup signifikan untuk semua fitur RFM. Cluster 0 dan 1 pada DBSCAN memiliki waktu pemrosesan yang serupa untuk Frequency, namun berbeda untuk Recency dan Monetery. Cluster 2 pada DBSCAN menunjukkan waktu pemrosesan yang lebih cepat untuk semua fitur, menunjukkan kemungkinan data yang lebih sederhana atau lebih sedikit, sedangkan untuk K-Means, Cluster 1 dan 3 memiliki waktu pemrosesan yang tinggi untuk Frequency dan Recency, sedangkan Cluster 0 dan 2 memiliki waktu yang lebih cepat. Hal ini menunjukkan variasi dalam kompleksitas atau jumlah data dalam setiap cluster, adapun langkah lain sebagai perbandingan algoritma dengan evaluasi dua parameter seperti tabel 8.

Tabel 8. Perbandingan Silhoutte Score Dengan DBI

Parameter Evaluasi	DBSCAN	K-Means
Silhoutte Score	0.699971	0.714917
Davies-Bouldin Index	0.390784	0.365776

Berdasarkan tabel diatas menunjukkan hasil evaluasi parameter dari dua algoritma clustering, DBSCAN dan K-Means, menggunakan dua parameter evaluasi utama: Silhouette Score dan Davies-Bouldin Index. Untuk Silhouette Score, K-Means memiliki nilai yang sedikit lebih tinggi (0.714917) dibandingkan dengan DBSCAN (0.699971), menunjukkan bahwa clustering yang dihasilkan oleh K-Means memiliki kualitas yang sedikit lebih baik dalam hal keseragaman internal dan perbedaan antar cluster. Untuk Davies-Bouldin Index, K-Means juga unggul dengan nilai yang lebih rendah (0.365776) dibandingkan DBSCAN (0.390784), menunjukkan bahwa cluster yang dihasilkan oleh K-Means lebih dekat satu sama lain dan lebih terpisah dibandingkan dengan cluster yang dihasilkan oleh DBSCAN. Secara keseluruhan, berdasarkan kedua parameter evaluasi ini, K-Means menunjukkan performa yang lebih baik dalam segmentasi pelanggan menggunakan data RFM, selain dari itu

adapun hasil analisa untuk segmentasi pelanggan berdasarkan hasil dari setiap cluster berdasarkan Recency, Frequency dan Monetery.

Tabel 9. Jumlah Pelanggan di Setiap Cluster

Segmentasi Berdasarkan Jumlah Pelanggan di Setiap Cluster			
Cluster	DBSCAN	Cluster	K-Means
	Jumlah		Jumlah
-1	2	0	4943
0	3722	1	3722
1	171	2	611
2	5554	3	173
Total	9449	Total	9449

Tabel 9 merupakan hasil segmentasi berdasarkan jumlah pelanggan di setiap klaster, algoritma DBSCAN memiliku 1 klaster ouliers dan 3 klaster dengan jumlah pelanggag terbanyak berada pada klaster 2, sedangkan K-Means memiliki 4 klaster dengan jumlah terbanyak berada di klaster 0.

Tabel 10. Analisa Recency

Analisa Segmentasi Pelanggan Berdasarkan Cluster Recency							
Cluster	DBSCAN			Cluster	K-Means		
	Mean	Min	Max		Mean	Min	Max
-1	0.93	0.93	0.93	0	0.68	0.45	1.00
0	0.93	0.86	0.93	1	0.93	0.86	0.93
1	0.93	0.90	0.93	2	0.21	0.00	0.41
2	0.63	0.00	1.00	3	0.93	0.90	0.93

Pada tabel 10 merupakan tabel Recency dari masing-masing algoritma, terdiri dari Mean, Min, dan Max. Nilai dari masing-masing keterangan berkisar antara 0 hingga 1, jika angka mendekati 1 berarti pelanggan memiliki waktu yang cuku lama menggunakan Transjakarta pada klaster tersebut, dari kedua metode algoritma di dapatkan sebuah segmentasi pelanggan berdasarkan klaster. DBSCAN dan K-Means mengungkapkan perbedaan penting dalam perilaku interaksi pelanggan. DBSCAN mengidentifikasi outliers dengan recency sangat tinggi, menunjukkan pelanggan yang baru saja berinteraksi, sementara juga membentuk cluster dengan variasi interaksi terbaru. Sebaliknya, K-Means membedakan lebih jelas antara pelanggan yang baru-baru ini berinteraksi dan yang tidak, dengan cluster yang mencerminkan recency rendah dan tinggi.

