

PERBANDINGAN K-MEANS DAN DEEP EMBEDDED CLUSTERING UNTUK SEGMENTASI PELANGGAN E-COMMERCE BERBASIS FEATURE EMBEDDING PRODUK

Saripudin¹, Martanto², Denni Pratama³, Raditya Danar Dana⁴.

Program Studi Teknik Informatika¹
Program Studi Manajemen Informatika²⁴
Program Studi Komputerisasi Akuntansi³

STMIK IKMI Cirebon
<https://ikmi.ac.id/page/18/?lang=de>
saripajja123@gmail.com

(*) Corresponding Author : saripajja123@gmail.com
Published : 30 Juni 2026

Abstract—Customer segmentation is an essential strategy in e-commerce data management to understand consumer behavior and support business decision-making, such as personalized promotions and the identification of high-value customers. However, conventional clustering methods such as K-Means have limitations in capturing non-linear relationships in complex and high-dimensional transaction data, which may lead to suboptimal segmentation results. Therefore, this study aims to quantitatively compare the performance of the K-Means algorithm and Deep Embedded Clustering (DEC) in customer segmentation by utilizing product feature embedding as a representation of customer features. The research methodology consists of data preprocessing, product embedding construction using Autoencoder, the application of K-Means and DEC algorithms for clustering, and cluster evaluation using an internal metric, namely the Silhouette Score. The results show that the Deep Embedded Clustering (DEC) method, with pretraining and joint optimization of embedding and clustering, produces more compact and well-separated clusters, achieving a higher Silhouette Score of 0.3729, compared to 0.3089 obtained by K-Means in high-dimensional embedding space. In addition, DEC demonstrates better stability with respect to initialization variations. Based on these findings, it can be concluded that the deep embedded clustering approach is more effective in generating customer segmentation for complex e-commerce datasets, although it requires higher computational costs and more extensive hyperparameter tuning.

Keywords: Customer Segmentation; Deep Embedded Clustering; K-Means; Feature Embedding; E-Commerce

Abstrak—Segmentasi pelanggan merupakan strategi penting dalam pengelolaan data e-commerce untuk memahami perilaku konsumen dan mendukung pengambilan keputusan bisnis, seperti personalisasi promosi serta identifikasi pelanggan bernilai tinggi. Namun, metode clustering konvensional seperti K-Means memiliki keterbatasan dalam menangkap hubungan non-linear pada data transaksi yang kompleks dan berdimensi tinggi, sehingga berpotensi menghasilkan segmentasi yang kurang optimal. Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa algoritma K-Means dan Deep Embedded Clustering (DEC) secara kuantitatif dalam melakukan segmentasi pelanggan dengan memanfaatkan feature embedding produk sebagai representasi fitur pelanggan. Metode penelitian meliputi tahap prapemrosesan data transaksi, pembangunan embedding produk menggunakan Autoencoder, penerapan algoritma K-Means dan DEC untuk proses clustering, serta evaluasi hasil kluster menggunakan metrik internal yaitu Silhouette Score. Hasil penelitian menunjukkan bahwa metode Deep Embedded Clustering (DEC) dengan pretraining dan optimasi joint embedding-clustering mampu menghasilkan kluster yang lebih kompak dan terpisah dengan nilai Silhouette yang lebih tinggi, yaitu sebesar 0,3729, dibandingkan dengan K-Means yang memperoleh nilai sebesar 0,3089 pada ruang embedding berdimensi tinggi. Selain itu, DEC juga menunjukkan stabilitas yang lebih baik terhadap variasi inisialisasi. Berdasarkan hasil tersebut, dapat disimpulkan bahwa pendekatan deep embedded clustering lebih efektif dalam menghasilkan segmentasi pelanggan pada dataset e-commerce yang kompleks, meskipun membutuhkan biaya komputasi dan proses tuning hyperparameter yang lebih besar.

