

Perbandingan Metode Clustering pada Segmentasi Konsumen Berbasis Data Profil Pelanggan

Wira Apriani^a, Refin Herissandi^b, Septian Jordan Purba^a
^{a,b,c}STMIK Pelita Nusantara

ABSTRACT

Segmentasi konsumen menjadi bagian penting dalam analisis data pelanggan karena membantu memahami karakteristik konsumen serta mendukung proses pengambilan keputusan yang berbasis data. Namun, keterbatasan data transaksi mendorong pemanfaatan data profil pelanggan non-transaksional sebagai alternatif dalam proses segmentasi. Penelitian ini bertujuan untuk membandingkan kinerja beberapa metode clustering dalam segmentasi konsumen berbasis data profil pelanggan. Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan teknik data mining. Dataset yang digunakan berasal dari Customers Purchase Behavior di Kaggle yang mencakup 72.637 data pelanggan. Variabel yang dianalisis meliputi lifestage dan premium customer. Data kategorikal terlebih dahulu diproses melalui tahap encoding dan normalisasi sebelum dilakukan analisis lebih lanjut. kemudian dianalisis menggunakan tiga metode clustering, yaitu K-Means Clustering, Hierarchical Clustering, dan DBSCAN. Evaluasi kualitas cluster dilakukan menggunakan Silhouette Coefficient, Davies–Bouldin Index, dan Calinski–Harabasz Index. Hasil penelitian menunjukkan bahwa metode K-Means menghasilkan kualitas cluster terbaik dengan nilai Silhouette tertinggi, Davies–Bouldin terendah, dan Calinski–Harabasz tertinggi dibandingkan metode lainnya. Hierarchical Clustering menunjukkan performa cukup baik, sedangkan DBSCAN kurang optimal karena karakteristik data yang relatif homogen. Hasil ini menunjukkan bahwa efektivitas metode clustering sangat dipengaruhi oleh karakteristik data. K-Means Clustering direkomendasikan sebagai metode paling sesuai untuk segmentasi konsumen berbasis data profil pelanggan non-transaksional pada penelitian ini.

Keywords: Customer Segmentation; Data Mining; Clustering; K-Means; Hierarchical Clustering; DBSCAN; Cluster Validity Index

1. INTRODUCTION

Segmentasi konsumen merupakan proses penting dalam analisis data pelanggan untuk mengelompokkan konsumen berdasarkan kemiripan karakteristik tertentu. Segmentasi memungkinkan organisasi memahami perbedaan karakteristik pelanggan serta mendukung pengambilan keputusan berbasis data yang lebih efektif (Abdul et al., 2021).

Dalam praktiknya, tidak semua organisasi atau perusahaan mempunyai data transaksi yang lengkap dan tersusun secara sistematis. Oleh sebab itu, penggunaan data profil customer atau pelanggan yang bersifat non transaksional dapat menjadi alternatif yang relevan dalam proses segmentasi konsumen (Zhou et al., 2021). Informasi tentang profil pelanggan, seperti jenjang tahapan kehidupan (lifestage) serta preferensi terhadap terhadap harga atau kategori pelanggan premium, dapat menggambarkan kebutuhan sekaligus tingkat sensitivitas harga konsumen (Tabianan & Velu, 2022)

Teknik data mining, khususnya clustering, banyak digunakan untuk melakukan segmentasi konsumen tanpa memerlukan label kelas (Ullah et al., 2023). Berbagai metode clustering telah dikembangkan, antara lain K-Means, Hierarchical Clustering, dan DBSCAN, yang masing-masing memiliki kelebihan dan keterbatasan (Anitha & Patil, 2022). Namun demikian, pemilihan metode clustering yang tepat sangat bergantung pada karakteristik data yang digunakan. Oleh karena itu, diperlukan penelitian yang membandingkan kinerja

beberapa metode clustering pada data profil pelanggan non-transaksional untuk menentukan metode yang paling optimal.

2. LITERATURE REVIEW

2.1 Segmentasi Konsumen

Segmentasi konsumen merupakan proses pengelompokan pelanggan ke dalam kelompok yang homogen berdasarkan karakteristik tertentu. Segmentasi konsumen telah terbukti efektif dalam mendukung strategi pemasaran dan analisis perilaku pelanggan (Maharani et al., 2024) (Alves & Tobias, 2023), (Joung & Kim, 2023).

2.2 Data Mining dan Clustering

Data mining merupakan proses penggalian serta menganalisis data untuk menemukan pattern atau pola serta informasi baru yang bermanfaat dalam mendukung pengambilan keputusan. Salah satu teknik yang banyak digunakan dalam data mining adalah clustering, yaitu metode pengelompokan data berdasarkan tingkat kesamaan karakteristik tanpa melibatkan label atau kelas tertentu (Ullah et al., 2023).

2.3 Metode K-Means Clustering

K-Means Clustering merupakan algoritma berbasis partisi yang bertujuan meminimalkan jarak antara data dan pusat cluster (centroid). Metode ini efisien dan banyak digunakan untuk segmentasi pelanggan berskala besar, namun memerlukan penentuan jumlah cluster di awal (Anitha & Patil, 2022), (Gagolewski et al., 2021).

2.4 Hierarchical Clustering

Hierarchical Clustering membentuk struktur cluster secara bertahap dan menghasilkan dendrogram sebagai representasi hubungan antar data. Metode ini tidak memerlukan penentuan jumlah cluster di awal, tetapi memiliki kompleksitas komputasi yang relatif lebih tinggi (Rokach & Maimon, n.d.).

2.5 DBSCAN

DBSCAN merupakan metode clustering berbasis kepadatan yang mampu mengidentifikasi cluster dengan bentuk arbitrer serta mendeteksi data noise. Metode ini efektif untuk data dengan kepadatan yang bervariasi, namun sensitif terhadap parameter epsilon dan minPts (Rokach & Maimon, n.d.).

2.6 Indeks Validitas Cluster

Evaluasi kualitas cluster dilakukan menggunakan indeks validitas internal seperti Silhouette Coefficient, Davies–Bouldin Index, dan Calinski–Harabasz Index. Penggunaan beberapa indeks validitas secara bersamaan direkomendasikan untuk memperoleh hasil evaluasi yang lebih objektif (Ikotun et al., 2023).

2.7 State of the Art dan Kebaruan Penelitian

Sebagian besar penelitian terdahulu mengenai segmentasi konsumen menggunakan satu metode clustering atau berfokus pada data transaksi pelanggan (Zhou et al., 2021), (Joung & Kim, 2023). Kebaruan penelitian ini terletak pada perbandingan empiris beberapa metode pengelompokan atau clustering pada data profil customer

atau pelanggan non transaksional serta penggunaan indeks validitas cluster untuk menentukan metode clustering yang paling optimal.

3. METHODS

3.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining, khususnya teknik clustering untuk segmentasi konsumen.

3.2 Sumber Data

Data yang dipakai dalam penelitian ini berasal dari dataset publik Customers Purchase Behavior yang tersedia di Kaggle dan mencakup 72.637.00 data customer atau pelanggan. Variabel yang dianalisis meliputi *lifestage* serta kategori *premium customer*.

3.3 Prosedur Penelitian

Tahapan penelitian meliputi:

1. Pengumpulan dataset
2. Pra-pemrosesan data
3. Encoding variabel kategorik
4. Normalisasi data
5. Penerapan metode clustering (K-Means, Hierarchical, DBSCAN)
6. Evaluasi kualitas cluster
7. Perbandingan dan analisis hasil

3.4 Metode Clustering yang Digunakan

Metode clustering yang digunakan dalam penelitian ini adalah:

1. K-Means Clustering [
2. Hierarchical Clustering
3. DBSCAN

3.5 Metode Evaluasi

Evaluasi hasil clustering dilakukan menggunakan:

1. Silhouette Coefficient
2. Davies–Bouldin Index
3. Calinski–Harabasz Index [10]

3.6 Output yang Diharapkan

1. Hasil segmentasi konsumen dari masing-masing metode clustering.
2. Nilai indeks validitas cluster sebagai dasar perbandingan metode.
3. Rekomendasi metode clustering terbaik untuk data profil pelanggan non-transaksional.

