Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.2024118208
Vol. 11, No. 6, Desember 2024, hlm. 1243-1252 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

SEGMENTASI PELANGGAN MAJALAH PADA SITUS WEB E-COMMERCE DENGAN K-MEANS++ DAN METODE RFM

Andrew Lomaksan Manuel Tampubolon[*1], Thio Marta Elisa Yuridis Butar Butar[2], Siti Rochimah[3]

[1,2,3]Institut Teknologi Sepuluh Nopember, Surabaya
Email: [1]6025231025@student.its.ac.id, [2]6025231009@student.its.ac.id, [3]siti@if.its.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 07 Desember 2023, diterima untuk diterbitkan: 20 November 2024)

Abstrak

Segmentasi pelanggan merupakan salah satu metode yang dapat diterapkan untuk memaksimalkan peluang bisnis. Hal tersebut dapat membantu bisnis agar tetap kompetitif dalam persaingan pasar. Penerapan *Artificial Intelligence* (AI) dapat membantu dalam memberikan pemahaman kepada pelaku bisnis tentang segmentasi pelanggan berdasarkan riwayat transaksi. Penelitian ini menerapkan metode *Recency*, *Frequency*, and *Monetary* (RFM) yang dipadukan dengan algoritma *clustering* K-Means++ untuk melakukan segmentasi pelanggan. *Silhouette score* menjadi indikator pemilihan nilai k yang paling optimal dalam menentukan jumlah *cluster*. Kerangka kerja CRISP-DM yang digunakan dalam makalah ini juga membantu mempertahankan proses analisis yang konsisten. Pendekatan statistik sederhana ddigunakan untuk mengklasifikasikan setiap fitur dalam RFM menjadi label *low*, *medium*, dan *high* dalam hal menangkap pola segmentasi pelanggan. Hasil eksperimen menunjukkan nilai k = 3 sebagai yang paling optimal berdasarkan nilai WSS sebesar 843,214747 dan *silhouette score* sebesar 0,638181. Eksperimen juga menunjukkan bahwa *cluster* 0 memiliki nilai RFM rata-rata sebesar 1,14 (*low*), 1,20 (*low*), dan 301.640 (*low*). *Cluster* 1 memiliki nilai RFM rata-rata sebesar 249,61 (*high*), 2,62 (*medium*), dan 799,934 (*medium*). *Cluster* 2 memiliki nilai RFM rata-rata sebesar 233,01 (*medium*), 6,41 (*high*), dan 2018,088 (*high*).

Kata kunci: *Segmentasi Pelanggan, E-Commerce, K-Means++, RFM, CRISP-DM.*

SEGMENTATION OF MAGAZINE SUBSCRIBERS ON E-COMMERCE WEBSITE USING K-MEANS++ AND RFM METHOD

Abstract

Customer segmentation is one method that can be applied to maximize business opportunities. It can help businesses remain competitive in the market competition. The application of Artificial Intelligence (AI) can assist in providing business stakeholders with an understanding of customer segmentation based on transaction history. This study applies the Recency, Frequency, and Monetary (RFM) method combined with the K-Means++ clustering algorithm for customer segmentation. The Silhouette score serves as an indicator for selecting the most optimal value of k to determine the number of clusters. The CRISP-DM framework used in this paper also helps maintain a consistent analysis process. A simple statistical approach is used to classify each RFM feature into low, medium, and high labels to capture customer segmentation patterns. Experimental results show that k = 3 is the most optimal value based on a WSS value of 843.214747 and a silhouette score of 0.638181. The experiments also indicate that Cluster 0 has average RFM values of 1.14 (low), 1.20 (low), and 301,640 (low). Cluster 1 has average RFM values of 249.61 (high), 2.62 (medium), and 799,934 (medium). Cluster 2 has average RFM values of 233.01 (medium), 6.41 (high), and 2018.088 (high).

Keywords: *Customer Segmentation, E-Commerce, K-Means++, RFM, CRISP-DM.*

1. PENDAHULUAN

E-commerce adalah sebuah platform digital yang memfasilitasi transaksi jual beli secara *online* antara penjual dan konsumen. Tujuan dari *e-commerce* adalah menjual produk dalam bentuk barang fisik dan layanan digital secara *online* (Ros'ario & Raimundo, 2021). Platform *e-commerce* turut menawarkan berbagai fitur dan kemudahan dalam berbelanja, seperti katalog produk dan keranjang belanja. Selain itu, platform *e-commerce* juga menyediakan berbagai opsi pembayaran bagi pelanggan. Berbagai fitur tambahan lainnya juga membantu pelanggan dalam menemukan, memilih, dan membeli produk atau layanan. Dalam hal struktur

bisnis, *e-commerce* mencakup beragam model pemasaran seperti pengecer *online*, layanan berlangganan (*subscription*), pasar digital, dan entitas B2B yang mengkhususkan diri dalam menjual produk atau layanan ke bisnis lain (Chen et al., 2022).

Agar tetap kompetitif, strategi bisnis yang efektif diperlukan untuk memaksimalkan peluang yang ada. Mengembangkan strategi tertentu yang seperti segmentasi pasar dan alokasi sumber daya yang praktis menjadi sangat penting demi meningkatkan angka penjualan produk. Selain itu, bisnis seperti *e-commerce* perlu untuk memperoleh keunggulan dalam kompetisi pasar (Dwivedi et al., 2021). Sebagai solusi, kecerdasan buatan dapat dimanfaatkan dalam bidang bisnis. Kecerdasan buatan dapat memberikan rekomendasi yang dipersonalisasi untuk meningkatkan pengalaman pelanggan dalam berbelanja (Verma et al., 2021; Haleem et al., 2022). Hal tersebut dapat membantu pelaku bisnis dalam mendapatkan pemahaman yang lebih dalam tentang perilaku konsumen (Fauzan & Davin, 2023; Verma et al., 2021). Melalui analisis data, kecerdasan buatan dapat membantu pelaku bisnis dalam mengidentifikasi tren aktivitas pengguna. Hal ini juga dapat membantu dalam memperoleh wawasan tentang preferensi dan karakteristik pelanggan.