Kata Kunci: Segmentasi Pelanggan; Deep Embedded Clustering; K-Means; Feature Embedding; E-Commerce.

INTRODUCTION

Perkembangan digitalisasi perdagangan telah mendorong bisnis untuk mencari cara yang lebih cerdas dan adaptif dalam memahami serta mengelompokkan pelanggan. Segmentasi pelanggan, yaitu proses membagi pasar yang heterogen menjadi subkelompok yang homogen, merupakan langkah dasar dalam personalisasi dan pemasaran prediktif (Abidar et al., 2025; Gomes & Meisen, 2023). Literatur terbaru menunjukkan bahwa penerapan *machine learning* (ML) telah menggantikan metode segmentasi statistik tradisional karena mampu mengungkap pola perilaku kompleks dalam *e-commerce* (Gallego et al., 2024; Kasem et al., 2024).

Penelitian pada periode 2021 hingga 2023 menunjukkan perkembangan metode clustering, khususnya pada berbagai varian K-Means dan pendekatan berbasis deep learning. Zhao et al. (2021) mengembangkan Extended Regularized K-Means untuk mengatasi tantangan data berdimensi tinggi. Wu et al. (2022) meningkatkan K-Medoids melalui optimasi ukuran jarak sehingga menghasilkan segmentasi yang lebih stabil. Ullah et al. (2023) mengintegrasikan pendekatan RFM dengan *machine learning* untuk mengidentifikasi pelanggan bernilai tinggi, sementara Mahfuza et al. (2022) memperkenalkan model LRFMV untuk segmentasi multivariabel. Tabianan et al. (2022) menggabungkan K-Means dengan data perilaku pelanggan serta dashboard analitik untuk meningkatkan interpretasi hasil [1], [2], [3], [4].

Integrasi antara *representation learning* dan *deep clustering* menjadi arah baru dalam pengembangan segmentasi pelanggan. Gallego et al. (2024) memanfaatkan *Non-Negative Matrix Factorization* (NMF) untuk menghasilkan representasi laten sebelum *clustering*, sedangkan Kasem et al. (2023) mengembangkan pipeline prediksi berbasis RFM *K-Means*. Tinjauan sistematis oleh Abidar et al. (2025) menegaskan efektivitas *machine learning* dalam segmentasi, namun juga menunjukkan keterbatasan penelitian yang mengkaji metode deep embedding pada data *e-commerce* yang kompleks. [5], [6], [7]

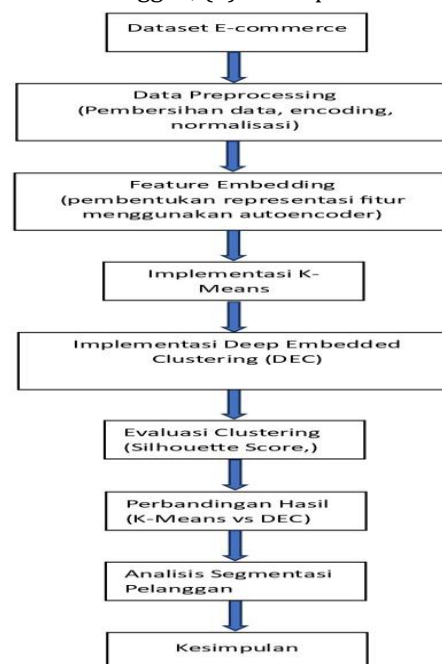
Walaupun berbagai kemajuan telah dicapai, sejumlah tantangan masih ditemukan, seperti tingginya dimensi data, sparsity fitur, serta penentuan jumlah *cluster* (K) yang optimal (Wu et al., 2022; Zhao et al., 2021). Selain itu, aspek interpretabilitas dan fleksibilitas model masih menjadi permasalahan utama (Gomes &

Meisen, 2023). Pendekatan berbasis deep learning melalui teknik *feature embedding* menawarkan potensi dalam merepresentasikan pola perilaku pelanggan secara lebih kompleks dan non-linear. Namun demikian, bukti empiris terkait efektivitas metode *deep embedded clustering*, khususnya dalam konteks data *e-commerce* berbasis *feature embedding* produk, masih terbatas dan belum banyak dibandingkan secara langsung dengan metode clustering konvensional seperti K-Means [8], [9], [10].

