VOL. 3. NO. 02 MARET 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

ANALISIS CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS UNTUK PENGELOMPOKAN DATA PENJUALAN E-COMMERCE

Gilang Saputra[1], Cep Lukman Rohmat[2], Arif Rinaldi Dikananda[3], Willy Prihartono[4].

Program Studi Teknik Informatika[1]
Program Studi Rekayasa Perangkat Lunak[23]
Program Studi Komputerisasi Akuntansi[4]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
gilangipan281@gmail.com

(*) Corresponding Author : gilangipan281@gmail.com
Published : 30 Maret 2026

Abstract— *The rapid growth of the E-Commerce industry generates massive volumes of transaction data daily, containing hidden patterns regarding purchasing behavior and product profitability crucial for business. However, companies often face challenges in processing this high-dimensional raw data into strategic insights, leading to inefficiencies in marketing allocation and stock management. This study aims to implement Data Mining techniques using the K-Means Clustering algorithm to group sales data based on transaction characteristics, adopting the Knowledge Discovery in Databases (KDD) methodology which includes data cleaning, feature selection, and standardization using Standard Scaler. Key attributes include product price, discount, quantity, total revenue, shipping cost, profit margin, and customer age. The optimal number of clusters (K) was determined by comparing the Elbow method and Silhouette Score. Evaluation results established K=4 as the optimal number of clusters, mapping data into: (1) "Premium/High Value" Cluster (highest average spending and margin, small volume); (2) "Mass Market" Cluster (largest volume, low value); (3) "Mid-Tier" Cluster (middle category); and (4) "Discount Seekers" Cluster (high discount dominance). It is concluded that K-Means is effective in separating transaction patterns into actionable segments, providing a foundation for management to design more targeted strategies, such as exclusive loyalty programs and efficient promotions.*

Keywords: K-Means Clustering, Data Mining, E-Commerce, KDD, Sales Segmentation.

Abstrak—Pesatnya pertumbuhan industri *E-Commerce* menghasilkan volume data transaksi yang masif, menyimpan pola tersembunyi mengenai perilaku pembelian dan profitabilitas produk yang krusial bagi bisnis. Namun, perusahaan sering kali menghadapi kendala dalam mengolah data mentah berdimensi tinggi ini menjadi wawasan strategis, mengakibatkan ketidakefisienan dalam alokasi pemasaran dan manajemen stok. Penelitian ini bertujuan untuk mengimplementasikan teknik *Data Mining* menggunakan algoritma *K-Means Clustering* guna mengelompokkan data penjualan berdasarkan karakteristik transaksi, mengadopsi metodologi *Knowledge Discovery in Databases* (KDD) yang meliputi pembersihan data, seleksi fitur, dan standardisasi menggunakan *Standard Scaler*. Atribut utama meliputi harga produk, diskon, kuantitas, total pendapatan, biaya pengiriman, margin keuntungan, dan usia pelanggan. Penentuan jumlah *cluster* optimal (K) dilakukan dengan membandingkan metode *Elbow* dan *Silhouette Score*. Hasil evaluasi menetapkan K=4 sebagai jumlah *cluster* paling optimal, memetakan data ke dalam: (1) *Cluster* "Premium/High Value" (rata-rata belanja dan margin tertinggi, volume kecil); (2) *Cluster* "Mass Market" (volume terbesar, nilai rendah); (3) *Cluster* "Mid-Tier" (kategori menengah); dan (4) *Cluster* "Discount Seekers" (didominasi diskon tinggi). Disimpulkan bahwa *K-Means* efektif dalam memisahkan pola transaksi menjadi segmen yang dapat ditindaklanjuti, memberikan landasan bagi manajemen untuk merancang strategi yang lebih terarah, seperti program loyalitas eksklusif dan promosi efisien.

Kata Kunci : K-Means Clustering, Data Mining, E-Commerce, KDD, Segmentasi Penjualan.

INTRODUCTION

Perkembangan e-commerce yang semakin pesat telah mengubah ekosistem bisnis digital secara signifikan, menghasilkan volume data transaksi yang sangat besar dan kompleks. Data ini mencerminkan pola perilaku pelanggan yang

dinamis dan tidak selalu mudah dipahami tanpa teknik analisis yang tepat. Dalam kondisi persaingan yang semakin kompetitif, perusahaan perlu memanfaatkan data penjualan untuk mengidentifikasi pola pembelian, memprediksi kebutuhan pasar, dan merancang strategi pemasaran yang lebih efektif. Analisis clustering menjadi salah satu pendekatan penting yang digunakan untuk menggali struktur tersembunyi dalam data tanpa label, sehingga perusahaan dapat mengelompokkan pelanggan atau produk berdasarkan kesamaan karakteristik tertentu.