Tabel 11. Analisa Frequency

Analisa Segmentasi Pelanggan Berdasarkan Cluster Frequency							
Cluster	DBSCAN			Cluster	K-Means		
	Mean	Min	Max		Mean	Min	Max
-1	0.87	0.87	0.87	0	0.68	0.45	1.00
0	0.46	0.32	1.00	1	0.93	0.86	0.93
1	0.45	0.34	0.54	2	0.21	0.00	0.41
2	0.04	0.00	0.30	3	0.93	0.90	0.93

Sedangkan pada tabel Frequency menunjukkan kedua algoritma DBSCAN dan K-Means menghasilkan klaster yang berbeda dalam hal Frequency transaksi pelanggan. DBSCAN mengidentifikasi outlier dengan frekuensi reaktif yang sangat tinggi (rata-rata 0,87) di cluster 1, serta membentuk cluster utama dengan frekuensi sedang hingga tinggi (rata-rata sekitar 0,45-0,46) dan satu cluster dengan frekuensi yang sangat rendah (rata-rata 0,04). Di sisi lain, K-Means membedakan dengan lebih jelas antara pelanggan dengan frekuensi reaktif yang sangat tinggi (rata-rata 0.93 di Cluster 1 dan 3) dan frekuensi reaktif yang rendah (rata-rata 0.21 di Cluster 2 dan 0.68 di Cluster 0). Hasil ini menunjukkan bahwa K-Means lebih baik dalam mengelompokkan pelanggan transjakarta berdasarkan perbedaan frekuensi yang lebih jelas, sementara DBSCAN cenderung mengelompokkan nasabah dengan frekuensi interaksi yang lebih cenderung mirip atau tidak terlalu bervariasi namun juga mampu mengidentifikasi outlier dengan frekuensi yang sangat tinggi, selanjutnya dihasilkan perbandingan Monetery ditunjukkan pada tabel 12.

Pada tabel Monetery adalah jumlah total uang yang telah dihabiskan oleh pelanggan selama periode tertentu. DBSCAN mengidentifikasi outlier dengan nilai moneter yang sangat tinggi di klaster -1 (rata-rata 0,92), pelanggan dengan nilai moneter menengah hingga tinggi di klaster 1 (rata-rata 0,85), dan nasabah dengan nilai moneter rendah di klaster 0 (rata-rata 0,08) dan klaster 2 (rata-rata 0,02). Sementara itu, K-Means membagi pelanggan menjadi empat kelompok dengan cara yang lebih terorganisir, klaster 0 dan 2 dengan nilai

moneter yang sangat rendah (rata-rata 0.02 dan 0.004), klaster 1 dengan nilai moneter rendah hingga sedang (rata-rata 0.08), dan klaster 3 dengan nilai moneter yang sangat tinggi (rata-rata 0.85). Hal ini menunjukkan bahwa K-Means lebih efektif dalam membedakan nasabah berdasarkan perbedaan nilai moneter, sedangkan DBSCAN cenderung mengelompokkan nasabah dengan pola pengeluaran yang lebih konsisten.

Tabel 12. Analisa Monetery

Analisa Segmentasi Pelanggan Berdasarkan Cluster Monetery							
Cluster	DBSCAN			Cluster	K-Means		
	Mean	Min	Max		Mean	Min	Max
-1	0.92	0.92	0.92	0	0.02	0.00	0.36
0	0.08	0.00	0.32	1	0.08	0.00	0.32
1	0.85	0.64	1.00	2	0.004	0.00	0.03
2	0.02	0.00	0.36	3	0.85	0.64	1.00

Berdasarkan analisis dari ketiga fitur yaitu Recency, Frequency dan Monetery, maka pelanggan dapat dikategorikan ke beberapa segmen berdasarkan klaster , adapun hasil dari setiap analisa segmentasi yang ditunjukkan pada tabel 13 dan 14.

Tabel 13. Karakteristik Segmentasi Pelanggan Algoritma DBSCAN

Klaster	Tipe Pelanggan	Karakteristik RFM
-1	Ouliers	Pengguna dengan interaksi terbaru yang sangat tinggi, frekuensi tinggi, dan jumlah uang yang tinggi.
0	Low Frequency Users	Pengguna dengan interaksi terbaru yang cukup lama (Recency tinggi), frekuensi rendah, dan nilai moneter rendah.
1	Medium Value Users	Pengguna dengan interaksi terbaru yang cukup lama (Recency tinggi), frekuensi sedang, dan nilai moneter tinggi.
2	Infrequent High Spenders	Pengguna dengan interaksi terbaru yang rendah (Recency rendah), frekuensi sangat rendah, tetapi nilai moneter tinggi.

