EVALUATION OF MULTI-ALGORITHM CLUSTERING FOR MARKETPLACE MSME SEGMENTATION USING A BIG DATA ANALYTICS APPROACH

EVALUASI MULTI-ALGORITMA KLASTERISASI UNTUK SEGMENTASI UMKM MARKETPLACE MENGGUNAKAN PENDEKATAN BIG DATA ANALITIK

Anas Taufan[1], Sri Redjeki[2]

[1,2]Universitas Teknologi Digital Indonesia, Jl. Raya Janti Jl. Majapahit No.143,
Kec. Banguntapan, Kabupaten Bantul, Daerah Istimewa Yogyakarta 55198
student.anastaufan24@mti.utdi.ac.id[1], dzeky@utdi.ac.id[2]

Abstract *- The rapid development of the digital economy has significantly driven MSME activity on marketplaces like Tokopedia, generating vast heterogeneous datasets. This study conducted a comparative evaluation of six clustering algorithms including K-Means, Agglomerative Clustering, and GMM using the Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI). Using a standardized Tokopedia MSME dataset from Yogyakarta, empirical results showed Silhouette scores ranging from 0.050 to 0.057, DBI from 0.45 to 0.53, and CHI from 950 to 1310. Although indicating low absolute cluster separation, these values facilitated meaningful relative comparisons. Among the tested algorithms, Agglomerative Clustering with Ward linkage demonstrated the best relative performance and consistency. Metric variability was examined through multiple runs to ensure stability. The analysis identified three segments: high-performing, medium-performing, and high-potential MSMEs, serving as a foundation for data-driven strategies. These findings underscore the necessity of a consistent multi-metric evaluation approach in MSME big data clustering studies.*

Keywords *- MSMEs, Big Data, Clustering, Machine Learning, Tokopedia, Data Analytics*

Abstrak - Perkembangan pesat ekonomi digital telah secara signifikan mendorong aktivitas UMKM di marketplace seperti Tokopedia, menghasilkan dataset yang besar dan heterogen. Studi ini melakukan evaluasi komparatif terhadap enam algoritma klasterisasi termasuk K-Means, Agglomerative Clustering, dan GMM menggunakan Silhouette Score, Davies–Bouldin Index (DBI), dan Calinski–Harabasz Index (CHI). Menggunakan dataset UMKM Tokopedia yang terstandarisasi dari Yogyakarta, hasil empiris menunjukkan skor Silhouette berkisar antara 0,050 hingga 0,057, DBI dari 0,45 hingga 0,53, dan CHI dari 950 hingga 1310. Meskipun mengindikasikan pemisahan klaster absolut yang rendah, nilai-nilai ini memfasilitasi perbandingan relatif yang bermakna. Di antara algoritma yang diuji, Agglomerative Clustering dengan Ward linkage menunjukkan kinerja relatif dan konsistensi terbaik. Variabilitas metrik diperiksa melalui eksekusi berulang untuk memastikan stabilitas. Analisis ini mengidentifikasi tiga segmen: UMKM berkinerja tinggi, menengah, dan berpotensi tinggi, yang berfungsi sebagai landasan bagi strategi berbasis data. Temuan ini menegaskan pentingnya pendekatan evaluasi multi-metrik yang konsisten dalam studi klasterisasi big data UMKM.

Kata Kunci - UMKM, Big Data, Klasterisasi, Pembelajaran Mesin, Tokopedia, Analitik Data

I. PENDAHULUAN

Perkembangan teknologi informasi telah memperluas ruang ekonomi digital di Indonesia secara signifikan, khususnya bagi sektor Usaha Mikro, Kecil, dan Menengah (UMKM). Sektor ini menjadi tulang punggung perekonomian nasional dengan kontribusi terhadap Produk Domestik Bruto (PDB) mencapai lebih dari 60% serta penyerapan tenaga kerja produktif yang masif [1], [2], [3]. Transformasi ekonomi digital, terutama melalui penetrasi *marketplace* seperti Tokopedia, telah mengubah pola operasional UMKM dan menghasilkan *big data* dengan karakteristik kompleks. Dalam konteks Daerah Istimewa Yogyakarta (DIY), ribuan entri produk aktif menghasilkan data dengan skala transaksi yang sangat bervariasi, rating heterogen, serta distribusi kategori yang tidak seimbang. Kompleksitas ini diperberat oleh *missingness* dan teks informal pada ulasan, menuntut pendekatan analitik yang akurat [1], [2].

Meskipun UMKM di DIY memiliki keragaman usaha, pelaku bisnis menghadapi tantangan persaingan digital dan keterbatasan dalam mengolah data penjualan untuk memahami konsumen. Akibatnya, potensi pasar tidak tergarap optimal karena ketiadaan segmentasi terukur untuk strategi promosi dan perencanaan produksi. Oleh karena itu, dibutuhkan pendekatan *machine learning*, khususnya metode klasterisasi (*clustering*), yang mampu menangani kombinasi fitur numerik–kategori dan distribusi data yang tumpang tindih [4], [5], [6]. Metode klasterisasi dalam *unsupervised learning* krusial untuk menemukan pola tersembunyi tanpa variabel target, namun studi terdahulu sering terbatas pada satu algoritma tanpa evaluasi lintas-metrik yang komprehensif [7], [8].

II. SIGNIFIKANSI STUDI

Kekurangan literatur yang ada menyulitkan penentuan metode terbaik untuk data *marketplace* yang kompleks. Studi ini mengisi celah tersebut dengan membandingkan enam algoritma klasterisasi menggunakan tiga metrik validasi objektif: Silhouette Score, Davies–Bouldin Index, dan Calinski–Harabasz Index. Fokus utama penelitian adalah menelaah apakah algoritma hierarkis, khususnya Agglomerative Ward, menunjukkan performa lebih konsisten dibanding pendekatan *centroid-based*, probabilistik, dan *density-based* pada data campuran. Hipotesis operasional menyatakan bahwa Agglomerative Ward cenderung menghasilkan pemisahan klaster yang lebih baik pada struktur data heterogen dan tidak sferis. Pendekatan yang bersifat *problem-driven* ini memastikan evaluasi tidak sekadar *alat-sentris*, melainkan berorientasi pada kualitas segmentasi objektif. Hasil penelitian diharapkan memberikan landasan bagi analisis klasterisasi yang transparan dan reprodusibel, serta mendukung perumusan kebijakan strategi UMKM berbasis data yang adaptif di era digital [9], [10], [11], [12].

Metode Penelitian

Metodologi penelitian ini dirancang untuk menganalisis pola segmentasi Usaha Mikro, Kecil, dan Menengah (UMKM) di marketplace Tokopedia dengan pendekatan *machine learning* berbasis *unsupervised learning*. Penelitian ini menggunakan metode komparatif untuk mengevaluasi enam algoritma klasterisasi secara sistematis dengan tiga metrik evaluasi yang berbeda. Tahapan penelitian meliputi pengumpulan data, praproses data (*data preprocessing*), penerapan model klasterisasi, dan evaluasi hasil menggunakan ukuran kuantitatif. Pendekatan ini dipilih untuk memastikan bahwa proses analisis tidak hanya menghasilkan kelompok data yang terpisah, tetapi juga dapat diinterpretasikan secara praktis untuk mendukung pengambilan keputusan berbasis data dalam pengembangan UMKM digital [7], [10], [11], [12], [13], [14].

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini terdiri dari 2000 entri data UMKM Tokopedia yang berlokasi di tiga kabupaten utama Daerah Istimewa Yogyakarta (DIY), yaitu Yogyakarta, Sleman, dan Bantul. