

Data Mining untuk Klasterisasi dan Klasifikasi Pelanggan Berdasarkan Pendapatan dan Transaksi Pembelanjaan Menggunakan Algoritma K-Means

Ikna Awaliyani¹

¹Fakultas Teknologi dan Informatika, Universitas Aisyah Pringsewu, Indonesia

Info Artikel

Riwayat Artikel:

Received April 29, 2025
Revised May 10, 2025
Accepted May 28, 2025

Abstract

In the era of technology and business, handling customers is very important as consumers and a significant source of marketing information. Customer evaluations have a major impact on marketing and service quality can be a benefit or a loss to a seller's reputation. Knowing key customer profiles supports effective marketing strategies with a deep understanding of customer preferences and behavior. This approach is supported by clustering technology for better analysis of customer information, improving management, and designing IT technology according to current developments. The results of data processing form four clusters, forming customer profiles based on family structure and income/expense levels. Understanding cluster characteristics is the basis for an effective marketing strategy. By customizing marketing tactics, targeting specific offers, and optimizing promotions, companies can build strong relationships, increase retention, and achieve business success. This approach provides deep insight, enables targeted decision making, and is responsive to evolving market dynamics.

Keywords: Data Mining, Clusterisasi, Customer

Corresponding Author:

Ikna Awaliyani
Email:
iknaawaliyani106@gmail.com



This is an open access article under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license.

Abstrak

Di era teknologi dan bisnis, penanganan pelanggan merupakan hal yang sangat penting sebagai konsumen dan sumber informasi pemasaran yang penting. Evaluasi pelanggan mempunyai dampak besar terhadap pemasaran dan kualitas pelayanan dapat berupa keuntungan atau kerugian bagi reputasi penjual. Mengetahui profil pelanggan utama mendukung strategi pemasaran yang efektif dengan pemahaman mendalam tentang preferensi dan perilaku pelanggan. Pendekatan ini didukung dengan *clustering technology* untuk analisis informasi pelanggan yang lebih baik, peningkatan manajemen, dan perancangan teknologi IT sesuai perkembangan saat ini. Hasil pengolahan data membentuk empat *cluster*, membentuk profil pelanggan berdasarkan struktur keluarga dan Tingkat pendapatan/pengeluaran. Memahami karakteristik kluster merupakan dasar dari strategi pemasaran yang efektif. Dengan menyesuaikan taktik pemasaran, menargetkan penawaran tertentu, dan mengoptimalkan promosi, perusahaan dapat membangun hubungan yang kuat, meningkatkan retensi, dan mencapai kesuksesan bisnis. Pendekatan ini memberikan wawasan mendalam, memungkinkan pengambilan keputusan yang tepat sasaran, dan responsif terhadap dinamika pasar yang terus berkembang.

Kata Kunci: Data Mining, Klasterisasi, Pelanggan

I. PENDAHULUAN

Dalam perkembangan jaman memperlakukan pelanggan dengan baik merupakan sesuatu yang penting [1]. Saat ini pelanggan berperan penting bukan hanya sebagai seorang *customer* melainkan juga sebagai bentuk pemasaran yang penting. Pelanggan dapat memberikan penilaian atau *review* yang dapat dijadikan sebagai keuntungan atau kerugian dalam proses pemasaran. Pelayanan yang baik dapat membuat pelanggan untuk tidak ragu dalam memberikan penilaian yang baik dan pelayanan yang buruk juga bisa menjadi *boomerang* bagi *seller* dalam pemasaran. Selain penilaian pelanggan serta pentingnya pelayanan bagi pelanggan. mengetahui profil dari pelanggan dapat menjadi salah satu bentuk pemasaran yang penting. Dengan mengetahui profil pelanggan, *seller* atau penjual dapat mengetahui jenis barang yang akan dibeli atau dapat menjadi bahan dalam pengelompokan pelanggan, baik pelanggan *exclusive* maupun *reguler*. Selain itu pengelompokan *customer* mampu meningkatkan nilai penjualan dengan melakukan analisis terhadap beberapa kriteria pembelian pelanggan serta promosi apa yang tepat yang akan di tawarkan ke customer [2]. Tujuan mempergunakan teknologi *clustering* ini memerlukan penggunaan pendekatan analitis untuk menilai informasi pelanggan yang lebih baik lagi. Menerapkan beberapa sistem data mining klasterisasi serta memperbaiki manajemen dan merancang teknologi IT sesuai dengan perkembangan jaman membantu *seller* dalam memperoleh pelanggan baru, membantu dalam mempertahankan pelanggan yang saat ini dan meningkatkan nilai seumur hidup pelanggan [3]. Beragamnya produk dan jasa yang tersedia di pasaran serta persaingan dalam