Kondisi ini menunjukkan adanya kesenjangan penelitian (*research gap*), terutama dalam mengevaluasi sejauh mana pendekatan *deep clustering* mampu meningkatkan kualitas segmentasi pelanggan dibandingkan metode tradisional pada data berdimensi tinggi. Oleh karena itu, penelitian ini dilakukan untuk membandingkan kinerja algoritma *K-Means* dan *Deep Embedded Clustering* (DEC) dalam segmentasi pelanggan *e-commerce* berbasis *feature embedding*.

MATERIALS AND METHODS

Penelitian ini mengikuti pendekatan eksperimental kuantitatif yang menerapkan alur kerja terintegrasi yang terdiri Beberapa tahapan: (1) Dataset *E-commerce*, (2) Data Preprocessing (Pembersihan data dan normalisasi), (3) *Feature Embedding* (pembentukan representasi fitur menggunakan *autoencoder*), (4) Implementasi *K-Means*, (5) Implementasi *Deep Embedded Clustering* (DEC) Evaluasi *Clustering* (*Silhouette Score*), (6) Perbandingan Hasil (*K-Means* vs DEC), (7) Analisis Segmentasi Pelanggan, (8) Kesimpulan.



Gambar 1. Alur Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang dimulai dari pengumpulan dataset pelanggan *e-commerce*, kemudian dilakukan proses *data preprocessing* untuk membersihkan dan menyiapkan data. Selanjutnya dilakukan pembentukan *Feature embedding* untuk merepresentasikan data pelanggan dalam bentuk vektor numerik. Setelah itu dilakukan proses *clustering* menggunakan algoritma *K-Means* dan *Deep Embedded Clustering (DEC)*. Hasil *clustering* kemudian dievaluasi menggunakan metrik evaluasi *clustering* dan dibandingkan untuk mengetahui metode yang menghasilkan segmentasi pelanggan yang lebih baik.

RESULTS AND DISCUSSION

Algoritma *K-Means* diterapkan pada data hasil *preprocessing*. Hasil *clustering* menunjukkan bahwa pembentukan *cluster* masih belum optimal, yang ditandai dengan adanya *overlap* antar *cluster* pada visualisasi data.

Hasil implementasi menunjukkan bahwa metode *K-Means* menghasilkan *cluster* yang masih tumpang tindih, sedangkan DEC menghasilkan *cluster* yang lebih kompak dan terpisah dengan jelas.

--- TAHAP 5: IMPLEMENTASI CLUSTERING ---

[TABEL] Sampel Hasil Pelabelan Pelanggan:

CustomerID	Recency	Frequency	Monetary	KMeans_Labels	DEC_Labels
12346.0	326	1	77183.60	0	3
12347.0	2	182	4310.00	2	3
12348.0	75	31	1797.24	0	1
12349.0	19	73	1757.55	0	1
12350.0	310	17	334.40	3	2
12352.0	36	85	2506.04	0	1
12353.0	204	4	89.00	3	2
12354.0	232	58	1079.40	0	1
12355.0	214	13	459.40	3	2
12356.0	23	59	2811.43	0	1

Gambar 2 Implementasi Clustering

Gambar 2 menunjukkan adanya perbedaan pelabelan pelanggan antara *K-Means* dan DEC, yang menunjukkan bahwa DEC lebih adaptif dalam mengelompokkan data.

Nilai *Silhouette Score* yang diperoleh sebesar 0.3089, yang menunjukkan bahwa kualitas *cluster* berada pada kategori sedang. Hal ini mengindikasikan bahwa *K-Means* memiliki keterbatasan dalam menangani data berdimensi tinggi dan hubungan *non-linear*.