Teknik machine learning, khususnya metode clustering, berperan penting dalam memahami perilaku pelanggan pada platform e-commerce. Di antara berbagai algoritma clustering, K-Means merupakan salah satu metode yang paling banyak digunakan karena kesederhanaan, efisiensi komputasi, serta kemampuannya menangani dataset berukuran besar. Penelitian menunjukkan bahwa K-Means mampu melakukan segmentasi pelanggan berdasarkan pola pembelian dan preferensi yang serupa, sehingga membantu perusahaan menyusun strategi pemasaran yang lebih terarah [1]. Keunggulan ini membuat K-Means menjadi pilihan yang relevan dalam analisis data transaksi digital.

Meskipun demikian, penerapan K-Means pada data berdimensi tinggi seringkali menghadapi tantangan, terutama ketika variabel-variabel saling berkorelasi. Karakteristik ini umum ditemukan dalam data e-commerce yang kaya akan fitur perilaku dan aktivitas pelanggan. Sejumlah penelitian mengusulkan variasi algoritma K-Means dengan pendekatan regularisasi untuk meningkatkan stabilitas dan akurasi hasil clustering ketika menangani data yang kompleks tersebut [2]. Hal ini menunjukkan bahwa metode clustering perlu beradaptasi dengan dinamika dan kebutuhan analisis pada lingkungan data modern.

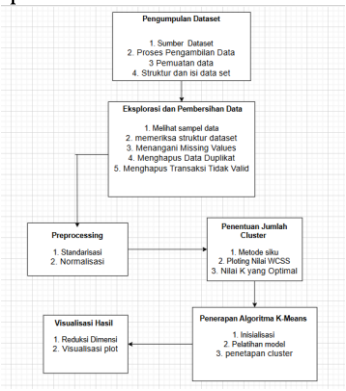
Selain itu, analisis nilai pelanggan menjadi bagian penting dari strategi bisnis e-commerce. Model analisis seperti RFM yang dipadukan dengan algoritma K-Means++ terbukti meningkatkan ketepatan dalam mengidentifikasi kelompok pelanggan bernilai tinggi maupun rendah [3]. Pendekatan ini memperkuat pemahaman bahwa clustering tidak hanya berfungsi untuk mengelompokkan data, tetapi juga sebagai alat strategis untuk mengidentifikasi nilai ekonomi pelanggan dan merancang intervensi pemasaran yang lebih efisien. Kombinasi teknik-teknik tersebut juga menunjang interpretasi data yang lebih akurat pada lingkungan dengan pola transaksi yang tidak stabil.

Dalam konteks kualitas analisis, penelitian terbaru menekankan pentingnya penggunaan kerangka evaluasi yang mempertimbangkan konteks bisnis dan karakteristik data. Perbandingan beberapa algoritma clustering menunjukkan bahwa efektivitas suatu metode tidak hanya ditentukan oleh hasil statistik, tetapi juga kesesuaiannya dengan tujuan segmentasi yang ingin dicapai [4]. Pendekatan ini menegaskan perlunya analisis yang lebih komprehensif untuk memastikan bahwa hasil clustering dapat memberikan kontribusi nyata bagi pengambilan keputusan strategis di perusahaan.

Di ranah pemasaran digital, teknik pembelajaran mesin seperti Non-negative Matrix Factorization (NMF) dan K-Means terus mendapatkan perhatian karena kemampuannya mengungkap pola laten dalam perilaku pelanggan [5]. Tren ini menunjukkan bahwa penggunaan algoritma clustering semakin menjadi kebutuhan dasar dalam pengelolaan data e-commerce. Seiring dengan meningkatnya kompleksitas data penjualan, analisis clustering berbasis K-Means menjadi semakin relevan untuk memberikan pemahaman yang mendalam mengenai struktur data serta mendukung strategi bisnis berbasis data. Oleh karena itu, penelitian mengenai pengelompokan data penjualan e-commerce menggunakan algoritma K-Means menjadi penting dan memiliki kontribusi signifikan baik bagi akademisi maupun praktisi.

MATERIALS AND METHODS

Desain penelitian yang digunakan adalah Clustering (Segmentasi Data), sebuah teknik *Unsupervised Machine Learning*. Algoritma utama yang diterapkan adalah K-Means Clustering.



Gambar 1. Alur Penelitian

Tahap pengumpulan dataset berfokus pada pemuatan data dan pembentukan struktur data awal yang akan digunakan dalam proses klastering. Data yang digunakan merupakan data sekunder yang diperoleh dari website Kaggle dan diimpor dalam bentuk file *ecommerce_sales_34500.csv*.