perdagangan manajemen hubungan pelanggan memiliki peran penting dalam identifikasi dan analisis pelanggan dalam penerapan strategi pemasaran [4]. Salah satu teknik yang berguna dalam analisis bisnis perilaku konsumen dan kategorisasi adalah segmentasi pelanggan [5]. Clusterisasi pelanggan bertujuan untuk pengelompokan pelanggan dalam *cluster* yang homogen. Pengelompokan *cluster* yang homogen memudahkan *seller* dalam menemukan pola sebagai bagian dari strategi pemasaran yang dapat diterapkan untuk meningkatkan pelanggan. Pendapatan dan pengeluaran dari pelanggan merupakan faktor krusial yang sangat penting untuk menentukan seberapa besar minat pelanggan dalam berbelanja dan barang apa saja yang bisa dibeli pelanggan berdasarkan pendapatan dan pengeluaran pelanggan [6]. Pelanggan dengan pendapatan yang tinggi akan diberikan penawaran sesuai dengan kebutuhannya begitupun dengan pelanggan yang membeli untuk kebutuhan pokok untuk meningkatkan minat pembelian *seller* akan menawarkan beberapa diskon untuk memberikan penilaian yang baik dan bermanfaat untuk menarik *customer* baru [7].

II. METODE

Metode penelitian yang diterapkan dalam studi ini adalah sebagai berikut:

2.1 Clustering

Clustering adalah metode yang digunakan untuk mengidentifikasi kelompok-kelompok umum yang terdiri dari kumpulan data. Setiap kelompok biasanya memiliki anggota yang lebih mirip satu sama lain dibandingkan dengan anggota dari kelompok lainnya. Sejak tahun 1970, segmentasi berbasis *clustering* telah menjadi sangat populer dalam penelitian data, terutama di bidang pemasaran. Meskipun *clustering* tidak termasuk dalam metode analisis data terstruktur, teknik ini sangat fleksibel dan penggunaannya sangat bergantung pada karakteristik data atau sampel yang dianalisis [8].

2.2 Metode Elbow

Metode *Elbow* merupakan teknik yang digunakan untuk menentukan jumlah *cluster* optimal dengan menganalisis persentase perbandingan terhadap jumlah *cluster*, yang diidentifikasi melalui titik siku pada grafik. Adapun langkah-langkah algoritma *Elbow* dalam menentukan nilai optimal *k* pada *K-Means* adalah sebagai berikut [5]:

- Menentukan nilai awal untuk *k*.
- Meningkatkan nilai *k* secara bertahap.
- Menghitung nilai *sum of squared errors* (SSE) untuk setiap *k* yang dicoba.
- Menganalisis pola penurunan SSE untuk mengidentifikasi nilai *k* yang menunjukkan penurunan drastis.
- Menentukan dan menetapkan nilai *k* pada titik di mana grafik membentuk siku.

Dalam metode *Elbow*, nilai *cluster* terbaik dipilih berdasarkan penurunan yang signifikan pada nilai *Sum of Square Error* (SSE) dan pembentukan siku. Penghitungan SSE menggunakan rumus (1):

$$SSE = \sum_{k=1}^K \sum_{x_i \in S_k} ||X_i - C_k||^2$$

2.3 K-Means

K-Means merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam beberapa *cluster*. Dalam metode ini, data yang memiliki karakteristik serupa akan ditempatkan dalam *cluster* yang sama, sedangkan data dengan karakteristik yang berbeda akan dikelompokkan ke dalam *cluster* yang berbeda [9]. Secara garis besar, langkah-langkah algoritma *K-Means* dapat dijelaskan sebagai berikut [10]:

- Menentukan nilai *K*, yaitu jumlah *cluster* yang diinginkan.
- Melakukan inisialisasi dengan memilih *K* centroid secara acak, kemudian menghitung jarak setiap data terhadap centroid tersebut menggunakan rumus jarak *Euclidean*.

$$d(x_i, \mu_j) = \sqrt{\sum (x_i - \mu_j)^2}$$

Keterangan :

x_i : Data Kinerja

μ_j : *Centroid* pada *cluster* ke -*j*s

Setiap data dipisahkan berdasarkan kedekatannya dengan *centroid* yaitu dengan mencari jarak terpendek.

