PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 668-673
E-ISSN: 3109-8886

Penerapan Teknik Clustering Menggunakan Algoritma K-Means Berbasis RFM untuk Segmentasi Pelanggan pada Dataset Online Retail II

[1]Andika Febrian Nurhidayat, [2]Muhammad Jehan Leon Kusuma, [3]Ridho Septiawan

[123]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Tangerang Selatan, Indonesia

[1]febrianandika854@gmail.com, [2]jehanlagi@gmail.com, [3]ridoseptiawan04@gmail.com

Abstract

The growth of transaction data in the retail sector provides valuable information that can be utilized to better understand customer behavior. This study aims to perform customer segmentation using the K-Means clustering algorithm based on Recency, Frequency, and Monetary (RFM) analysis on the Online Retail II dataset. The research process includes data cleaning, handling missing values, removing duplicate records, treating outliers using the Interquartile Range (IQR) method, and constructing RFM features. The processed data were then normalized and grouped using the K-Means algorithm. The optimal number of clusters was determined using the Elbow Method, while the clustering performance was evaluated using the Silhouette Score. The results indicate that the optimal clustering structure consists of three customer groups with a Silhouette Score of 0.4596, suggesting a reasonably good clustering quality. The identified segments are VIP Customers, Regular Customers, and At Risk Customers. VIP Customers demonstrate high purchase frequency and spending value, whereas At Risk Customers exhibit low transaction activity and a higher likelihood of customer attrition. The findings can support data-driven marketing strategies, customer retention programs, and business decision-making processes in the retail industry.

Keywords: *Clustering, K-Means, RFM, Customer Segmentation, Online Retail II*

Abstrak

Pertumbuhan data transaksi pada sektor retail menghasilkan informasi yang dapat dimanfaatkan untuk memahami perilaku pelanggan secara lebih mendalam. Penelitian ini bertujuan melakukan segmentasi pelanggan menggunakan teknik clustering dengan algoritma K-Means berbasis Recency, Frequency, dan Monetary (RFM) pada dataset Online Retail II. Tahapan penelitian meliputi pembersihan data, penanganan nilai hilang, penghapusan duplikasi, penanganan outlier menggunakan metode Interquartile Range (IQR), serta pembentukan fitur RFM. Data yang telah diproses kemudian dinormalisasi dan dikelompokkan menggunakan algoritma K-Means. Penentuan jumlah cluster dilakukan menggunakan Elbow Method, sedangkan kualitas hasil clustering dievaluasi menggunakan Silhouette Score. Hasil penelitian menunjukkan bahwa jumlah cluster optimal adalah tiga kelompok pelanggan dengan nilai Silhouette Score sebesar 0,4596 yang menunjukkan kualitas pemisahan cluster yang cukup baik. Segmentasi yang dihasilkan terdiri atas VIP Customer, Regular Customer, dan At Risk Customer. Kelompok VIP memiliki frekuensi transaksi dan nilai pembelian yang tinggi, sedangkan kelompok At Risk menunjukkan aktivitas transaksi yang rendah dan berpotensi berhenti berbelanja. Hasil penelitian ini dapat dimanfaatkan sebagai dasar penyusunan strategi pemasaran yang lebih tepat sasaran, peningkatan loyalitas pelanggan, serta pengambilan keputusan bisnis berbasis data pada perusahaan retail.

Kata Kunci: Clustering, K-Means, RFM, Segmentasi Pelanggan, Online Retail II

A. PENDAHULUAN

Persaingan dalam industri retail menuntut perusahaan untuk memahami perilaku pelanggan secara lebih mendalam agar strategi pemasaran yang diterapkan dapat berjalan secara efektif. Aktivitas transaksi yang dilakukan pelanggan setiap hari menghasilkan data dalam jumlah besar yang dapat dimanfaatkan sebagai sumber informasi untuk mendukung pengambilan keputusan bisnis. Pengolahan data transaksi yang tepat memungkinkan perusahaan memperoleh gambaran mengenai karakteristik pelanggan, pola pembelian, serta tingkat kontribusi pelanggan terhadap pendapatan perusahaan (Anitha & Patil, 2022).

Salah satu pendekatan yang banyak digunakan dalam pemanfaatan data pelanggan adalah segmentasi pelanggan. Segmentasi pelanggan bertujuan mengelompokkan pelanggan berdasarkan karakteristik dan perilaku yang serupa sehingga perusahaan dapat memberikan perlakuan yang berbeda pada setiap kelompok pelanggan. Penerapan segmentasi pelanggan terbukti membantu perusahaan dalam meningkatkan efektivitas pemasaran, mempertahankan pelanggan potensial, serta

Diterima Redaksi: 01-06-2026 | Selesai Revisi: 07-06-2026 | Diterbitkan Online: 08-06-2023

mengoptimalkan strategi bisnis yang dijalankan (Alzami et al., 2023).

Dalam proses segmentasi pelanggan, model Recency, Frequency, dan Monetary (RFM) sering digunakan untuk menggambarkan perilaku pelanggan berdasarkan riwayat transaksi yang dimiliki. Recency menunjukkan waktu transaksi terakhir pelanggan, Frequency menunjukkan intensitas transaksi yang dilakukan, sedangkan Monetary menggambarkan total nilai transaksi pelanggan. Kombinasi ketiga indikator tersebut mampu memberikan representasi yang lebih komprehensif terhadap aktivitas pelanggan sehingga banyak digunakan dalam penelitian customer segmentation (Winaryanti et al., 2025).

Teknik clustering merupakan salah satu metode data mining yang dapat digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik tanpa memerlukan label kelas sebelumnya. Salah satu algoritma clustering yang paling banyak digunakan adalah K-Means karena memiliki proses komputasi yang relatif sederhana, mampu menangani data numerik dengan baik, serta menghasilkan cluster yang mudah diinterpretasikan. Penelitian sebelumnya menunjukkan bahwa kombinasi metode RFM dan algoritma K-Means mampu menghasilkan segmentasi pelanggan yang efektif pada berbagai kasus bisnis retail maupun e-commerce (Awaliyah, 2024).

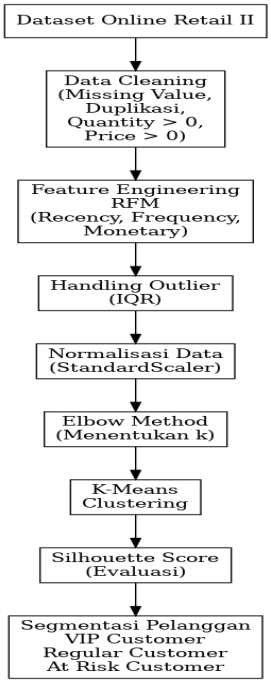
Dataset yang digunakan dalam penelitian ini adalah Online Retail II yang berisi data transaksi pelanggan pada perusahaan retail online. Dataset tersebut memiliki atribut yang mendukung pembentukan nilai RFM sehingga sesuai digunakan untuk proses segmentasi pelanggan. Sebelum dilakukan proses clustering, data terlebih dahulu melalui tahapan pembersihan data, penanganan nilai hilang, penghapusan data duplikat, dan penanganan outlier untuk meningkatkan kualitas data yang digunakan dalam proses analisis.

Penelitian ini bertujuan menerapkan teknik clustering menggunakan algoritma K-Means berbasis RFM untuk melakukan segmentasi pelanggan pada dataset Online Retail II. Hasil segmentasi yang diperoleh diharapkan dapat membantu perusahaan dalam mengidentifikasi kelompok pelanggan berdasarkan perilaku transaksi sehingga dapat digunakan sebagai dasar penyusunan strategi pemasaran yang lebih tepat sasaran, peningkatan loyalitas pelanggan, serta pengambilan keputusan bisnis yang berbasis data.

B. METODE

Penelitian ini memanfaatkan dataset Online Retail II yang berisi riwayat transaksi pelanggan pada perusahaan retail online. Dataset tersebut terdiri atas atribut *Invoice*, *StockCode*, *Description*, *Quantity*, *InvoiceDate*, *Price*, *Customer ID*, dan *Country*. Sebelum dilakukan pengolahan, jumlah data yang tersedia mencapai 541.910 transaksi.

Tahapan penelitian diawali dengan proses persiapan data untuk memastikan kualitas informasi yang digunakan pada tahap analisis. Alur penelitian yang diterapkan ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian Segmentasi Pelanggan Menggunakan Algoritma K-Means Berbasis RFM

Tahap awal pengolahan data dilakukan melalui proses *data cleaning*. Pada tahap ini dilakukan penghapusan data duplikat, penghapusan baris yang tidak memiliki identitas pelanggan (*Customer ID*), penyaringan transaksi dengan nilai *Quantity* kurang dari atau sama dengan nol, serta penyaringan data dengan nilai *Price* kurang dari atau sama dengan nol. Selain itu, atribut *InvoiceDate* dikonversi ke format tanggal agar dapat digunakan dalam proses perhitungan waktu transaksi pelanggan.