Analisis dengan kecerdasan buatan menjadi penting dengan mendorong bisnis menerapkan keputusan berbasis data. Analsis data dapat membantu pengambilan keputusan yang lebih cepat dan mendapatkan wawasan tentang karakteristik pelanggan (Verma et al., 2021; Haleem et al., 2022). Proses tersebut mencakup strategi data, rekayasa data, tata kelola, manajemen perubahan, dan budaya (Haleem et al., 2022). Menurut laporan AI: *Built to Scale* dari Accenture, 84 persen eksekutif bisnis percaya bahwa kecerdasan buatan dapat membantu mencapai tujuan pertumbuhan mereka. Namun, 76 persen mengaku membutuhkan bantuan untuk meningkatkan kecerdasan buatan di seluruh bisnis mereka (Fauzan & Davin, 2023). Oleh karena itu, bisnis harus melakukan inovasi untuk mengembangkan strategi yang komprehensif. Kecerdasan buatan perlu untuk dimanfaatkan dengan baik agar tetap terdepan dan kompetitif di pasar.

Kecerdasan buatan diperlukan dalam mempersonalisasikan pelanggan agar tetap kompetitif di pasar. Personalisasi tersebut dapat dilakukan dengan melakukan segmentasi terhadap pelanggan, menemukan pola-pola dengan mengelompokkan pelanggan berdasarkan perilaku berbelanja. Segmentasi pelanggan sendiri dapat diartikan sebagai proses mengklasifikasikan pelanggan dengan karakteristik serupa ke dalam segmen serupa (Sarker, 2022). Penggunaan algoritma pengelompokan dapat membantu untuk lebih memahami karakteristik pelanggan, baik dalam hal demografi dan perilaku dinamis pelanggan (Verma et al., 2021). Mengingat data transaksi penjualan tidak memiliki label (kelas) data, maka algoritma *unsupervised machine learning* adalah algoritma yang sesuai untuk kasus tersebut. Beberapa algoritma *unsupervised machine learning* dapat membantu dalam menemukan titik pusat data yang terkait erat (Verma et al., 2021; Sarker, 2021). Salah satu algoritma pengelompokan seperti K-Means dapat dimanfaatkan untuk kasus tersebut. Selain itu algoritma tersebut juga cocok digunakan untuk kasus segmentasi pelanggan (Siebert et al., 2018). Metode ini juga dapat dikombinasikan dengan metode lain seperti Recency, Frequency, dan Monetary (RFM). Metode RFM dapat membantu mengevaluasi pelanggan berdasarkan perilaku belanja mereka (Christy et al., 2021). Metode ini juga dapat digunakan untuk meningkatkan analisis pelanggan dan prediksi pengalaman berbelanja (Khajvand et al., 2011). Informasi mengenai bagaimana pelanggan berbelanja dapat dikumpulkan dengan menganalisis riwayat transaksi pelanggan dengan memanfaatkan data transaksi yang tersimpan dalam *database*.

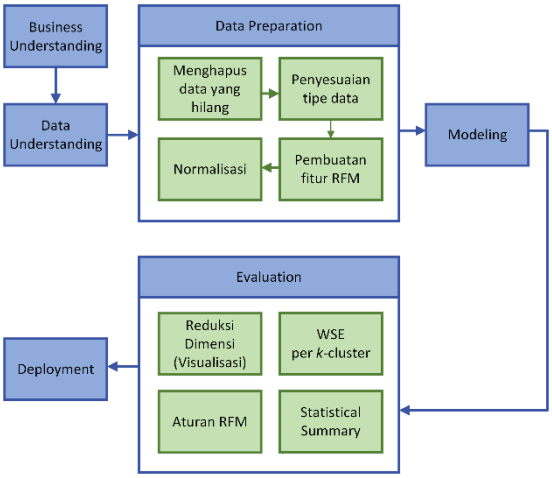
Dalam proses analisis, beberapa penelitian turut menerapkan kerangka kerja seperti *Standard Procedure for Data Mining* (CRISP-DM) untuk memastikan langkah kerja yang terstruktur. CRISP-DM merupakan kerangka kerja *data mining* yang populer serta menawarkan proses yang baik dalam hal *business understanding* hingga *deployment* (Martínez-Plumed et al., 2019). Kerangka kerja CRISP-DM juga digunakan bersamaan dengan analisis data dengan metode RFM dalam beberapa studi kasus yang berhubungan dengan penjualan. Beberapa penelitian diantaranya yaitu penerapan pada studi kasus *workshop* sepeda motor pada masa pandemi COVID-19 dalam hal mencari konsumen yang paling potensial (Mauritsius et al., 2023), manufaktur produk kayu dalam menemukan pelanggan paling menguntungkan dan tidak menguntungkan (Dzulhaq et al., 2019), dan peternakan unggas (Mirantika & Rijanto., 2023).

Untuk membantu *e-commerce* penjualan majalah dalam menemukan pola pembelian dari pelanggan, maka dilakukan penelitian yang melibatkan proses segmentasi pelanggan. Eksperimen yang dilakukan melibatkan data transaksi penjualan pada *e-commerce* penjualan majalah. Metode yang digunakan yaitu RFM yang dikombinasikan dengan algoritma *clustering* menggunakan K-Means++. Algoritma K-Means++ dipilih ketimbang K-Means karena menawarkan proses inisialisasi awal *centroid* yang lebih baik dan proses konvergensi yang lebih singkat (Bahmani et al., 2012). Metode RFM digunakan untuk membentuk fitur dalam bentuk jumlah pembelian (*recency*), rasio pembelian (*frequency*), dan total pembelian (*monetary*). Fitur tersebut nantinya akan di klasterisasi dengan algoritma yang umum seperti K-Means++ yang merupakan bentuk pengembangan dari algoritma K-Means *clustering*. Proses eksperimen juga melibatkan proses menemukan nilai

k (jumlah *cluster*) yang paling optimal berdasarkan *silhouette score*. Proses normalisasi juga diterapkan agar data yang diolah memiliki distribusi yang normal. Dalam memvisualisasikan pola data, dilakukan proses reduksi dimensi terlebih dahulu dengan *Principal Component Analysis* (PCA) dengan mereduksi data menjadi 2 dimensi. Hal tersebut berguna dalam mendapatkan perbandingan data secara visual sebelum dan sesudah proses klasterisasi. Hasil akhir dari eksperimen adalah pola-pola data secara *recency*, *frequency*, dan *monetary*, bagaimana pola untuk tiap klaster yang menggambarkan pola belanja untuk setiap pelanggan. Kerangka kerja CRISP-DM diterapkan dalam memastikan proses analisis dilakukan secara terstruktur dan runtut.