3. Centroid baru diperbarui dengan menghitung rata-rata dari seluruh data dalam masing-masing *cluster* menggunakan rumus berikut:

$$\mu_j(t+1) = \frac{1}{N_{sj}} \sum_{j \in s_j} x_j$$

Keterangan :

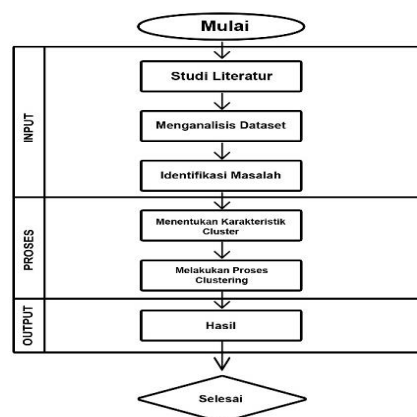
μ_j : Centroid baru pada iterasi (t+1)

N_{sj} : Data pada *cluster* s_j

4. Jika terdapat perubahan data dalam setiap *cluster*, maka langkah 2 hingga 5 akan terus diulangi sampai tidak terjadi lagi perubahan keanggotaan dalam masing-masing *cluster*.

a. Alur Penelitian

Penelitian ini dibagi ke dalam tiga tahapan utama, yaitu tahap *input*, tahap proses, dan tahap *output*. Setiap tahap memiliki urutan kegiatan masing-masing, yang ditampilkan pada Gambar 2.1 di bawah ini sebagai rangkaian langkah-langkah dalam pelaksanaan penelitian.



Gambar 2.1 Alur Penelitian

III. HASIL DAN PEMBAHASAN

Clustering, juga dikenal sebagai analisis *cluster*, merupakan topik yang sangat signifikan dalam bidang data mining. Proses ini melibatkan partisi kumpulan data menjadi beberapa kelompok, di mana titik-titik data dalam setiap kelompok atau *cluster* lebih erat terhubung satu sama lain daripada dengan kelompok atau *cluster* lainnya. Pada tahap ini, konsep-konsep kunci seperti Customer Relationship Management (CRM), segmentasi pelanggan, dan penerapan segmentasi pelanggan dieksplorasi dan dibahas secara rinci. Pentingnya konsep-konsep ini bagi dunia bisnis dan organisasi juga ditekankan [11].

Target pemasaran dan segmentasi pelanggan sering digunakan secara bergantian karena keduanya sangat terkait erat. Sasaran pemasaran melibatkan pengelompokan pembeli berdasarkan karakteristik tertentu yang menjadi fokus perhatian perusahaan [6]. Dengan kata lain, *clustering* dijelaskan sebagai proses penggabungan titik-titik data dengan mengidentifikasi korespondensi berdasarkan atribut yang ada dalam data mentah. Meskipun demikian, fokus utamanya adalah menemukan jumlah dan jenis *cluster* yang sesuai dan memiliki pandangan yang mendalam untuk keperluan analisis [12]. Proses ini merupakan tugas berulang di mana data mentah yang besar dipindai untuk mengidentifikasi pola dan kesamaan. Dalam mencari pengetahuan yang signifikan dari data yang tidak terorganisir, algoritma *clustering* tertentu beserta parameter yang sesuai mungkin menjadi pilihan yang optimal dalam konteks pasar.

3.1 Dataset

Data yang digunakan pada penelitian ini merupakan data yang di ambil dari <https://www.kaggle.com/> data yang diberikan merupakan data dalam bentuk .csv. Dalam data tersebut terdapat data pelanggan yang didalamnya terdapat ID customer yang unik, tahun kelahiran pelanggan, Pendidikan terakhir, Status pernikahan Pendapatan tahunan

pelanggan, Jumlah anak, jumlah tanggungan remaja, Tanggal pendaftaran menjadi *customer*, jumlah hari terakhir pembelian, dan jumlah pembelian *wine*.