Perbandingan performa kedua algoritma dapat dilihat pada tabel berikut:

Tabel 1 Perbandingan Kedua Algoritma

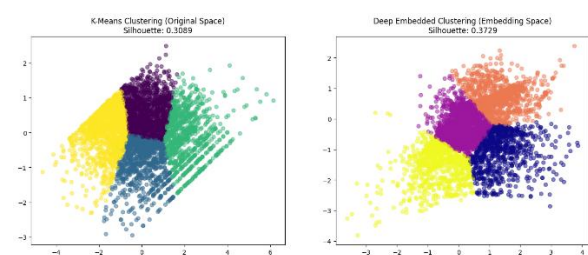
Peneliti	Metode	Hasil
----------	--------	-------

[11]	K-Means + RFM	Menghasilkan 4 cluster dengan kontribusi revenue tertinggi $\pm 60\%$ pada satu segmen.
Penelitian ini (K-means)	K-Means	0.3089
Penelitian ini (DEC)	Deep Embedded Clustering	0.3729

Berdasarkan Tabel 1 penelitian terdahulu oleh Yang (2025) menunjukkan bahwa metode *K-Means* dengan pendekatan RFM mampu melakukan segmentasi pelanggan, namun memiliki keterbatasan dalam menangkap pola non-linear pada data. Hal ini sejalan dengan hasil penelitian ini, di mana *K-Means* menghasilkan nilai *Silhouette Score* yang masih relatif rendah.

Sebaliknya, metode *Deep Embedded Clustering (DEC)* yang digunakan dalam penelitian ini menunjukkan performa yang lebih baik, dengan nilai *Silhouette Score* yang lebih tinggi serta pemisahan *cluster* yang lebih jelas. Hal ini menunjukkan bahwa penggunaan *feature embedding* mampu meningkatkan kualitas segmentasi pelanggan *e-commerce*.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa metode DEC memiliki performa yang lebih baik dibandingkan *K-Means* dalam menghasilkan segmentasi pelanggan yang optimal.



Gambar 3 Perbandingan hasil

Gambar 3 tersebut memperlihatkan perbandingan hasil *clustering* menggunakan metode *K-Means* pada ruang data asli (*original space*) dan *Deep Embedded Clustering (DEC)* pada ruang *embedding (embedding space)*, yang divisualisasikan dalam bentuk scatter plot dua dimensi.

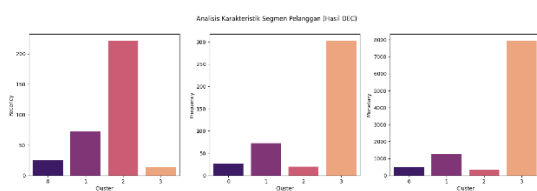
Pada hasil *K-Means*, terlihat bahwa pembagian *cluster* masih cenderung tumpang tindih antar kelompok. Meskipun terdapat pemisahan warna yang menunjukkan empat *cluster* batas antar *cluster* belum terbentuk secara tegas. Hal ini

mengindikasikan bahwa *K-Means* mengalami kesulitan dalam memisahkan data yang memiliki distribusi kompleks pada ruang asli. Nilai *Silhouette Score* sebesar 0.3089 juga menunjukkan bahwa kualitas *clustering* masih berada pada kategori sedang, dengan tingkat pemisahan antar *cluster* yang belum optimal.

Sebaliknya, pada hasil *Deep Embedded Clustering* (DEC), pola sebaran data menunjukkan pemisahan *cluster* yang lebih jelas dan lebih kompak. Setiap *cluster* tampak lebih terkonsentrasi dan memiliki batas yang relatif tegas dibandingkan dengan *K-Means*. Hal ini disebabkan oleh penggunaan *embedding* yang mampu merepresentasikan struktur laten data dengan lebih baik sebelum proses *clustering* dilakukan. Nilai *Silhouette Score* yang lebih tinggi, yaitu 0.3729, mengindikasikan bahwa DEC menghasilkan kualitas *clustering* yang lebih baik, baik dari segi kohesi (kedekatan dalam *cluster*) maupun separasi (jarak antar *cluster*).