2. METODE PENELTIAN

Penelitian menggunakan metode CRISP-DM yang terdiri dari 5 langkah mulai dari *business understanding* hingga *evaluation*. Proses tersebut meliputi *business understanding*, *data understanding*, *modeling*, *evaluation*, and *deployment* (Wirth & Hipp, 2000). Namun proses *deployment* tidak dilakukan pada penelitian ini. Proses yang diterapkan pada penelitian terbatas pada *evaluasi*, yaitu dihasilkan *cluster* yang representatif untuk mensegmentai pelanggan. Selain itu pada beberapa proses CRISP-DM berisi sub-proses lain seperti *preparation* and *evaluation* data, seperti yang ditunjukkan pada gambar 1. Berikut adalah proses yang diterapkan dalam penelitian:



Gambar 1. Alur penelitian berdasarkan metode CRISP-DM

2.1. *Business Understanding*

Situs *e-commerce* penjualan majalah menyediakan berbagai produk majalah cetak dan majalah elektronik untuk segmen usia yang berbeda dengan segmen kelompok anak-anak, pria, dan wanita. Proses yang terjadi dalam *e-commerce* adalah serangkaian proses dari pelanggan mendaftar, memasukkan barang ke dalam keranjang, *checkout*, dan pembayaran hingga transaksi selesai, baik itu transaksi berhasil, kedaluarsa, atau dibatalkan. Sebagai aspek yang dapat menarik pelanggan, *e-commerce* melakukan promosi seperti diskon produk, bundling produk, bonus hadiah, dan voucher diskon. Namun, saat ini, promosi belum dilakukan sesuai target pelanggan. Hal ini diperlukan untuk merencanakan proses promosi kepada pelanggan potensial dengan tujuan tidak hanya meningkatkan pendapatan, Hal ini juga diharapkan untuk menawarkan berbagai promosi yang menarik pelanggan.

2.2. *Data Understanding*

Data dalam *e-commerce* terdiri dari transaksi pelanggan terhadap *e-magazine* dan produk cetak. Transaksi dimulai dari 1 Maret 2021 hingga 1 Maret 2023. Data diperoleh dari *database* melalui proses kueri SQL (*Structured Query Language*) *e-commerce* penjualan majalah dan data tidak tersedia secara publik. Kriteria data riwayat transaksi hanya transaksi berhasil yang terdiri dari data 5.847 baris dan 11 atribut seperti tanggal transaksi, tanggal pembuatan akun, kode transaksi, email, id pengguna, merek, harga berlangganan, jumlah item, harga satuan, jenis produk, dan jenis langganan. Secara umum, data yang tersedia mewakili komponen yang diperlukan dalam metode RFM. Semua atribut ini nantinya akan digunakan dalam proses pra pemrosesan data.

2.3. *Data Preparation*

Data yang disimpan dalam database *e-commerce* belum mengalami pemrosesan seperti pada data yang disimpan di *data warehouse*. Bentuk data seperti tipe data, format, dan ukuran, dan menghapus nilai *null* menjadi perhatian dalam proses *data mining* dalam data transaksi *e-commerce*. Dalam proses ini akan disesuaikan data dengan tipe data yang sesuai untuk diproses dengan benar. Kemudian, data rahasia akan disamarkan terlebih dahulu, seperti trans_id dan user_id. Atribut baru dibentuk dengan nama produk, yang merupakan kombinasi dari merek, durasi berlangganan, dan jenis produk (E-magazine atau cetak), seperti pada deskripsi pada tabel 1.

Tabel 1. Atribut Dataset

Nama Atribut	Deskripsi
trans_date	Tanggal transaksi
acc_created_date	Tanggal pembuatan akun
trans_id	Nomor transaksi
user_id	Id unik transaksi
brand	Merk produk
product_name	Nama produk dan lama langganan
quantity	Jumlah item transaksi
sub_total	Total pembayaran per produk

Penelitian ini menggunakan data kategorik dan data numerik. Data kategorik seperti user_id digunakan sebagai *unique identifier*. Fitur lain dari himpunan data yang tidak digunakan akan dihilangkan. Untuk menghasilkan fitur baru yang dapat diproses dengan metode RFM, metode analisis

berdasarkan data riwayat transaksional untuk pemasaran (Hughes, 1994). *Recency* adalah jarak antara pembelian awal dan pembelian berikutnya. *Frequency* adalah perhitungan berapa kali pelanggan berbelanja dalam periode yang telah ditentukan. *Monetary* adalah hasil perhitungan total transaksi nasabah dalam periode yang telah ditentukan.

Jika pelanggan memiliki beberapa transaksi, nilai *recency* didasarkan pada tanggal dan frekuensi transaksi terakhir. Sedangkan *monetary* adalah nilai total dari keseluruhan pengeluaran. Semua fitur pertama-tama akan dinormalisasi dengan StandardScaler untuk menormalkan data sehingga meminimalkan nilai kesalahan (Thara et al., 2019). Kemudian, data tersebut akan diolah dengan reduksi dimensi dengan Principal Component Analysis (PCA) menjadi data dua dimensi.