Proses pemuatan data dilakukan menggunakan library Pandas ke dalam sebuah *dataframe* agar data dapat dikelola dan dianalisis secara efisien.

Setelah data berhasil dimuat, dilakukan tahap persiapan fitur dengan memilih variabel-variabel kuantitatif yang relevan untuk proses klastering. Variabel yang digunakan meliputi *price, discount, quantity, total_amount, shipping_cost, profit_margin,* dan *customer_age*. Pemilihan fitur ini bertujuan untuk merepresentasikan karakteristik transaksi dan pelanggan secara numerik sehingga dapat diproses oleh algoritma K-Means.

Tahap eksplorasi dan pembersihan data bertujuan untuk memastikan kualitas data sebelum dilakukan pemodelan. Eksplorasi awal dilakukan menggunakan fungsi *.head()* dan *.info()* untuk memahami struktur data, tipe variabel, serta jumlah entri yang tersedia. Selanjutnya, dilakukan proses pembersihan data yang mencakup pengecekan dan penanganan *missing values* serta pengendalian *outliers* ekstrem pada variabel kuantitatif guna menghindari bias dalam hasil klastering.

Karena algoritma K-Means sangat sensitif terhadap skala data, tahap *preprocessing* menjadi langkah yang wajib dilakukan. Data numerik yang telah dipilih distandarisasi menggunakan metode *StandardScaler* dari library Scikit-learn. Proses ini mengubah data sehingga memiliki nilai rata-rata nol dan deviasi standar satu, sesuai dengan rumus Z=(x−μ)/σZ = (x - \mu) / \sigmaZ=(x−μ)/σ, sehingga setiap variabel memiliki kontribusi yang seimbang dalam perhitungan jarak Euclidean.

Sebelum menerapkan algoritma K-Means, jumlah klaster optimal ditentukan menggunakan metode *Elbow*. Metode ini menghitung nilai *Within-Cluster Sum of Squares* (WCSS) untuk berbagai nilai K, misalnya dari K=1 hingga K=10. Nilai WCSS kemudian diplot ke dalam grafik, dan jumlah klaster optimal ditentukan pada titik di mana penurunan WCSS mulai melambat secara signifikan, membentuk pola menyerupai siku (*elbow*).

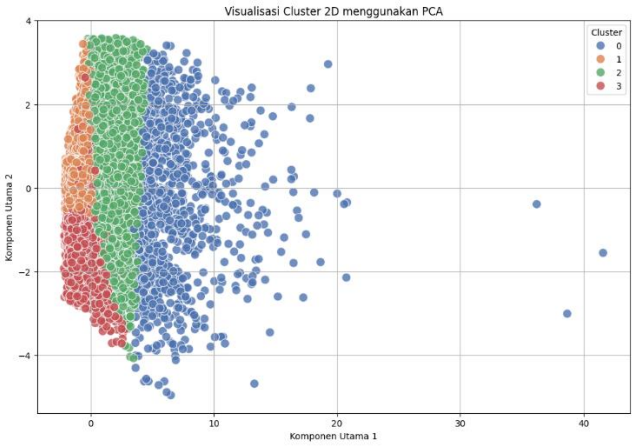
Setelah jumlah klaster optimal diperoleh, algoritma K-Means diterapkan dengan inisialisasi menggunakan metode *k-means++* untuk meningkatkan kualitas pusat klaster awal. Model dilatih menggunakan data yang telah distandarisasi, dan setiap data kemudian diberi label klaster sesuai hasil pengelompokan. Untuk memvisualisasikan hasil klastering, digunakan teknik *Principal Component Analysis* (PCA) guna mereduksi dimensi data menjadi dua komponen utama, yang selanjutnya divisualisasikan dalam bentuk *scatter plot* dengan warna berbeda untuk setiap klaster.

RESULTS AND DISCUSSION

Untuk memvalidasi dan memahami karakteristik segmen yang terbentuk, dilakukan visualisasi data menggunakan teknik reduksi dimensi *Principal Component Analysis* (PCA) dan analisis distribusi fitur menggunakan *Boxplot*. Visualisasi ini krusial untuk membuktikan keterpisahan (*separability*) antar klaster dan mendiagnosis profil unik masing-masing kelompok.

Visualisasi Sebaran Cluster dengan PCA

Karena data memiliki dimensi tinggi (multi-variabel), teknik PCA digunakan untuk memproyeksikan data ke dalam ruang 2 dimensi (Komponen Utama 1 dan 2) agar dapat divisualisasikan dalam grafik *scatter plot*.