Berdasarkan banyaknya kategori atau atribut data yang digunakan, pada penelitian ini analisi segmentasi konsumen yang dipergunakan adalah data pendapatan tahunan pelanggan.

3.2 Cleaning Data

Cleaning data merupakan data yang memiliki nilai pendapatan yang hilang, data yang diambil merupakan data pelanggan baru. Proses *cleaning* data dapat ditunjukkan pada gambar 2.2

data.describe()

	Income	Kidhome	Teenhome	Recency	Wines	Fruits
count	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000
mean	52247.251354	0.441787	0.505415	49.012635	305.091606	26.356047
std	25173.076661	0.536896	0.544181	28.948352	337.327920	39.793917
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	35303.000000	0.000000	0.000000	24.000000	24.000000	2.000000
50%	51381.500000	0.000000	0.000000	49.000000	174.500000	8.000000
75%	68522.000000	1.000000	1.000000	74.000000	505.000000	33.000000
max	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000

8 rows x 26 columns

Gambar 2.2 Cleaning Data

3.3 Processing Data

Pada proses ini dilakukan pemeriksaan tanggal tercatat terbaru dan terlama menskalakan fitur menggunakan scaler standar dan membuat kerangka dataset untuk pengurangan dimensi kategori dataset : *education, income, kithome, tinline, recency, wines, fruits, meat, fish, sweets, number store purchahses, number web visit month, customer for, age, spent, living with, children, family size, his parent*. Proses *processing* data dapat ditunjukkan pada gambar 2.3.

Dataframe to be used for further modelling:

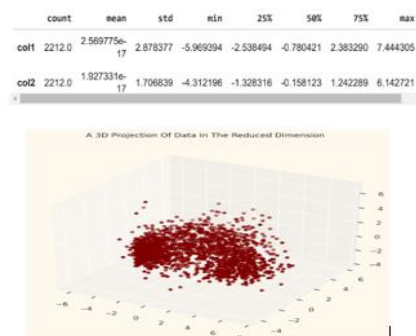
	Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat
0	-0.893586	0.287105	-0.822754	-0.929699	0.310353	0.977660	1.552041	1.690293
1	-0.893586	-0.260882	1.040021	0.908097	-0.380813	-0.872618	-0.637461	-0.718230
2	-0.893586	0.913196	-0.822754	-0.929699	-0.795514	0.357935	0.570540	-0.178542
3	-0.893586	-1.176114	1.040021	-0.929699	-0.795514	-0.872618	-0.561961	-0.655787
4	0.571657	0.294307	1.040021	-0.929699	1.554453	-0.392257	0.419540	-0.218684

5 rows x 23 columns

Gambar 2.3 Reprocessing Data

3.4 Dimensionality Reduction

Pada tahap ini akan dilakukan pengurangan dimensi, ada banyak faktor yang menjadi dasar klasifikasi akhir, semakin tinggi jumlah fitur semakin sulit untuk di kasifikasikan banyak dari fitur berkorelasi dan menjadi mubazir. maka dari itu dilakukan pengurangan dimensi sebelum memasukan ke klasifikasi [8]. Reduksi dimesi adalah proses pengurangan jumlah variabel acak yang dipertimbangkan dengan memperoleh sekumpulan variabel utama. Proses pengurangan dimensi dapat dirunjukkan pada gambar 2.4.

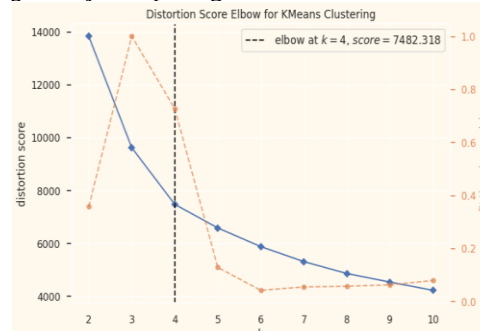


Gambar 2.4 Dimensionality Reduction

3.5 Clustering

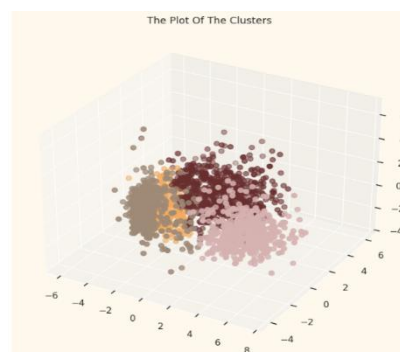
Setelah mereduksi atribut menjadi 3 dimensi selanjutnya dilakukan pengelompokan dengan melibatkan penggabungan mencapai jumlah kluster yang diinginkan tercapai. Langkah-langkah *clustering* :