Karakteristik pelanggan kemudian dibentuk menggunakan pendekatan *Recency, Frequency, Monetary* (RFM). Pendekatan ini digunakan untuk menggambarkan perilaku pelanggan melalui tiga indikator utama, yaitu jarak waktu sejak transaksi terakhir (*Recency*), jumlah transaksi yang pernah dilakukan (*Frequency*), dan total nilai pembelian pelanggan (*Monetary*) (Anitha & Patil, 2022). Pembentukan fitur RFM dilakukan dengan mengelompokkan seluruh transaksi berdasarkan *Customer ID* sehingga setiap pelanggan direpresentasikan oleh satu baris data.

Untuk mengurangi pengaruh nilai ekstrem terhadap hasil pengelompokan, dilakukan deteksi dan penghapusan *outlier* menggunakan metode *Interquartile Range* (IQR). Data yang berada di luar batas distribusi yang ditentukan tidak disertakan dalam proses analisis. Setelah seluruh tahapan pembersihan data selesai dilakukan, jumlah transaksi yang memenuhi kriteria analisis menjadi 392.693

transaksi. Selanjutnya, proses transformasi RFM menghasilkan 3.710 data pelanggan yang digunakan sebagai masukan pada proses clustering.

Tabel 1. Ringkasan Jumlah Data Hasil Pengolahan

Tahap Pengolahan Data	Jumlah Data
Dataset Awal	541.910
Setelah Data Cleaning	392.693
Setelah Transformasi RFM dan Penghapusan Outlier	3.710

Tabel 1 menunjukkan perubahan jumlah data pada setiap tahapan pengolahan. Penurunan jumlah data terjadi sebagai konsekuensi dari proses pembersihan data dan penghapusan data yang tidak memenuhi kriteria analisis. Hasil akhir yang digunakan dalam proses clustering berupa 3.710 data pelanggan yang telah direpresentasikan dalam bentuk nilai RFM.

Sebelum proses pengelompokan dilakukan, data terlebih dahulu dinormalisasi menggunakan *StandardScaler*. Langkah ini bertujuan menyamakan skala antarvariabel sehingga perbedaan rentang nilai pada atribut *Recency*, *Frequency*, dan *Monetary* tidak mendominasi proses pembentukan cluster.

Proses segmentasi pelanggan dilakukan menggunakan algoritma K-Means. Algoritma ini mengelompokkan objek berdasarkan kedekatan karakteristik sehingga anggota dalam satu cluster memiliki tingkat kemiripan yang lebih tinggi dibandingkan dengan anggota pada cluster lain. Pemilihan K-Means didasarkan pada kemampuannya dalam menangani data numerik dan kemudahannya dalam menginterpretasikan hasil segmentasi pelanggan (Alzami et al., 2023).

Jumlah cluster yang digunakan tidak ditentukan secara langsung, melainkan diperoleh melalui *Elbow Method*. Metode ini digunakan untuk mengamati perubahan nilai *Sum of Squared Error* (SSE) pada beberapa alternatif jumlah cluster sehingga dapat diperoleh nilai cluster yang paling representatif. Setelah proses clustering selesai dilakukan, kualitas hasil pengelompokan dievaluasi menggunakan *Silhouette Score*. Nilai tersebut digunakan untuk mengukur tingkat kedekatan data pada cluster yang sama sekaligus tingkat pemisahan antarcluster (Awaliyah, 2024).

Tahap akhir penelitian dilakukan dengan menganalisis karakteristik masing-masing cluster berdasarkan nilai rata-rata *Recency*, *Frequency*, dan *Monetary*. Berdasarkan hasil tersebut, setiap kelompok pelanggan diberikan label interpretatif yang terdiri atas *VIP Customer*, *Regular Customer*, dan *At Risk Customer* untuk memudahkan proses penyusunan rekomendasi bisnis.