Secara keseluruhan, visualisasi ini menunjukkan bahwa metode DEC lebih unggul dibandingkan *K-Means* dalam menangani data dengan pola kompleks, karena mampu menghasilkan segmentasi yang lebih terstruktur, terpisah dengan baik, dan memiliki kualitas *clustering* yang lebih tinggi.

Hasil segmentasi menunjukkan bahwa setiap *cluster* memiliki karakteristik pelanggan yang berbeda berdasarkan nilai RFM. Perbedaan ini dapat dimanfaatkan untuk menyusun strategi pemasaran yang lebih tepat sasaran, seperti identifikasi pelanggan loyal, pelanggan berisiko *churn*, serta penentuan strategi promosi.



Gambar 4 Analisis segmentasi pelanggan

Gambar 4 tersebut menampilkan hasil analisis karakteristik segmen pelanggan menggunakan metode *Deep Embedded Clustering* (DEC) berdasarkan tiga indikator utama, yaitu *recency*, *frequency*, dan *monetary* (RFM). Pada aspek *recency*, *cluster* 2 memiliki nilai tertinggi, yang menunjukkan bahwa pelanggan dalam kelompok ini relatif paling lama tidak melakukan transaksi, sehingga dapat dikategorikan sebagai pelanggan tidak aktif. Sebaliknya, *cluster* 3 memiliki nilai *recency* paling rendah, menandakan bahwa pelanggan dalam kelompok ini baru saja melakukan transaksi dan tergolong aktif.

Pada indikator *frequency*, *cluster* 3 menunjukkan nilai paling tinggi dibandingkan *cluster* lainnya, yang mengindikasikan bahwa pelanggan dalam segmen ini paling sering melakukan pembelian. Sementara itu, *cluster* 2 memiliki nilai *frequency* yang rendah, memperkuat indikasi bahwa kelompok ini kurang aktif dalam bertransaksi.

Dari sisi *monetary*, *cluster* 3 kembali mendominasi dengan nilai yang sangat tinggi, menunjukkan bahwa pelanggan dalam segmen ini memberikan kontribusi pendapatan terbesar. *Cluster* 1 berada pada tingkat menengah baik dalam *frequency* maupun *monetary*, sehingga dapat dikategorikan sebagai pelanggan potensial. Adapun *cluster* 0 menunjukkan nilai yang relatif rendah pada ketiga indikator, yang mencerminkan pelanggan dengan tingkat aktivitas dan kontribusi yang rendah.

Secara keseluruhan, hasil ini menunjukkan bahwa *cluster* 3 merupakan segmen pelanggan paling bernilai (*high-value customers*), *cluster* 1 sebagai pelanggan potensial, *cluster* 0 sebagai pelanggan dengan kontribusi rendah, dan *cluster* 2 sebagai pelanggan yang cenderung tidak aktif atau berisiko *churn*.

CONCLUSION

Berdasarkan hasil penelitian dan pembahasan mengenai perbandingan algoritma *K-Means* dan *Deep Embedded Clustering* (DEC) untuk segmentasi pelanggan *e-commerce* berbasis *feature embedding*, maka dapat ditarik beberapa kesimpulan sebagai berikut:

Algoritma *K-Means* memiliki keterbatasan dalam menangani data pelanggan *e-commerce* yang bersifat kompleks, berdimensi tinggi, dan memiliki hubungan non-linear. Hal ini ditunjukkan oleh hasil *clustering* yang masih mengalami *overlap* antar *cluster* serta nilai *Silhouette Score* sebesar 0,3089 yang menunjukkan kualitas cluster pada kategori sedang.