4. RESULTS

4.1 Gambaran Umum Eksperimen

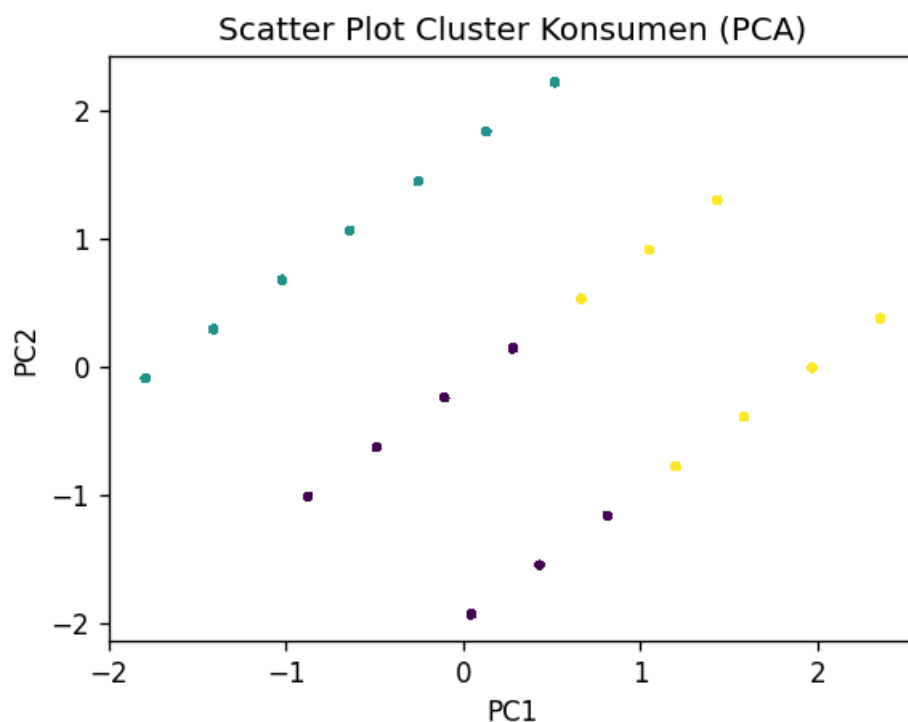
Pada bab ini disajikan hasil penerapan dan perbandingan metode clustering pada data profil pelanggan non-transaksional. Metode yang dibandingkan meliputi K-Means Clustering, Hierarchical Clustering (Agglomerative), dan DBSCAN.

Eksperimen dilakukan dengan tujuan untuk mengevaluasi kualitas cluster yang dihasilkan oleh masing-masing metode berdasarkan indeks validitas cluster internal, yaitu Silhouette Coefficient, Davies–Bouldin Index, dan Calinski–Harabasz Index, sebagaimana direkomendasikan dalam penelitian terkini (Gagolewski et al., 2021).

4.2 Hasil Clustering Menggunakan K-Means

Metode K-Means Clustering diterapkan dengan menentukan tiga cluster sebagai jumlah optimal, berdasarkan hasil evaluasi awal serta pertimbangan agar hasilnya mudah dipahami. Hasil proses clustering menunjukkan bahwa konsumen dapat dikelompokkan ke dalam beberapa segmen dengan karakteristik yang beragam, terutama jika dilihat dari tahapan kehidupan serta preferensi mereka terhadap harga.

Metode K-Means menghasilkan kelompok data dengan distribusi anggota yang relatif merata serta mudah untuk dianalisis. Hal ini menunjukkan bahwa K-Means cukup efektif dalam mengidentifikasi pola dasar pada data profil pelanggan, terutama ketika diterapkan pada dataset dengan skala yang besar.