Tabel 14. Karakteristik Segmentasi Pelanggan

Klaster	Tipe Pelanggan	Karekteristik RFM
0	Moderate Users	Pengguna dengan interaksi terbaru yang moderat (Recency sedang), frekuensi tinggi, dan nilai moneter sedang.
1	Low Spend Frequent Users	Pengguna dengan interaksi terbaru yang sangat tinggi (Recency rendah), frekuensi penggunaan yang sangat tinggi, tetapi nilai moneter rendah.
2	Low Engagement Users	Pengguna dengan interaksi terbaru yang rendah (Recency sangat tinggi), frekuensi penggunaan yang rendah, dan nilai moneter yang sangat rendah.
3	High Value Users	Pengguna dengan interaksi terbaru yang tinggi (Recency rendah), frekuensi penggunaan yang sangat tinggi, dan nilai moneter yang tinggi.

5. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan dari kedua algoritma yaitu DBSCAN dan K-Means dalam Segmentasi Pelanggan dengan Metode Recency, Frequency, dan Monetery (RFM). Berdasarkan analisis, K-Means menunjukkan hasil yang lebih unggul dibandingkan DBSCAN dalam hal kualitas clustering. K-Means memiliki Silhouette Score sebesar 0.714917, lebih tinggi dibandingkan DBSCAN yang memiliki nilai 0.699971, mengindikasikan bahwa clustering K-Means memiliki keseragaman internal yang lebih baik dan perbedaan antar cluster yang lebih jelas. Davies-Bouldin Index (DBI) juga menunjukkan bahwa K-Means lebih efektif dengan nilai 0.365776 dibandingkan DBSCAN yang memiliki nilai 0.390784, menandakan bahwa cluster yang dihasilkan oleh K-Means lebih kompak dan terpisah dengan baik. Heatmap time analysis mengungkapkan bahwa K-Means memiliki waktu pemrosesan yang lebih cepat dan konsisten di berbagai cluster, sementara DBSCAN menunjukkan variasi waktu pemrosesan yang lebih signifikan, terutama dalam mengidentifikasi outliers. Dalam penerapan pada segmentasi pelanggan Transjakarta, K-Means lebih efektif dalam membedakan pelanggan berdasarkan recency, frequency, dan monetary value.

Saran yang dapat penulis sampaikan yaitu, untuk penelitian selanjutnya, disarankan untuk mengeksplorasi penggunaan algoritma clustering lainnya seperti Hierarchical Clustering atau Gaussian Mixture Models (GMM) untuk segmentasi pelanggan Transjakarta. Selain itu, penelitian dapat diperluas dengan menggabungkan data tambahan seperti feedback pelanggan, pola perjalanan harian, dan data demografis untuk mendapatkan pemahaman yang lebih komprehensif tentang perilaku dan kebutuhan

pelanggan. Menggunakan data dalam periode waktu yang lebih panjang juga dapat memberikan wawasan tentang tren dan perubahan perilaku pelanggan dari waktu ke waktu. Terakhir, implementasi metode machine learning yang lebih canggih, seperti teknik deep learning, dapat dieksplorasi untuk meningkatkan akurasi dan efektivitas segmentasi pelanggan.

REFERENSI

[1] N. Anche Natalia Fransisca and N. A. Natali F, "Analisis Dampak Service Quality terhadap Customer Trust, Customer Satisfaction dan Customer Loyalty Bus Transjakarta," *Ecodemica: Jurnal Ekonomi, Manajemen dan Bisnis*, vol. 7, no. 1, 2023, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ecodemicaWebsite:https://ejournal.bsi.ac.id/ejurnal/index.php/ecodemica

[2] D. A. Susanto, "Segmentasi Pelanggan Menggunakan Algoritma K-Means dan Model RFM (Studi Kasus: Industri Pengolahan Limbah Rumah Tangga)," *Ilmiah KOMPUTASI, Volume 21 No : 2, Juni 2022, p-ISSN 1412-9434/e-ISSN 2549-7227*, 2022.

[3] A. Satriawan, R. Andreswari, and O. N. Pratiwi, "SEGMENTASI PELANGGAN TELKOMSEL MENGGUNAKAN METODE CLUSTERING DENGAN RFM MODEL DAN ALGORITMA K-MEANS TELKOMSEL CUSTOMER SEGMENTATION USING CLUSTERING METHOD WITH RFM MODEL AND K-MEANS ALGORITHM," 2021.