2.4. Modeling

Model *machine learning* yang digunakan adalah proses *clustering* dengan menggunakan K-Means++ untuk mengatasi permasalahan algoritma K-Means klasik. Bentuk K-Means klasik secara acak menginisialisasi *centroid* awal, menghasilkan hasil pengelompokan yang berbeda yang dapat menyebabkan kesalahan atau memperlambat algoritma untuk mencapai konvergensi (Gao et al., 2021). Iterasi yang ditetapkan akan membatasi K-Means ++ hingga 300 kali hingga model mencapai konvergensinya. Model optimal dihasilkan dari pemilihan parameter optimal nilai *k*. Untuk proses evaluasi, parameter n_init digunakan sebagai parameter optimal ditentukan menggunakan visualisasi nilai WSS sehingga dihasilkan grafik untuk evaluasi, kemudian dengan *elbow method* dan perhitungan berdasarkan *silhouette score* terbaik (Rousseeuw, 1987). Algoritma K-Means dengan parameter terbaik nantinya akan digunakan untuk mengelompokkan data pelanggan.

2.5. Evaluation

Dengan menetapkan nilai *k*, *cluster* akan dibentuk menggunakan data yang telah melalui pra-pemrosesan dan *dimension reduction* dengan PCA. *Principal Component Analysis* (PCA) adalah metode untuk mengekstraksi pola dalam fitur dengan mengurangi dimensi. PCA menggabungkan beberapa variabel menjadi dua atau tiga komponen. PCA bekerja dengan mengurangi variasi data dan menemukan pola yang kuat dari kumpulan data (Wold, Esbensen & Geladi, 1987). Fitur *monetary*, *recency*, dan *frequency* yang sebelumnya dinormalisasi akan direduksi (*monetary*, *recency*, dan *frequency*. user_id tidak termasuk didalamnya. Pemilihan dua dimensi pada eksperimen bertujuan untuk meminimalkan dimensi komponen dan proses visualisasi data. Atas maksud tersebut, maka data akan direduksi menjadi 2 dimensi. Kemudian data dikelompokkan menurut *cluster* masing-masing, di mana setiap pengguna akan diberi label kelas untuk divisualisasi. Data dikelompokkan berdasarkan kategori. Setiap *cluster* dapat dikelompokkan menjadi rendah, sedang, dan tinggi berdasarkan nilai rata-rata setiap *cluster*. Pertama, hitung nilai kuartil 1 dari setiap fitur himpunan data. Kedua, hitung nilai kuartil 3 dari setiap fitur himpunan data. Ketiga, data yang nilai rata-ratanya di bawah kuartil satu dikategorikan sebagai *low*. Keempat, data yang nilai rata-ratanya berada di kisaran kuartil 1 hingga kuartil tiga dikategorikan sebagai *medium*. Kelima, data yang nilai rata-ratanya di atas kuartil tiga dikategorikan sebagai *high*.

3. KAJIAN PUSTAKA

Dedi et al. mempelajari segmentasi pelanggan berdasarkan nilai *Recency*, *Frequency*, and *Monetary* (RFM) dengan K-means tradisional dan Fuzzy C-means. Data yang digunakan dalam penelitian adalah data transaksional untuk menganalisis perilaku pelanggan. Makalah ini mengusulkan metode baru dengan memilih *centroid* awal di K-Means dan bertujuan untuk menyegmentasikan pelanggan dengan iterasi dan waktu yang berkurang. Algoritma RM K-Means yang diusulkan menghabiskan lebih sedikit waktu dan mengurangi iterasi, membuatnya lebih efektif. Segmentasi berfokus pada kebaruan, frekuensi, dan nilai moneter, memungkinkan perusahaan untuk menyesuaikan strategi pemasaran berdasarkan perilaku pembelian (Dedi et al., 2019).

Christy et al. mengelompokkan pelanggan menjadi beberapa kelompok berdasarkan nilai *Recurrence*, *Frequency*, dan *Monetary* (RFM). Algoritma pengelompokan seperti K-means dan Fuzzy C-means digunakan untuk menganalisis perilaku konsumen dengan data transaksi jual beli. Pemilihan *centroid* awal mampu mengurangi jumlah iterasi dan waktu yang dibutuhkan untuk segmentasi pelanggan. Hasil dibandingkan dengan metode konvensional dengan mempertimbangkan kekompakan kluster, waktu eksekusi, dan jumlah iterasi. Hasil studi menunjukkan bahwa *clustering* membantu perusahaan memberikan rekomendasi produk, mengidentifikasi tren, dan menyesuaikan program pemasaran (Christy et al., 2021).

Jinfeng Zhou dkk. mendemonstrasikan model RFMT (*Recency*, *Frequency*, *Monetary*, and *Interpurchase Time*) untuk menyegmentasikan pelanggan di industri ritel. Model ini dapat digunakan untuk mengidentifikasi berbagai kelompok pelanggan berdasarkan perilaku pembelian mereka. Hasil segmentasi dapat memberikan rekomendasi untuk strategi bisnis, seperti mengalokasikan sumber daya pemasaran dan rekomendasi produk yang disesuaikan. Namun, perlu ada lebih banyak informasi tentang metode yang digunakan untuk mengumpulkan data pelanggan. Selain itu, makalah ini juga perlu memberikan informasi tentang ukuran sampel yang digunakan dalam penelitian. Hasil penelitian dapat menunjukkan informasi yang dapat

membantu mengevaluasi reliabilitas dan generalisasi hasil segmentasi (Zhou, Wei & Xu, 2021).

Rahim et al. menggunakan model *Recency*, *Frequency*, dan *Monetary* (RFM) yang dikombinasikan dengan model pembelajaran mesin seperti Multi-Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk mengklasifikasikan pelanggan. Penelitian ini bertujuan untuk mengklasifikasikan pelanggan berdasarkan perilaku belanja di industri ritel (Rahim et al., 2021).

Anitha et al. menerapkan konsep *business intelligence* dalam mengidentifikasi pelanggan potensial dengan menyediakan data yang relevan dan tepat waktu untuk industri ritel. Studi ini menganalisis riwayat transaksi untuk mengidentifikasi perilaku belanja konsumen dan keuntungan bisnis. Model RFM (Recency, Frequency, dan Monetary) dikombinasikan dengan algoritma K-Means untuk menyegmentasikan pelanggan. Analisis RFM menghitung kebaruan, frekuensi, dan nilai moneter setiap pelanggan untuk proses segmentasi. Algoritma K-Means memilih jumlah *cluster* berdasarkan *silhouette score* berdasarkan nilai RFM (Anitha & Patil, 2022).