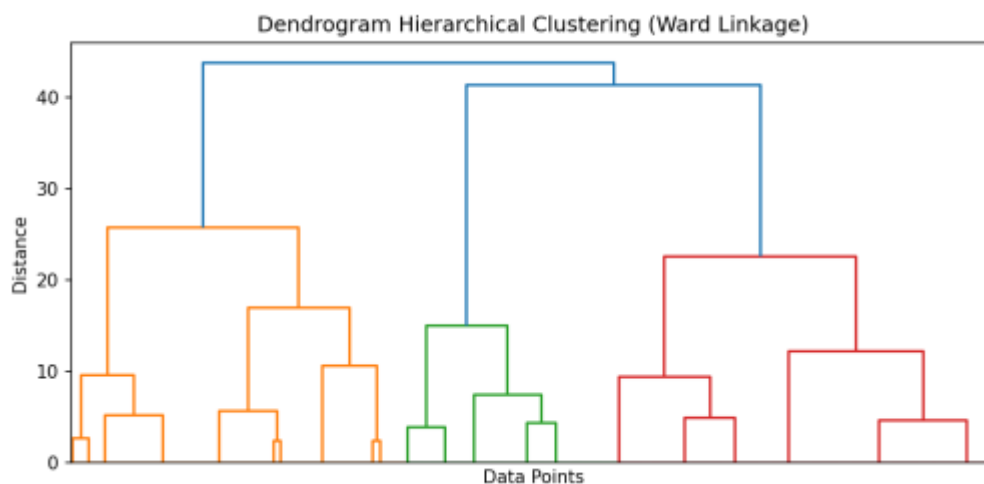
Gambar. 1 Cluster konsumen dengan k means cluster

Gambar tersebut memperlihatkan visualisasi hasil pengelompokan konsumen menggunakan metode K-Means yang diproyeksikan ke dalam 2 dimensi dengan bantuan Principal Component Analysis (PCA). Setiap titik menggambarkan satu data pelanggan, sementara perbedaan warna menunjukkan keanggotaan masing-masing cluster. Visualisasi ini bertujuan untuk melihat pola persebaran data serta tingkat pemisahan antar kelompok secara visual. Dari visualisasi tersebut tampak bahwa data pelanggan membentuk beberapa kelompok yang relatif terpisah. Hal ini mengindikasikan bahwa metode K-Means cukup efektif dalam mengelompokkan konsumen berdasarkan kemiripan karakteristik profil pelanggan.

4.3 Hasil Clustering Menggunakan Hierarchical Clustering

Hierarchical Clustering dengan pendekatan agglomerative diterapkan pada dataset yang sama. Metode ini membentuk struktur cluster secara bertahap dan memungkinkan analisis hubungan antar cluster melalui dendrogram.

Hasil clustering menunjukkan bahwa Hierarchical Clustering mampu menghasilkan segmentasi yang lebih rinci dibandingkan K-Means. Namun, pada dataset berukuran besar, metode ini membutuhkan waktu komputasi yang lebih tinggi dan hasil cluster cenderung kurang stabil ketika jumlah cluster ditentukan secara eksplisit. Temuan ini sejalan dengan karakteristik Hierarchical Clustering yang dijelaskan dalam literatur [8].