- Metode *elbow* untuk menentukan jumlah cluster yang dibentuk. Berdasarkan hasil dari proses metode *elbow* pada gambar diagram menunjukkan bahwa terdapat 4 *distortion score* yang akan menjadi sebuah *cluster* optimal untuk data ini. Proses *clustering* ditunjukkan pada gambar 2.5.



Gambar 2.5 Clustering

- Dilakukan Pengelompokan *Agglomerative*
Pengelompokan seperti pada gambar 2.6 dibawah ini sehingga diperoleh *scatterplot* yang memiliki beberapa kecenderungan berdasarkan perbedaan warna pada *scatterplot*.



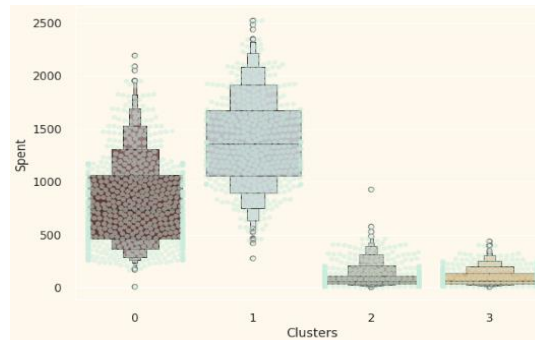
Gambar 2.6 Pengelompokan Agglomerative

- Memeriksa cluster yang terbentuk melalui *scatterplot*.
Berdasarkan gambar 2.6 diatas terdapat perbedaan warna diperoleh kecenderungan pembelian yang dilakukan oleh *customer*.

3.6 Evaluating Model

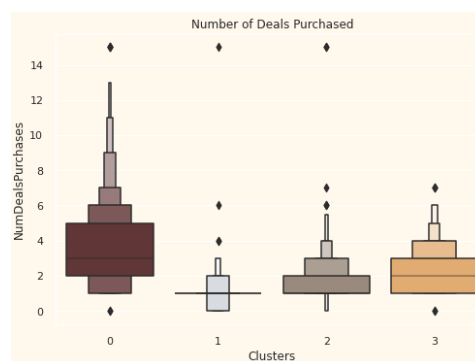
Tujuan pada evaluasi model ini untuk mempelajari pola-pola pada *cluster* yang terbentuk dan mengetahui sifat dari pola *cluster* tersebut.

- Melihat distribusi grup cluster berdasarkan distribusi grup kluster setiap kluster terdistribusi secara merata.
- Berdasarkan plot pendapatan vs pengeluaran menunjukkan terdapat 4 kelompok
Kelompok 0: merupakan kelompok *cus-tomer* dengan pengeluaran tinggi dan pendapatan rata-rata
Kelompok 1: merupakan *customer* dengan pengeluaran tinggi dan pen-dapatan tinggi
Kelompok 2: merupakan *customer* dengan pengeluaran rendah dan pendapatan rendah
Kelompok 3: merupakan *customer* dengan pengeluaran tinggi dan pen-dapatan rendah



Gambar 2.7 Evaluating Model

Dapat terlihat jelas bahwa kluster 1 merupakan Kumpulan pelanggan terbesar pada penelitian ini. Dan diikuti oleh kluster 0. Selanjutnya menelusuri pengeluaran untuk setiap cluster sebagai strategi pemasaran yang ditargetkan.