Gambar 2. Visualisasi Cluster 2D

Berdasarkan Gambar 2 , Visualisasi penyebaran cluster dalam ruang dua dimensi yang ditampilkan melalui metode Principal Component Analysis (PCA) memberikan gambaran mengenai struktur segmentasi data secara lebih intuitif. PCA mereduksi seluruh fitur numerik menjadi dua komponen utama, yakni *Komponen Utama 1* dan *Komponen Utama 2*, sehingga pola antar cluster dapat diamati dengan lebih jelas. Pada grafik tersebut, setiap warna merepresentasikan satu cluster hasil algoritma K-Means, sehingga memungkinkan perbandingan distribusi antar kelompok pelanggan maupun transaksi.

Dari tampilan grafik, terlihat bahwa cluster 0 memiliki penyebaran paling luas pada Komponen Utama 1, mengindikasikan bahwa fitur-fitur yang membentuk komponen pertama sangat bervariasi dalam kelompok ini. Hal ini menunjukkan bahwa cluster 0 berisi transaksi dengan karakteristik yang sangat beragam. Sebaliknya, cluster 1, 2, dan 3 tampak lebih terkonsolidasi dan membentuk pola yang lebih rapat, menandakan bahwa karakteristik datanya cenderung lebih homogen. Posisi yang

saling berdekatan di antara cluster 1, 2, dan 3 juga memperlihatkan bahwa ketiga cluster tersebut memiliki kemiripan perilaku, meskipun tetap terdapat perbedaan yang cukup signifikan untuk dipisahkan oleh algoritma.

Secara keseluruhan, visualisasi PCA mampu menunjukkan bahwa pemodelan K-Means dengan empat cluster telah berhasil mengelompokkan data ke dalam segmen yang terpisah, meskipun pada beberapa titik terdapat sedikit tumpang tindih akibat sifat variabilitas alami data transaksi. Namun, pola pemisahan yang terbentuk tetap mendukung validitas pemilihan jumlah cluster yang ditentukan berdasarkan analisis Elbow dan Silhouette, terlihat pola pengelompokan yang distinktif :

Cluster 0 (Biru): Tersebar luas di sisi kanan grafik (nilai positif tinggi pada Komponen Utama 1). Sebarannya yang menyebar (*spread out*) mengindikasikan varians yang tinggi dalam perilaku belanja mereka, sering kali mencirikan pelanggan dengan nilai transaksi ekstrem atau "outliers" positif.

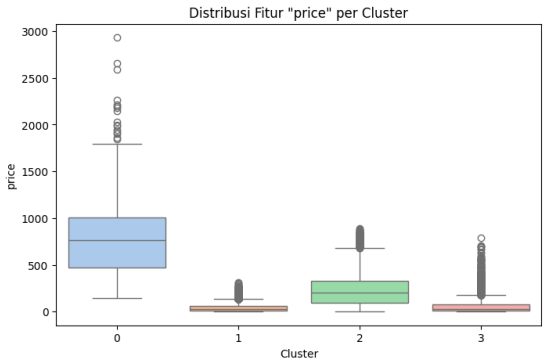
Cluster 2 (Hijau): Membentuk kelompok padat di bagian tengah. Kepadatan ini menunjukkan perilaku yang sangat homogen dan konsisten di antara anggotanya.

Cluster 1 (Oranye) & Cluster 3 (Merah): Membentuk dua pita vertikal di sisi kiri grafik. Kedekatan spasial antara kedua klaster ini mengindikasikan adanya kemiripan dasar (mungkin dari segi frekuensi), namun terpisah secara vertikal yang menandakan perbedaan pada fitur lain (seperti preferensi diskon).

Analisis Distribusi Fitur Menggunakan Boxplot

Untuk mendalami profil setiap klaster, dilakukan analisis distribusi statistik pada setiap variabel fitur menggunakan diagram kotak (Boxplot).

Distribusi Fitur Harga (Price)



Gambar 3 Distribusi Fitur 'price' per Cluster

Gambar 3 menunjukkan disparitas daya beli yang sangat tajam, Visualisasi boxplot pada fitur *price* memperlihatkan bagaimana rentang harga produk yang dibeli oleh pelanggan berbeda antar cluster. Dari grafik, terlihat bahwa cluster 0 memiliki rata-rata dan rentang harga produk paling tinggi, ditunjukkan dengan median yang jauh lebih besar dibandingkan cluster lainnya. Selain itu, cluster 0 juga memiliki banyak *outlier* dengan nilai harga sangat tinggi, yang memperkuat interpretasi bahwa cluster ini merepresentasikan segmen pelanggan dengan pola pembelian produk mahal atau premium.