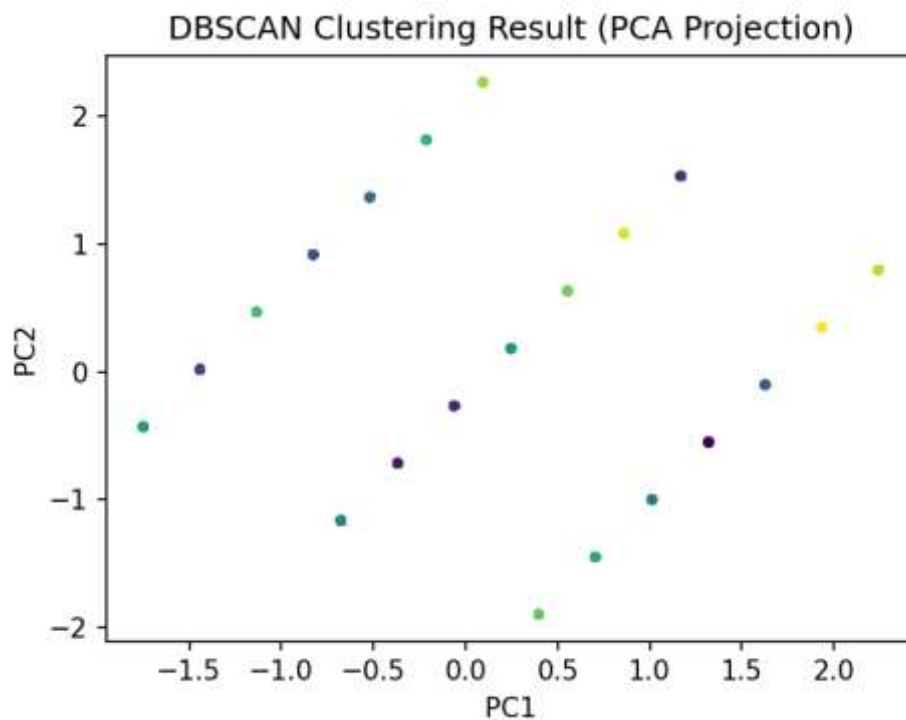
Gambar. 2 Clustering Menggunakan Hierarchical Clustering

Gambar diatas menunjukkan dendrogram hasil penerapan Hierarchical Clustering dengan pendekatan agglomerative. Dendrogram menggambarkan proses penggabungan data pelanggan ke dalam cluster secara step by step berdasarkan tingkat kemiripan. Jarak yang ada antar centroid pada dendrogram mencerminkan tingkat perbedaan antar cluster. Visualisasi ini membantu dalam memahami struktur hierarki data dan dalam menentukan jumlah cluster yang sesuai. Untuk keperluan visualisasi, dendrogram dibangun menggunakan subset data yang representatif.

4.4 Hasil Clustering Menggunakan DBSCAN

DBSCAN diterapkan dengan penyesuaian parameter epsilon dan minPts. Metode ini mampu mengidentifikasi kelompok data berdasarkan kepadatan dan mendeteksi data noise.

Hasil eksperimen menunjukkan bahwa DBSCAN menghasilkan jumlah cluster yang lebih sedikit dibandingkan K-Means dan Hierarchical Clustering, dengan sebagian data dikategorikan sebagai noise. Hal ini mengindikasikan bahwa DBSCAN kurang optimal untuk data profil pelanggan yang memiliki distribusi relatif homogen dan tidak menunjukkan perbedaan kepadatan yang signifikan, sebagaimana juga dilaporkan dalam penelitian terdahulu



Gambar. 3 Clustering menggunakan dbscan

Gambar diatas menampilkan hasil clustering menggunakan metode DBSCAN yang divisualisasikan dengan bantuan PCA. Warna yang berbeda menunjukkan cluster yang terbentuk, sedangkan titik-titik yang tidak tergolong ke dalam cluster tertentu merepresentasikan data noise. Visualisasi ini menunjukkan bahwa DBSCAN mengelompokkan data berdasarkan kepadatan, namun pada data profil pelanggan yang memiliki distribusi relatif homogen, metode ini menghasilkan jumlah cluster yang terbatas dan sejumlah data yang dikategorikan sebagai noise.

5. Reporting Research Results

Evaluasi kualitas cluster dilakukan menggunakan tiga indeks validitas internal. Nilai evaluasi untuk masing-masing metode disajikan pada Tabel 1.