Agustino dkk. menggunakan metode RFM dan K-Means dalam memprofil pelanggan pada platform edukasi pada sebuah *start-up* digital. Pada penelitian diusulkan beberapa metrik seperti *Elbow Method*, *Silhouette Scores*, dan *Davis Bouldin Index* dalam menemukan jumlah cluster yang sesuai. Dataset yang digunakan berdasarkan transaksi pada *start-up* tersebut dengan jumlah data sebanyak 283 dan fitur yang terdiri atas email, id transaksi, waktu transaksi, tipe transfer, dan jumlah transfer. Dalam proses normalisasi data digunakan *library* Standar Scaler terhadap fitur *recency*, *frequency*, dan *monetary* yang telah dihitung sebelumnya berdasarkan fitur pada dataset. Hasil dalam penelitian tersebut menunjukkan bahwa 2 *cluster* adalah yang paling optimal, dimana pelanggan pada *cluster* 1 menunjukkan pembelian lebih dari satu kali walaupun dengan nominal pembelian yang sedikit (Agustino et al., 2022).

Monalisa dkk. dalam penelitiannya melakukan identifikasi terhadap pelanggan yang memiliki prospek dengan metode RFM dan DBSCAN *algorithm*. Dengan menggunakan *Silhouette Index* didapatkan *cluster* yang paling optimal yaitu 5. Penelitian tersebut juga melakukan pendekatan analisis demografi terhadap gender, usia, pekerjaan, alamat, dan status pernikahan dalam menganalisis frekuensi pembelian antara *cluster* pertama (memiliki prospek) dengan *cluster loyal customers* (Monalisa et al., 2023).

Al-Yasir dkk. melakukan penelitian dengan menerapkan metode Fuzzy C-Means dan RFM dalam menganalisis loyalitas pelanggan B2B. (Al-Yasir et al., 2024). Penelitian menggunakan *Davies Bouldin Index* (DBI) dalam mendapatkan *cluster* yang sesuai antara 2-10. Didapatkan nilai DBI 0,4908 dan dua sebagai jumlah *cluster* yang sesuai. Digunakan 4 indikator RFM berdasarkan rata-rata dari tiap *Recency*, *Frequency*, dan *Monetary*, yaitu *loyal customer*, *lost customer*, *new customer*, dan *prospect customer* (Monalisa & Fitri, 2019). Didapatkan 2 *cluster*, yaitu *loyal customer* dan *lost customer* sebagai hasil analisis (Al-Yasir et al., 2024).

4. HASIL

4.1. Bentuk Data

Dilakukan pra-pemrosesan untuk beberapa fitur yang berbentuk kategorik dan numerik. Fitur user_id tetap dipertahankan sebagai *unique identifier* untuk setiap transaksi. Setelah proses pra-pemrosesan, 3.900 baris data dihasilkan fitur baru dalam bentuk fitur *monetary*, *frequency*, dan *recency* dalam bentuk numerik. Metode normalisasi data numerik menghasilkan distribusi normal dengan metode *standard scaler*. Rumus *standard scaler* dapat dijelaskan seperti rumus berikut (Thara et al., 2019):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

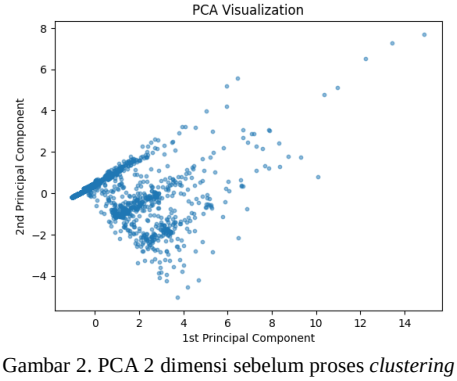
di mana x adalah nilai awal, μ adalah mean, dan σ adalah standar deviasi. Normalisasi ini bertujuan untuk mengubah nilai rata-rata distribusi menjadi nol dan standar deviasinya menjadi 1. Dengan demikian, data yang diperoleh memiliki distribusi yang seragam sehingga data dapat digunakan untuk proses *clustering*. Beberapa sampel data dapat dilihat pada tabel 2.

Tabel 2. Fitur Setelah Proses Normalisasi

user_id	recency	frequency	monetary
001b5eddf7	0,4387	0,4416	-0,2096
0023cedbf9	-0,9347	-0,4402	-0,2096
0024ae4898	0,0752	0,4416	-0,2096
002d8b5849	-0,4700	-0,4402	-0,2096

4.2. Reduksi Dimensi

Setelah didapatkan fitur berdasarkan *recency*, *frequency*, dan *monetary* dilakukan proses reduksi dimensi data menjadi 2 dimensi. Tujuan dari hal tersebut yaitu memudahkan dalam proses visualisasi dan menerangkan *cluster* data. Grafik *scatter plot* digunakan untuk. Pada gambar 2, dapat dilihat bahwa data semakin berhimpit tetapi belum ada pengelompokan yang jelas.