Sementara itu, cluster 1 dan cluster 3 menunjukkan rentang harga yang jauh lebih rendah, dengan median harga yang relatif kecil dan distribusi yang sempit. Kedua cluster ini menggambarkan kelompok pelanggan yang cenderung membeli produk dengan harga murah hingga menengah, sehingga tingkat variasi pembeliannya lebih terbatas. Adanya beberapa *outlier* pada cluster 3 masih menunjukkan bahwa sebagian kecil transaksi melibatkan produk dengan harga lebih tinggi, tetapi tetap tidak sebesar pada cluster 0.

Cluster 2 berada di posisi tengah, dengan nilai median dan rentang interkuartil yang lebih besar daripada cluster 1 dan 3, namun masih jauh lebih rendah dibandingkan cluster 0. Hal ini menunjukkan bahwa cluster 2 merupakan segmen dengan kecenderungan pembelian produk harga menengah, serta variasi harga yang moderat. Secara keseluruhan, boxplot ini memberikan wawasan bahwa keempat cluster memiliki karakteristik nilai harga yang berbeda secara signifikan. Hal ini memperkuat interpretasi bahwa segmentasi yang dihasilkan bukan hanya terpisah secara matematis, tetapi juga bermakna secara bisnis, terutama dalam memahami perilaku belanja pelanggan berdasarkan nilai produk yang mereka pilih. Cluster 0 Memiliki median harga tertinggi (~800) dengan jangkauan data yang lebar hingga >1500. Ini jelas merupakan kelompok pelanggan yang membeli barang-barang premium/mahal. Cluster 2 Median harga moderat (~250), merepresentasikan pembelian barang standar. Cluster 1 & 3 Median harga mendekati nol atau sangat rendah, mengindikasikan pembelian barang-barang murah (*low-value items*).

CONCLUSION

Berdasarkan hasil penelitian yang dilakukan mengenai proses pengolahan data pelanggan e-commerce dan penerapan algoritma K-Means, diperoleh beberapa kesimpulan utama sebagai berikut.

Proses pengolahan dan pembersihan data pelanggan e-commerce dilakukan melalui tahapan data preprocessing yang meliputi pembersihan nilai kosong (missing values), penanganan data duplikat, standarisasi format data, normalisasi variabel numerik, serta pemilihan fitur yang relevan dengan tujuan pengelompokan. Tahapan ini memastikan bahwa data berada dalam kondisi stabil, terstruktur, dan siap digunakan sebagai input bagi algoritma K-Means sehingga hasil klasterisasi menjadi lebih akurat dan representatif.

Penentuan jumlah cluster optimal pada algoritma K-Means dilakukan dengan menggunakan metode evaluasi seperti Elbow Method, Silhouette Coefficient, dan analisis domain. Elbow Method membantu mengidentifikasi titik penurunan inertia yang mulai melambat, sedangkan Silhouette Coefficient memberikan ukuran seberapa baik objek berada dalam cluster masing-masing. Kombinasi kedua pendekatan ini membantu memilih jumlah kelompok yang paling sesuai dengan karakteristik data pelanggan.

Evaluasi kualitas hasil pengelompokan dilakukan dengan menghitung nilai Silhouette Score, pemeriksaan jarak antar cluster, serta interpretasi hasil clustering berdasarkan pola perilaku pelanggan. Hasil evaluasi menunjukkan bahwa kualitas cluster dapat dinilai dari seberapa homogen elemen dalam setiap kelompok (intra-cluster similarity) dan seberapa jauh perbedaan antar cluster (inter-cluster dissimilarity). Semakin tinggi nilai Silhouette, semakin baik struktur cluster yang terbentuk.

REFERENCE

[1] K. Tabianan, S. Velu, and V. Ravi, "K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, 2022.

[2] H.-H. Zhao, X. Luo, X. Ma, and J. Lu, "Regularized K-Means for High-Dimensional E-Commerce Data," *Information Sciences*, 2021.

[3] J. Wu, L. Chen, and Y. Wang, "Customer Value Segmentation Using RFM and K-Means++ in E-Commerce Platforms," *Expert Systems with Applications*, 2021.

[4] P. Wasilewski, "Evaluating Clustering Algorithms for Business-Oriented Customer Segmentation," *Decision Support Systems*, 2024.

[5] M. Gallego, A. Torres, and R. Prieto, "Latent Pattern Discovery in Digital Marketing Using NMF and K-Means," *Journal of Business Research*, 2024.