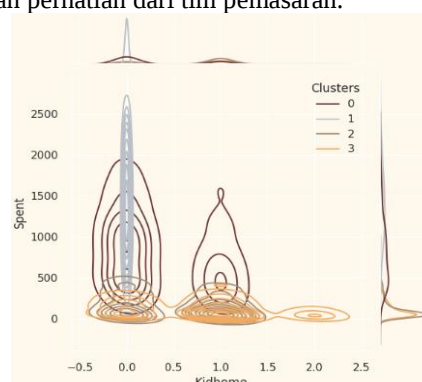


Gambar 2.8 Cluster Pengeluaran

Sejauh ini pada program promo sebelumnya, tanggapan terhadap promosi yang dilakukan oleh *seller* masih kurang. Berdasarkan dataset tidak ada kelompok kluster yang mengambil setiap promo sehingga diperlukan promosi yang lebih tepat sasaran dan terencana.

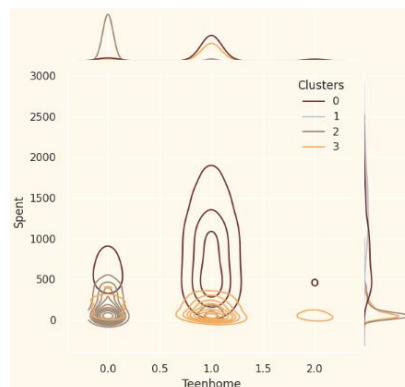
3.7 Profiling

Dengan terbentuknya *cluster* dan melihat kebiasaan pembeli, dapat diambil kesimpulan tentang siapa pelanggan utama dan siapa yang membutuhkan perhatian dari tim pemasaran.



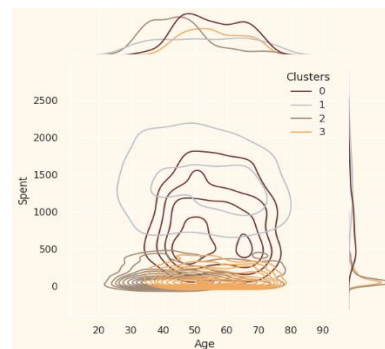
Gambar 2.9 Profiling

Kelompok keluarga 1 dengan pengeluaran tinggi dan pendapatan tinggi dan memiliki anak di rumah akan lebih banyak pengeluaran untuk berbelanja. Selanjutnya diikuti oleh kelompok 0, dengan pengeluarannya tinggi pendapatan rata-rata.



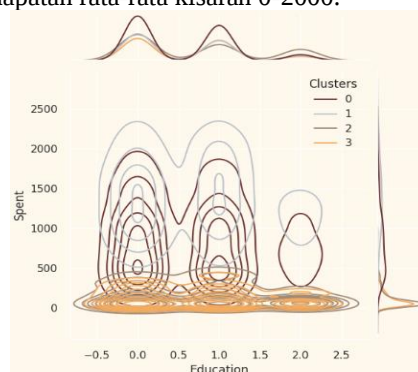
Gambar 2.10 Kelompok Pengeluaran Tinggi

Keluarga kelompok 0 dengan pengeluaran tinggi dan penghasilan rata-rata yang memiliki remaja dirumah memiliki kecenderungan untuk lebih banyak pengeluaran pembelian dibanding dengan keluarga kelompok 3 dengan pengeluaran tinggi dan pendapatan rendah.



Gambar 2.11 Cluster Kelompok Usia

Berdasarkan diagram hasil klaster kelompok usia 30-80 memiliki kecenderungan masuk kedalam kelompok keluarga 1 dengan pengeluaran tinggi dan pendapatan tinggi dengan kisaran 500-2500 dilanjutkan keluarga dengan kelompok 0 dengan pengeluaran tinggi dan pendapatan rata-rata kisaran 0-2000.



Gambar 2.12 Cluster Pendidikan Rata-rata

Berdasarkan Pendidikan rata-rata memiliki pengaruh terhadap pengeluaran keluarga. Keluarga kelompok 1 dengan Pendidikan rata-rata klaster 1,0 memiliki kecenderungan untuk lebih banyak melakukan pembelian dengan kisaran 500-2500, dilanjutkan dengan keluarga kelompok 0 dengan pembelian rata-rata 0-2000, selanjutnya keluarga kelompok 2 merupakan keluarga dengan pengeluaran rendah dan pendapatan rendah yang memungkinkan untuk melakukan pembelian antara 0-500.