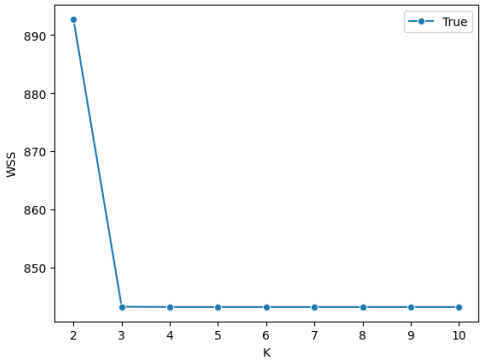
Gambar 2. PCA 2 dimensi sebelum proses *clustering*

4.3. Analisis Jumlah *Cluster*

Sebelum proses *clustering*, pertama dilakukan proses analisis cluster. Proses analisis *cluster* bertujuan untuk menemukan jumlah *cluster* atau nilai *k* yang optimal. Parameter *k* adalah input yang digunakan sebagai nilai parameter fungsi K-Means pada *library* sklearn. Analisis cluster menggunakan *elbow method* dan *silhouette score* yang telah dilakukan pada tahap sebelumnya. Nilai (Dalam Jumlah Kuadrat) WSS dihasilkan dengan melakukan iterasi sebanyak sepuluh kali untuk setiap *cluster*. Pada gambar 3, dapat dilihat bahwa memplot nilai WSS untuk setiap kelompok menghasilkan *elbow method*. Dengan begitu, kita dapat menentukan bahwa nilai *k* terbaik berdasarkan *elbow method* adalah k dengan nilai 3. Mendapatkan *silhouette score* terhadap semua transaksi *k* dari satu hingga sepuluh dapat dilakukan dengan menggunakan rumus 2 berikut (Rousseeuw, 1987):

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

di mana $a(i)$ adalah rata-rata perbedaan/jarak ke semua titik objek dan $b(i)$ adalah nilai minimum antara jarak rata-rata sampel dari sampel ke *cluster* lain. Dengan menghitung *silhouette score* (misalnya, *k* dalam K-Means), *silhouette score* tertinggi menunjukkan jumlah *cluster* terbaik.



Gambar 3. *Elbow method*

Tabel 3. *Silhouette Score*

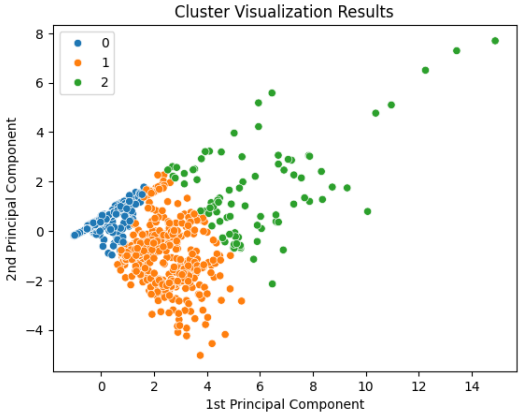
k	wss	silhouette score
3	843,214747	0,638181
4	843,165970	0,638099
5	843,165970	0,638099
6	843,165970	0,638099
7	843,165970	0,638099
8	843,165970	0,638099
9	843,165970	0,638099
10	843,165970	0,638099
2	892,717979	0,637027

Kemudian dilakukan proses analisis *cluster dengan* menggunakan metode *silhouette score*. Iterasi sepuluh kali dilakukan untuk mendapatkan nilai *k* yang sesuai. Mengacu pada tabel 3, dapat disimpulkan bahwa nilai *k* yang sesuai berdasarkan *silhouette score* adalah *k* = 3, dengan nilai 0,63818. *Silhouette score* terbaik adalah yang paling signifikan dibandingkan dengan jumlah *cluster* lainnya.

4.4. *Clustering*

Clustering dilakukan dengan menggunakan metode K-Means++ untuk mendapatkan pola segmentasi pelanggan. Parameter yang diterapkan pada fungsi K-Means ++ adalah n_init = 3 nilai *k*). Selain itu, iterasi maksimum ditentukan dengan menentukan max_iter = 300. Ini bertujuan untuk menentukan batas iterasi hingga mencapai konvergensi. Proses pembentukan model akan membagi data menjadi tiga kelompok yang berbeda. Model K-Means akan menghasilkan label untuk setiap *record data*.

Untuk itu, fitur label akan dibentuk untuk menginisiasi data yang disertakan dalam nomor *cluster*. gambar 4 menunjukkan bagaimana *cluster* didistribusikan berdasarkan pola data. Dengan membagi *cluster* menjadi tiga, maka akan terbentuk 3 *centroid*. Titik data akan dikelompokkan berdasarkan jarak dengan *centroid* terdekat. Hanya *cluster* yang terbentuk untuk saat ini; Kemudian, akan dilakukan proses untuk menemukan pola dari masing-masing *cluster*.



Gambar 4. *Plotting data* untuk tiap cluster

4.5. Pola untuk tiap *cluster*

Menentukan pola masing-masing *cluster* dilakukan dengan mencari nilai mean dari masing-masing *cluster* berdasarkan fitur. Kemudian nilai tersebut akan menjadi patokan untuk menentukan apakah *cluster* dengan fitur RFM-nya dapat dikategorikan sebagai *low, medium,* atau *high*. Setiap fitur dihitung terlebih dahulu. Dikategorikan sebagai *low* ketika di bawah Q1, *medium* antara Q1 dan Q3, dan 'high' ketika lebih dari Q3.

Tabel 4 menunjukkan bagaimana pola data didasarkan pada nilai rata-rata untuk setiap *cluster* terhadap fitur-fiturnya. Nilai *recency* dapat dikategorikan sebagai 'rendah' berarti jarak antara satu pembelian dan pembelian berikutnya cenderung cepat. Nilai *recency high* berarti bahwa waktu antara satu pembelian dan pembelian berikutnya cenderung

lambat. Nilai *frequency* dikategorikan sebagai *low* jika Anda jarang melakukan pembelian, *frequency high* berarti pembelian yang sering. *Monetary* dikategorikan *low* jika nilai uang yang dikeluarkan untuk melakukan pembelian cenderung rendah, sedangkan *high* berarti nilai pembelian tinggi. Nasabah yang diinginkan adalah nasabah yang memiliki nilai *recency* rendah, *frequency* tinggi, dan *monetary* tinggi. Berdasarkan proses kategorisasi yang telah dijelaskan, karakteristik masing-masing *cluster* dapat digambarkan sebagai berikut:

Tabel 4. Karakteristik RFM untuk Tiap *Cluster*

cluster	recency	frequency	monetary
0	low (1,14)	low (1,20)	low (301.640)
1	high (249,61)	medium (2,62)	medium (799.934)
2	medium (233,01)	high (6,41)	high (2.018.088)

pada data yang belum dinormalisasi, pola-pola tersebut dengan nilai mean masing-masing *cluster* untuk setiap fitur RFM dapat dijelaskan pada Tabel 5 sebagai berikut:

Tabel 5. Segmentasi RFM pada Data Riil

segmen	recency	frequency	monetary
Low	< 117,07	< 1,91	< 550.787
Medium	≥ 117,07 dan ≤ 249,61	≥ 1,91 dan ≤ 6,41	≥ 550.787 dan ≤ 2.018.088
High	≥ 249,61	≥ 6,41	≥ 2.018.088

Pelanggan pada *cluster* ini berjumlah 3.303 pelanggan dari total 3.900 pelanggan yang dianalisis. Jumlah ini sekitar 84,69% dari total pelanggan yang dianalisis atau mayoritas pembeli di *e-commerce*. Ini menunjukkan karakteristik pelanggan pada *cluster* 0. Nilai pembelian rata-rata pada *cluster* ini adalah Rp. 301.640, Hal ini dapat diklasifikasikan sebagai nilai pembelian yang rendah. Frekuensi rata-rata pembelian adalah 1,20 kali. Ini diklasifikasikan sebagai frekuensi pembelian yang rendah. Nilai rata-rata waktu yang cukup cepat hingga pembelian berikutnya adalah 1,14 hari. Berdasarkan nilai-nilai tersebut, dapat dikatakan bahwa pelanggan di *cluster* ini cenderung melakukan pembelian dalam waktu singkat, bahkan dengan nilai pembelian yang rendah.

Pelanggan pada *cluster* ini berjumlah 511 pelanggan dari total 3.900 pelanggan yang dianalisis. Jumlah ini sekitar 13,10% dari total pelanggan yang dianalisis. Rata-rata nilai pembelian di *cluster* ini adalah Rp. 799.934 atau tergolong nilai pembelian menengah. Rata-rata frekuensi pembelian adalah 2,62 kali dan rata-rata nilai waktu hingga pembelian berikutnya cukup lama, yaitu 249,61 hari. Berbeda dengan pembeli pada *cluster* sebelumnya, karakteristik pembeli pada *cluster* ini membeli produk dengan nilai sedang namun akan kembali membeli produk dalam jangka waktu yang lama.

Pelanggan pada *cluster* ini berjumlah 86 pelanggan dari total 3.900 pelanggan yang dianalisis. Jumlah ini sekitar 2,20% dari total pelanggan yang dianalisis. Rata-rata nilai pembelian di *cluster* ini yaitu Rp. 2.018.088 atau dengan nilai yang besar dengan frekuensi pembelian rata-rata 6,41 kali, dan nilai rata-rata waktu hingga pembelian berikutnya cukup lama, yaitu 233,01 hari. Karakter pelanggan berikut ini unik karena frekuensinya hingga 3 kali lipat dari *cluster* lainnya, bahkan dengan nilai beli yang besar. Hanya saja jangka waktu pembeliannya cukup lama untuk pembelian berikutnya.

5. KESIMPULAN DAN SARAN

Proses segmentasi pelanggan *e-commerce* dapat dilakukan dengan menggunakan metode RFM. Dengan menerapkan algoritma K-Means++, terbentuk tiga *cluster* berbeda dengan karakteristiknya masing-masing. *Clustering* dengan nilai $k = 3$ menghasilkan nilai WSS 843,214747 dan *silhouette score* 0,638181 sebagai k terbaik. Dengan membagi *cluster* menjadi tiga, didapatkan pola-pola pembelian untuk setiap segmen pelanggan. *Cluster* 0 memiliki nilai RFM rata-rata masing-masing 1,14 (*low*), 1,20 (*low*), dan 301,640 (*low*). *Cluster* 1 memiliki nilai RFM rata-rata masing-masing 249,61 (*high*), 2,62 (*medium*), dan 799,934 (*medium*). *Cluster* 2 memiliki nilai RFM rata-rata masing-masing 233,01 (*medium*), 6,41 (*high*), dan 2018,088 (*high*). Segmentasi dan karakteristik segmen itu sendiri cukup representatif untuk menjelaskan bagaimana setiap *cluster* dipolakan. Dengan begitu, bisnis seperti penjualan majalah *e-commerce* mendapatkan wawasan tentang perilaku belanja pelanggan. Hasilnya dapat digunakan untuk keputusan dalam berbagai proses marketing, termasuk promosi.

Mengingat penelitian ini menggunakan perhitungan *range* untuk menentukan kategori RFM, kedepannya dapat ditentukan dengan menggabungkan metode yang telah diterapkan dengan cara lain, seperti *association rule mining* (Chen & Gunawan, 2023). Akan menarik jika proses segmentasi pelanggan menggunakan algoritma yang berbeda, seperti K-Medoids dan DBSCAN, untuk mengetahui seberapa baik masing-masing algoritma melakukannya dengan menghitung nilai DBI atau *silhouette score* (Aryuni, Madyatmadja & Miranda, 2018).

DAFTAR PUSTAKA

AGUSTINO, D. P., HARSEMADI, I. G., & BUDAYA, I. G. B. A. 2022. Edutech Digital Start-Up Customer Profiling Based On RFM Data Model Using K-Means Clustering. Journal Of Information Systems and Informatics, 4(3), 724-736.

AL-YASIR, A. Y., AFDAL, M., ZARNELLY, Z., & MARSAL, A. 2024. Analisis Loyalitas Pelanggan Business To Business

Berdasarkan Model RFM Menggunakan Algoritma Fuzzy C-Means: Business to Business Customer Loyalty Analysis Based on RFM Model Using Fuzzy C-Means Algorithm. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(1), 359-365.

ANITHA, P., PATIL, M. M. 2022. RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University - Computer and Information Sciences, 34(5), 1785–1792.

ARYUNI, M., MADYATMADJA, E. D., MIRANDA, E. 2018, September. Customer segmentation in XYZ bank using K-means and K-medoids clustering. In 2018 International Conference on Information Management and Technology (ICIMTech) (pp. 412-416). IEEE.