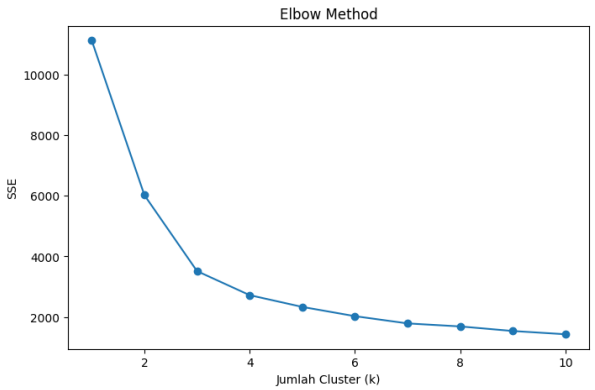
C. HASIL DAN PEMBAHASAN

1. Hasil Preprocessing Data

Tahap awal penelitian dilakukan dengan membersihkan dataset Online Retail II yang berjumlah 541.910 transaksi. Proses pembersihan mencakup penghapusan data duplikat, penghapusan nilai hilang pada atribut *Customer ID*, penghapusan transaksi dengan nilai *Quantity* dan *Price* yang tidak valid, serta penanganan *outlier* menggunakan metode *Interquartile Range* (IQR). Setelah proses tersebut dilakukan, diperoleh 392.693 transaksi yang memenuhi kriteria analisis. Selanjutnya data transaksi ditransformasikan menggunakan metode RFM sehingga menghasilkan 3.710 data pelanggan yang digunakan sebagai masukan pada proses clustering.

2. Penentuan Jumlah Cluster

Jumlah cluster ditentukan menggunakan *Elbow Method* dengan mengamati perubahan nilai *Sum of Squared Error* (SSE) pada beberapa alternatif jumlah cluster.

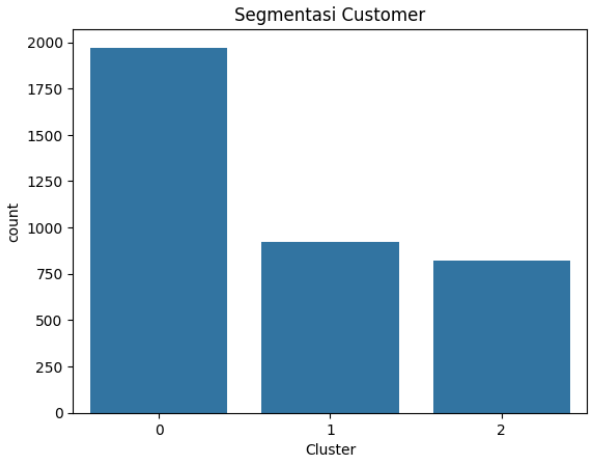


Gambar 2. Hasil Elbow Methode

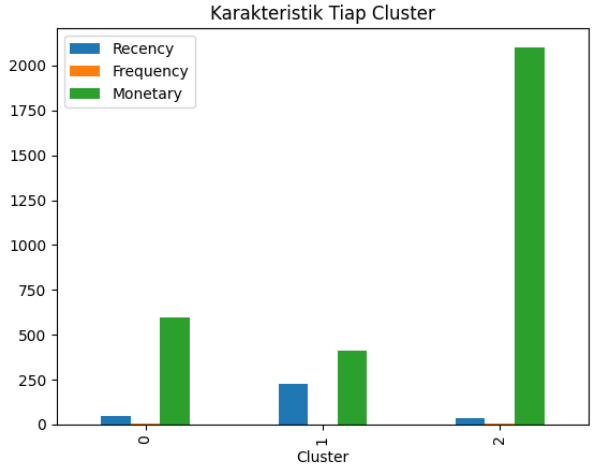
Berdasarkan Gambar 2, terjadi penurunan nilai SSE yang cukup tajam pada jumlah cluster 1 hingga 3. Setelah titik tersebut, penurunan SSE berlangsung lebih landai sehingga membentuk pola siku (*elbow*). Kondisi ini menunjukkan bahwa tiga cluster merupakan jumlah yang paling representatif untuk digunakan pada proses segmentasi pelanggan. Oleh karena itu, penelitian ini menggunakan nilai $k = 3$ pada algoritma K-Means.

3. Hasil Segmentasi Pelanggan

Setelah jumlah cluster ditentukan, proses pengelompokan pelanggan dilakukan menggunakan algoritma K-Means.



Gambar 3. Hasil Segmentasi Pelanggan Berdasarkan Cluster



Gambar 4. Karakteristik Tiap Cluster

Gambar 3 menunjukkan distribusi pelanggan pada masing-masing cluster. Cluster 0 memiliki jumlah anggota paling banyak dibandingkan cluster lainnya. Sementara itu, cluster 1 dan cluster 2 memiliki jumlah anggota yang lebih sedikit dengan karakteristik transaksi yang berbeda. Hasil tersebut menunjukkan bahwa algoritma K-Means berhasil membentuk kelompok pelanggan berdasarkan kemiripan perilaku transaksi.