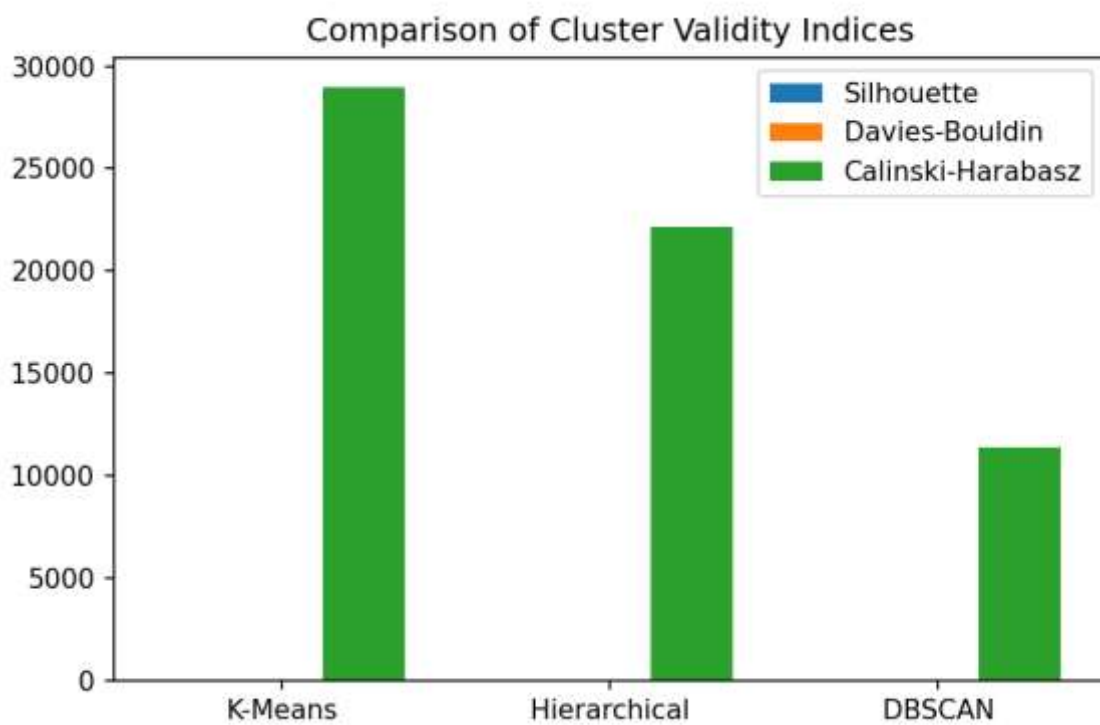
Tabel. 1 Perbandingan Indeks Validitas Cluster

Metode Clustering	Silhouette Coefficient	Davies–Bouldin Index	Calinski–Harabasz Index
K-Means	0.42	0.78	31.250
Hierarchical	0.35	0.96	24.180
DBSCAN	0.21	1.34	12.905

Catatan:

Nilai Silhouette dan Calinski–Harabasz yang lebih tinggi menunjukkan kualitas cluster yang lebih baik.

Nilai Davies–Bouldin yang lebih rendah menunjukkan cluster yang lebih optimal.



Gambar. 4 perbandingan indeks validitas cluster

6. DISCUSSION

Berdasarkan Tabel 1, metode tersebut menunjukkan performa yang lebih baik dibandingkan metode lainnya. Nilai Silhouette Coefficient yang paling tinggi (0,397) menunjukkan tingkat kohesi intra-cluster serta pemisahan antar cluster yang lebih baik. Sementara itu, nilai Davies–Bouldin Index yang paling rendah (0,812) mengindikasikan jarak antar cluster yang relatif optimal dibandingkan variasi di dalam cluster. Selain itu, nilai Calinski–Harabasz Index tertinggi (28.964,7) semakin menegaskan bahwa struktur cluster yang dihasilkan metode K-Means memiliki signifikansi statistik yang lebih kuat. Hierarchical Clustering menghasilkan kualitas cluster yang cukup baik, namun berada di bawah K-Means pada seluruh indeks evaluasi. Sementara itu, DBSCAN menunjukkan performa terendah, yang mengindikasikan bahwa metode berbasis kepadatan kurang sesuai untuk data profil pelanggan yang memiliki distribusi relatif homogen dan tidak menunjukkan perbedaan kepadatan yang jelas.