BAHMANI, B., MOSELEY, B., VATTANI, A., KUMAR, R., & VASSILVITSKII, S. 2012. Scalable k-means++. *arXiv preprint arXiv:1203.6402*.

CHEN, A. H.-L., GUNAWAN, S. 2023. Enhancing Retail Transactions: A Data-Driven Recommendation Using Modified RFM Analysis and Association Rules Mining. Applied Sciences, 13(18), 10057

CHEN, Y., LI, M., SONG, J., MA, X., JIANG, Y., WU, S., CHEN, G. L. 2022. A study of cross-border E-commerce research trends: Based on knowledge mapping and literature analysis. Frontiers in Psychology, 13.

CHRISTY, A. J., UMAMAKESWARI, A., PRIYATHARSINI, L., NEYAA, A. 2021. RFM ranking – An effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences, 33(10), 1251–1257.

DEDI, M. I. DZULHAQ, K. W. SARI, S. RAMDHAN, R. TULLAH, SUTARMAN. Customer Segmentation Based on RFM Value Using K-Means Algorithm. 2019. Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1-7.

DWIVEDI, Y. K., ISMAGILOVA, E., HUGHES, D. L., CARLSON, J., FILIERI, R., JACOBSON, J., JAIN, V., KARJALUOTO, H., KEFI, H., KRISHEN, A. S., KUMAR, V., RAHMAN, M. M., RAMAN, R., RAUSCHNABEL, P. A., ROWLEY, J., SALO, J., TRAN, G. A., WANG, Y. 2021. Setting the future of digital and social media marketing research: Perspectives and research propositions. International Journal of Information Management, 59, 102168.

DZULHAQ, M. I., SARI, K. W., RAMDHAN, S., & TULLAH, R. 2019. Customer segmentation based on RFM value using K-means algorithm. In *2019 Fourth International Conference on Informatics and Computing (ICIC)* (pp. 1-7). IEEE.

FAUZAN M., DAVIN. 2023. Unlocking Digital Business Success: Leveraging Artificial Intelligence in Social Media Analytics for Enhanced Customer Insights and Engagement.

GAO, M., PAN, S., CHEN, S., LI, Y., PAN, N., PAN, D., SHEN, X. 2021. Identification Method of Electrical Load for Electrical Appliances Based on K-Means ++ and GCN. IEEE Access, 9, 27026–27037.

HALEEM, A., JAVAID, M., ASIM QADRI, M., PRATAP SINGH, R., SUMAN, R. 2022. Artificial intelligence (AI) applications for marketing: A literature-based study. International Journal of Intelligent Networks, 3, 119–132.

HUGHES, A.M. 1994. Strategic Database Marketing. Probus Publishing Company, Chicago.

JAIN, V., WADHWANI, K., EASTMAN, J. K. 2023. Artificial intelligence consumer behavior: A hybrid review and research agenda. Journal of Consumer Behavior.

KHAJVAND, M., ZOLFAGHAR, K., ASHOORI, S., ALIZADEH, S. 2011. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science, 3, 57–63.

MARTÍNEZ-PLUMED, F., CONTRERAS-OCHANDO, L., FERRI, C., HERNÁNDEZ-ORALLO, J., KULL, M., LACHICHE, N., ... & FLACH, P. 2019. CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, *33*(8), 3048-3061.

MAURITSIUS, T., TURMAWAN, A., & HERDIANSYAH, H. 2023. Customer Segmentation Based On RFM: A Case Study in the context of Pandemic. In 2023 5th International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-6). IEEE.

MIRANTIKA, N., & RIJANTO, E. 2023. Comparative Analysis Of K-Means And K-Medoids Algorithms in Determining Customer Segmentation Using RFM Model. *Journal Of Engineering Science and Technology*, *18*(5), 2340-2351.

MONALISA, S., JUNIARTI, Y., SAPUTRA, E., MUTTAKIN, F., & AHSYAR, T. K. 2023. Customer Segmentation With RFM Models And Demographic Variable Using DBSCAN Algorithm. TELKOMNIKA (Telecommunication Computing Electronics And Control), 21(4), 742-749.

PEDREGOSA, F., VAROQUAUX, GA"EL, GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

RAHIM, M. A., MUSHAFIQ, M., KHAN, S., ARAIN, Z. A. 2021. RFM-based repurchase behavior for customer classification and segmentation. Journal of Retailing and Consumer Services, 61, 102566.

ROUSSEEUW, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

ROSÁRIO, A., RAIMUNDO, R. 2021. Consumer Marketing Strategy and E-commerce in the Last Decade: A Literature Review. Journal of Theoretical and Applied Electronic Commerce Research, 16(7), 3003–3024.

MONALISA S. AND KURNIA F.. 2019. Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour. Telkomnika (Telecommunication Comput. Electron. Control., vol. 17, no. 1, pp. 110–117, 2019, doi: 10.12928/TELKOMNIKA.v17i1.9394.

SARKER, I. H. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160.

SARKER, I. H. 2022. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. SN Computer Science, 3(2), 158.

SIEBERT, M., KOHLER, C., SCERRI, A., & TSATSARONIS, G. Technical Background and Methodology for the Elsevier's Artificial Intelligence Report. 2018.

THARA D.K., B.G, PREMASUDHA., XIONG, F. 2019. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognition Letters, 128, 544–550.

VERMA, S., SHARMA, R., DEB, S., MAITRA, D. 2021. Artificial intelligence in marketing: Systematic review and future research direction. International Journal of Information Management Data Insights, 1(1), 100002.

WIRTH, R., HIPP, J. 2000, April. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of knowledge discovery and Data Mining (Vol. 1, pp. 29-39).

WOLD, S., ESBENSEN, K., GELADI, P. 1987. Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1–3), 37–52.

ZHOU, J., WEI, J., XU, B. 2021. Customer segmentation by web content mining. Journal of Retailing and Consumer Services, 61, 102588.

Halaman ini sengaja dikosongkan