4. Evaluasi Hasil Clustering

Kualitas hasil clustering dievaluasi menggunakan *Silhouette Score*. Berdasarkan hasil pengujian diperoleh nilai *Silhouette Score* sebesar **0,4596**.

Nilai tersebut menunjukkan bahwa hasil pengelompokan memiliki kualitas yang cukup baik. Sebagian besar data memiliki tingkat kedekatan yang lebih tinggi dengan anggota pada cluster yang sama dibandingkan dengan cluster lainnya. Dengan demikian, hasil segmentasi yang diperoleh dapat digunakan sebagai dasar dalam proses analisis pelanggan.

5. Analisis Karakteristik Cluster

Karakteristik setiap cluster dianalisis menggunakan rata-rata nilai *Recency*, *Frequency*, dan *Monetary*.

Tabel 2. Profil Cluster Berdasarkan Nilai RFM

Cluster	Recency	Frequency	Monetary
0	48,74	2,12	596,54
1	226,80	1,49	412,57
2	36,66	6,05	2101,79

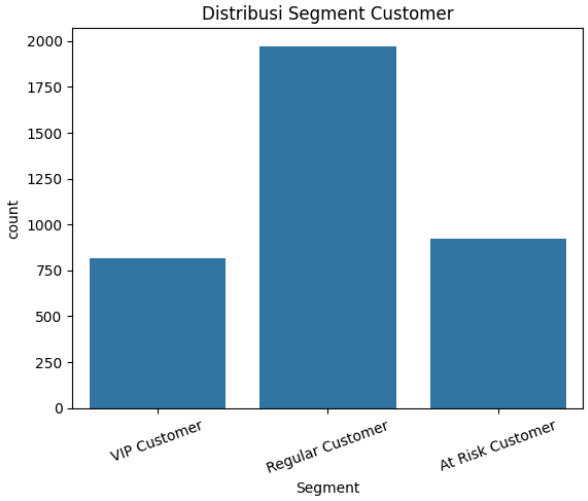
Berdasarkan Tabel 2 dan Gambar 4, cluster 2 memiliki nilai *Frequency* dan *Monetary* tertinggi dibandingkan cluster lainnya. Kondisi tersebut menunjukkan bahwa pelanggan pada cluster ini melakukan transaksi lebih sering dan memiliki total nilai pembelian yang lebih besar. Oleh karena itu, cluster 2 dikategorikan sebagai **VIP Customer**.

Cluster 0 menunjukkan nilai *Recency*, *Frequency*, dan *Monetary* yang berada pada tingkat menengah. Karakteristik tersebut menggambarkan pelanggan yang masih aktif melakukan transaksi namun belum memiliki kontribusi sebesar kelompok VIP. Dengan demikian, cluster 0 dikategorikan sebagai **Regular Customer**.

Sementara itu, cluster 1 memiliki nilai *Recency* tertinggi dan nilai *Frequency* serta *Monetary* yang relatif rendah. Hal ini menunjukkan bahwa pelanggan pada cluster tersebut sudah cukup lama tidak melakukan transaksi dan memiliki tingkat aktivitas yang rendah. Oleh karena itu, cluster 1 dikategorikan sebagai **At Risk Customer**.

6. Distribusi Segmen Pelanggan

Hasil segmentasi pelanggan selanjutnya direpresentasikan dalam bentuk kategori pelanggan yang lebih mudah dipahami.



Gambar 5. Distribusi Segmen Pelanggan

Berdasarkan Gambar 5, kelompok **Regular Customer** merupakan segmen dengan jumlah pelanggan terbanyak. Kondisi ini menunjukkan bahwa sebagian besar pelanggan masih berada pada tingkat aktivitas transaksi yang normal. Selain itu, terdapat kelompok **VIP Customer** yang memberikan kontribusi pembelian tinggi serta kelompok **At Risk Customer** yang memerlukan perhatian khusus karena berpotensi berhenti melakukan transaksi.

7. Implikasi dan Rekomendasi Bisnis

Hasil segmentasi pelanggan dapat dimanfaatkan untuk mendukung penyusunan strategi bisnis yang lebih terarah. Kelompok **VIP Customer** dapat dipertahankan melalui program loyalitas, pemberian penghargaan pelanggan, dan penawaran eksklusif. Kelompok **Regular Customer** dapat ditingkatkan nilainya melalui strategi *upselling* dan *cross-selling*. Sementara itu, kelompok **At Risk Customer** dapat menjadi sasaran program retensi seperti pemberian voucher, promosi khusus, maupun komunikasi pemasaran yang lebih intensif untuk mendorong pelanggan kembali melakukan transaksi.