Penerapan *feature embedding* menggunakan *autoencoder* terbukti mampu meningkatkan representasi data pelanggan. Model *autoencoder* berhasil melakukan reduksi dimensi dan mempertahankan informasi penting dalam bentuk representasi laten, yang ditunjukkan oleh proses pelatihan yang stabil tanpa indikasi *overfitting*.

Metode *Deep Embedded Clustering* (DEC) menunjukkan performa yang lebih baik dibandingkan *K-Means* dalam proses segmentasi pelanggan. Hal ini ditunjukkan oleh nilai *Silhouette Score* sebesar 0,3729, yang lebih tinggi dibandingkan *K-Means*, serta hasil *cluster* yang lebih kompak dan terpisah secara jelas.

Deep Embedded Clustering (DEC) mampu menangkap pola hubungan *non-linear* antar variabel pelanggan, khususnya pada atribut *Recency*, *Frequency*, dan *Monetary* (RFM), sehingga menghasilkan segmentasi yang lebih representatif dan informatif dibandingkan metode *K-Means* yang berbasis jarak.

REFERENCE

- [1] E. Suprpto, "Pengelompokan Potensi Padi di Indonesia Menggunakan K-Means Cluster," *Jurnal Ilmiah Populer*, vol. 5, 2021.
- [2] D. Ramdhan, G. Dwilestari, R. D. Dana, A. Ajiz, and K. Kaslani, "Clustering Data Persediaan Barang Dengan Menggunakan Metode K-Means," *MEANS (Media Informasi Analisa dan Sistem)*, vol. 7, no. 1, pp. 1–9, 2022, doi: 10.54367/means.v7i1.1826.
- [3] M. G. Pradana, A. C. Nurcahyo, P. H. Saputro, U. A. Ata, and U. A. Ata, "Penerapan Metode K-Means Klustering untuk Menentukan Kepuasan Pelanggan," *Citec Journal*, vol. 7, no. 1, pp. 42–50, 2020.
- [4] Y. A. Wijaya *et al.*, "K-MEANS DI SEKOLAH MENENGAH KEJURUAN WAHIDIN KOTA CIREBON," vol. 6, no. 2, pp. 552–559, 2022.
- [5] Y. Dharma Putra, M. Sudarma, and I. B. A. Swamardika, "Clustering History Data Penjualan Menggunakan Algoritma K-Means," *Majalah Ilmiah Teknologi Elektro*, vol. 20, no. 2, p. 195, 2021, doi: 10.24843/mite.2021.v20i02.p03.
- [6] D. Anggarwati, O. Nurdiawan, I. Ali, and D. A. Kurnia, "Penerapan Algoritma K-Means Dalam Prediksi Penjualan Karoseri," vol. 1, no. 2, pp. 58–62, 2021.
- [7] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 145, pp. 4–16, 2021.
- [8] Y. Hartati, S. Defit, and G. W. Nurcahyo, "Klasterisasi Bibit Terbaik Menggunakan Algoritma K-Means dalam Meningkatkan Penjualan," *Jurnal Informatika Ekonomi Bisnis*, vol. 3, pp. 4–10, 2021, doi: 10.37034/infeb.v3i1.56.
- [9] I. Reisandi, Daryana, F. Sri Mulyati, and M. Fauzi, "Implementasi Clustering K-Means Terhadap Penilaian Kinerja Karyawan PT XYZ," *Jurnal Sosial dan Teknologi (SOSTECH)*, vol. 1, no. 8, pp. 757–767, 2021.
- [10] D. Gustian and M. S. Al-Farits, "Data Mining Untuk Melihat Minat Belajar Siswa Menerapkan Metode K-Means," *Journal of Information System Research (JOSH)*, vol. 4, no. 3, pp. 775–784, Apr. 2023, doi: 10.47065/josh.v4i3.3218.
- [11] Y. Yang, "E-commerce Customer Segmentation Using Machine Learning," vol. 0, pp. 102–109, 2025, doi: 10.54254/2754-1169/2025.BL29390.