IV. SIMPULAN

Dengan memanfaatkan hasil pemrosesan data yang menghasilkan empat kelompok atau klaster, kita dapat membuat profil pelanggan yang didasarkan pada struktur keluarga dan tingkat pendapatan atau pengeluaran mereka. Pemahaman yang mendalam terhadap karakteristik masing-masing klaster ini memberikan landasan yang kokoh untuk merencanakan strategi pemasaran yang lebih efektif. Dengan mengetahui preferensi, kebutuhan, dan perilaku konsumen dalam setiap kelompok, perusahaan dapat menyesuaikan taktik pemasaran, menargetkan penawaran spesifik, dan mengoptimalkan upaya promosi. Pendekatan ini membantu membangun hubungan yang lebih kuat dengan pelanggan, meningkatkan retensi, dan pada gilirannya, meningkatkan kesuksesan bisnis secara keseluruhan.

DAFTAR PUSTAKA

- [1] C. Wijaya, A. S, and W. Hasanah, "Pelanggan dan Kepuasan," *J. Dharmawangsa*, pp. 16–25, 2018.
- [2] S. Tripathi, A. Bhardwaj, and P. E, "Approaches to Clustering in Customer Segmentation," *Int. J. Eng. Technol.*, vol. 7, no. 3.12, p. 802, 2018, doi: 10.14419/ijet.v7i3.12.16505.
- [3] S. Wahyuni, T. T. Wulansari, and F. Fahrullah, "Segmentasi Pelanggan Berdasarkan Analisis Recency, Frequency, Monetary Menggunakan Algoritma K-Means Pada CV. Toedjoe Sinar Group," *J. Rekayasa Teknol. Inf.*, vol. 7, no. 2, p. 180, 2023, doi: 10.30872/jurti.v7i2.8748.
- [4] S. Sharyanto and D. Lestari, "Penerapan Data Mining Untuk Menentukan Segmentasi Pelanggan Dengan Menggunakan Algoritma K-Means dan Model RFM Pada E-Commerce," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 4, p. 866, 2022, doi: 10.30865/jurikom.v9i4.4525.
- [5] Z. J. Lee, C. Y. Lee, L. Y. Chang, and N. Sano, "Clustering and classification based on distributed automatic feature engineering for customer segmentation," *Symmetry (Basel)*, vol. 13, no. 9, pp. 1–16, 2021, doi: 10.3390/sym13091557.
- [6] B. Y. Putra, F. Y. Azzahra, and I. A. Erlanda, "Klasterisasi Pengunjung Mall Menggunakan Algoritma K-Means Berdasarkan Pendapatan Dan Pengeluaran," *J. Inform. dan Tek. Elektro Terap.*, vol. 11, no. 3s1, 2023, doi: 10.23960/jitet.v11i3s1.3392.
- [7] I. Maskanah, "Segmentasi Pelanggan Toko Purnama dengan Algoritma K-Means dan Model RFM untuk Perancangan Strategi Pemasaran," *INOVTEK Polbeng - Seri Inform.*, vol. 5, no. 2, p. 218, 2020, doi: 10.35314/isi.v5i2.1443.
- [8] R. Butler, E. Hinton, M. Kirwan, and A. Salih, "Customer Behaviour Classification Using Simulated Transactional Data," *Eur. Model. Simul. Symp. EMSS*, no. September, pp. 0–10, 2022, doi: 10.46354/i3m.2022.emss.039.
- [9] S. A. Kusuma Melati and Arief Wibowo, "Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model Qrf Pada Perusahaan Rintisan Penyedia Tenaga Kerja," *J. Ilm. Ilmu Komput.*, vol. 6, no. 2, pp. 15–19, 2020, doi: 10.35329/jiik.v6i2.138.
- [10] J. Reddy, "Customer Segmentation Using Clustering Jaswanth Reddy".
- [11] M. Nirmala and M. Shah Makzoom, "Application Development for Customer Segmentation Using an Unsupervised Learning Algorithm," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 4099, pp. 127–133, 2023, doi: 10.32628/ijrsrset2310215.
- [12] S. Hartini, W. Gata, S. Kurniawan, H. Setiawan, and K. Novel, "Cosmetics Customer Segmentation and Profile in Indonesia Using Clustering and Classification Algorithm," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012001.
- [13] A. Unique, "濟無No Title No Title No Title," vol. 15, no. 0, pp. 1–23, 2016.