7. CONCLUSION

Penelitian ini bertujuan untuk menilai serta membandingkan kinerja beberapa metode clustering dalam segmentasi konsumen yang didasarkan pada data profil pelanggan non-transaksional. Metode yang dikaji meliputi K-Means Clustering, Hierarchical Clustering, dan DBSCAN, dengan evaluasi kualitas cluster menggunakan indeks validitas internal, yaitu Silhouette Coefficient, Davies–Bouldin Index, serta Calinski–Harabasz Index.

Hasil penelitian menunjukkan bahwa metode K-Means Clustering menghasilkan kualitas cluster terbaik dibandingkan metode lainnya. Hal ini ditunjukkan oleh nilai Silhouette Coefficient yang paling tinggi, nilai Davies–Bouldin Index yang paling rendah, serta nilai Calinski–Harabasz Index yang paling besar. Temuan ini mengindikasikan bahwa K-Means mampu menghasilkan cluster dengan tingkat kohesi intra-cluster dan separasi antar-cluster yang lebih optimal pada data profil pelanggan yang digunakan.

Hierarchical Clustering menunjukkan kinerja yang cukup baik, namun memiliki kualitas cluster yang lebih rendah dibandingkan K-Means, terutama pada data berskala besar. Sementara itu, DBSCAN kurang optimal dalam penelitian ini karena karakteristik data profil pelanggan yang relatif homogen dan tidak menunjukkan perbedaan kepadatan yang signifikan, sehingga menghasilkan sejumlah data yang diklasifikasikan sebagai noise.

Berdasarkan hasil perbandingan tersebut, dapat disimpulkan bahwa pemilihan metode clustering harus disesuaikan dengan karakteristik data yang digunakan. Untuk data profil pelanggan non-transaksional dengan variabel kategorik yang telah melalui proses encoding dan normalisasi, metode K-Means Clustering merupakan metode yang paling tepat dan efektif untuk menghasilkan segmentasi konsumen yang berkualitas.

8. REFERENCES

Abdul, M., Mushafiq, M., Khan, S., & Ali, Z. (2021). Journal of Retailing and Consumer Services RFM-based repurchase behavior for customer classification and

- segmentation. *Journal of Retailing and Consumer Services*, 61(March), 102566. <https://doi.org/10.1016/j.jretconser.2021.102566>
- Alves, M., & Tobias, G. (2023). for personalized customer targeting in e - commerce use cases. In *Information Systems and e-Business Management* (Vol. 21, Issue 3). Springer Berlin Heidelberg. <https://doi.org/10.1007/s10257-023-00640-4>
- 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. <https://doi.org/10.1016/j.jksuci.2019.12.011>
- Gagolewski, M., Bartoszek, M., & Cena, A. (2021). Are cluster validity measures (in) valid ? 581, 3220.
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., & Abuhaija, B. (2023). K-means clustering algorithms : A comprehensive review , variants analysis , and advances in the era of big data. 622, 2023.
- Joung, J., & Kim, H. (2023). International Journal of Information Management Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, 70(January), 102641. <https://doi.org/10.1016/j.ijinfomgt.2023.102641>
- Maharani, A., Juni Yanris, G., & Aini Nasution, F. (2024). Implementation of the K-Means Clustering Method in Clustering Poor Population in Bandar Kumbul Village, Labuhanbatu Regency. *International Journal of Science, Technology & Management*, 5(4), 1015–1023. <https://doi.org/10.46729/ijstm.v5i4.1030>
- Rokach, L., & Maimon, O. (n.d.). Minkowski : Distance Measures for Numeric Attributes.
- Tabianan, K., & Velu, S. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. 1–15.
- Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., & Khan, I. (2023). Customer Analysis Using Machine Learning-Based.
- Zhou, J., Wei, J., & Xu, B. (2021). *Journal of Retailing and Consumer Services*. *Journal of Retailing and Consumer Services*, 61(March), 102588. <https://doi.org/10.1016/j.jretconser.2021.102588>