D. PENUTUP

Simpulan

Penelitian ini berhasil menerapkan teknik clustering menggunakan algoritma K-Means berbasis RFM pada dataset Online Retail II untuk melakukan segmentasi pelanggan. Setelah melalui tahapan pembersihan data, pembentukan fitur RFM, penanganan outlier, dan normalisasi data, diperoleh 3.710 data pelanggan yang digunakan dalam proses clustering. Berdasarkan hasil Elbow Method, jumlah cluster yang paling sesuai adalah tiga cluster.

Evaluasi menggunakan Silhouette Score menghasilkan nilai sebesar 0,4596 yang menunjukkan kualitas clustering yang cukup baik. Hasil segmentasi menghasilkan tiga kelompok pelanggan, yaitu VIP Customer, Regular Customer, dan At Risk Customer. Kelompok VIP Customer memiliki frekuensi transaksi dan nilai pembelian tertinggi, sedangkan kelompok At Risk Customer menunjukkan aktivitas transaksi yang rendah dan berpotensi berhenti melakukan pembelian. Hasil penelitian ini dapat dimanfaatkan sebagai dasar dalam penyusunan strategi pemasaran yang lebih efektif dan terarah.

Saran

Penelitian selanjutnya dapat mengembangkan segmentasi pelanggan menggunakan dataset yang lebih terbaru atau menambahkan atribut lain yang dapat menggambarkan perilaku pelanggan secara lebih rinci. Selain itu, penggunaan algoritma clustering lain seperti K-Medoids, DBSCAN, atau Hierarchical Clustering dapat dilakukan untuk membandingkan kualitas hasil segmentasi yang diperoleh.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada **Ibu Mufidah Karimah, S.Kom., M.Kom.** selaku dosen pengampu mata kuliah Data Mining yang telah memberikan arahan, bimbingan, serta masukan selama proses pembelajaran dan pelaksanaan penelitian ini. Ucapan terima kasih juga disampaikan atas dukungan dan motivasi yang diberikan dalam penyusunan proyek segmentasi pelanggan menggunakan algoritma K-Means berbasis RFM pada dataset Online Retail II. Bimbingan yang diberikan sangat membantu penulis dalam memahami konsep data mining serta menyelesaikan penelitian ini dengan baik.

E. DAFTAR PUSTAKA

Alzami, F., Nurdin, N., & Kurniawan, A. (2023). *Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce*. ILKOM Jurnal Ilmiah, 15(1), 32–44.

Anitha, P., & Patil, M. M. (2020). *RFM Model for Customer Purchase Behavior Using K-Means Algorithm*. Journal of King Saud University - Computer and Information Sciences, 32(8), 863–870. https://doi.org/10.1016/j.jksuci.2019.12.011

Awaliyah, D. A. (2024). *Optimizing Customer Segmentation in Online Retail Using K-Means Clustering*. Scientific Journal of Informatics, 11(1), 45–56.

Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Nikmah, T. L. (2023). *Customer Segmentation Based on Loyalty Level Using K-Means and LRFM Feature Selection in Retail Online Store*. Jurnal ELTIKOM, 7(1), 21–28. https://doi.org/10.31961/eltikom.v7i1.648

Nugraha, A., Prasetyo, E., & Putra, R. (2025). *K-Means Clustering Interpretation Using Recency, Frequency, and Monetary Analysis*. TELKOMNIKA (Telecommunication Computing Electronics and Control), 23(1), 155–164.

Syahra, Y. (2025). *Customer Segmentation Using RFM and K-Means Clustering*. Sinkron: Jurnal dan Penelitian Teknik Informatika, 10(1), 1–12.

Tan, P. N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson Education.

Winaryanti, H. S., Suryanto, A., & Rahmawati, D. (2025). *Customer Segmentation Using K-Means Clustering with RFM Method*. Journal of Applied Intelligent System, 9(1), 1–12.

Zahro, G. W. N. W., Prasetyo, A., & Firmansyah, R. (2025). *Integration of RFM Method and K-Means Clustering for Customer Segmentation Effectiveness*. Dinda: Data Science and Information Technology Journal, 5(1), 12–21.