[4] C. R. I. Pata, Statiswaty, and N. Ransi, "SEGMENTASI PEMETAAN PELANGGAN POTENSIAL MENGGUNAKAN ALGORITMA DBSCANDENGAN RFM MODEL BERBASIS WEB," *Vol. 1 No. 2 (2023): Volume 1 Nomor 2 Tahun 2023* , vol. 1, 2023, Accessed: Oct. 21, 2023. [Online]. Available: http://animator.uho.ac.id/index.php/journal

[5] S. Ika Murpratiwi, I. Gusti Agung Indrawan, and A. Aranta, "ANALISIS PEMILIHAN CLUSTER OPTIMAL DALAM SEGMENTASI PELANGGAN TOKO RETAIL," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 18, no. 2, 2021.

[6] K. E. Setiawan and A. Kurniawan, "PENGELOMPOKAN RUMAH SAKIT DI JAKARTA MENGGUNAKAN MODEL DBSCAN, GAUSSIAN MIXTURE, DAN HIERARCHICAL CLUSTERING," *Jurnal Informatika Terpadu*, vol. 9, no. 2, pp. 149–156, 2023, [Online]. Available: https://journal.nurulfikri.ac.id/index.php/JIT

[7] R. C. P. Ipa, Statiswaty, and R. Natalis, "SEGMENTASI PEMETAAN PELANGGAN POTENSIAL MENGGUNAKAN ALGORITMA DBSCANDENGAN RFM MODEL BERBASIS WEB," *ANIMATOR*, vol. 1, no. 2, pp. 63–71, 2023, Accessed: Nov. 03, 2023. [Online]. Available: http://animator.uho.ac.id/index.php/journal

[8] R. A. Farissa, R. Mayasari, and Y. Umaidah, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient," 2021. doi: 10.30871/jaic.v5i1.3237.

[9] T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D. S. Man, and T. G. Le, "An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry," *Business Systems Research*, vol. 14, no. 1, pp. 26–53, Sep. 2023, doi: 10.2478/bsrj-2023-0002.

[10] H. Imam, P. Yenik, and A. M. Muhamad, "Pengaruh Pembangunan Infrastruktur Transportasi Berkelanjutan terhadap Mobilitas dan Lingkungan di Kalimantan," *Jurnal Multidisiplin West Science*, vol. 02, no. 10, pp. 908–917, 2023, doi: https://doi.org/10.58812/jmws.v2i10.705.

[11] A. Levana and S. Hartono, "Analisis Perilaku Konsumen Terhadap Niat Menggunakan Transportasi Publik PT. Transportasi Jakarta," *LOCUS:Penelitian & Pengabdian*, vol. 2, no. 9, 2023, Accessed: Nov. 04, 2023. [Online]. Available: https://locus.rivierapublishing.id/index.php/jl

[12] M. Harahap, Y. Lubis, and Z. Situmorang, "Data Science bidang Pemasaran : Analisis Prilaku Pelanggan," *Data Sciences Indonesia (DSI)*, vol. 1, no. 1, pp. 21–32, Nov. 2021, doi: 10.47709/dsi.v1i1.1194.

[13] A. M. T I Sambi Ua *et al.*, "Penggunaan Bahasa Pemrograman Python Dalam Analisis Faktor Penyebab Kanker Paru-Paru Universitas Bina Nusantara," *Jurnal Publikasi Teknik Informatika (JUPTI)*, vol. 2, no. 2, 2023, doi: 10.55606/jupti.v2i2.1742.

[14] F. Defina, S. Alhamdani, A. A. Dianti, and Y. Azhar, "Kredit Menggunakan Metode K-Means Clustering," MEI, 2021. doi: 10.14421/jiska.2021.6.2.70-77.

[15] M. Jordy, A. Triayudi, and I. D. Sholihati, "Analisis Segmentasi Recency dan Customer Value Pada AVANA Indonesia Dengan Algoritma K-Means dan Model RFM (Recency, Frequency and Monetary)," *Journal of Information System Research (JOSH)*, vol. 4, no. 2, pp. 579–589, Jan. 2023, doi: 10.47065/josh.v4i2.2950.

[16] N. R. Dwitya and W. Istiono, "Consumer Segmentation of Emina Cosmetics Optimal and Relevant Approach of RFM+Lifetime Analysis," *SAGA: Journal of Technology and Information Systems*, 2023, doi: 10.58905/SAGA.vol1i3.171.

[17] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 1, p. 32, Apr. 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[18] A. Alamsyah *et al.*, "Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm," *Scientific Journal of Informatics*, vol. 9, no. 2, pp. 189–196, Nov. 2022, doi: 10.15294/sji.v9i2.39437.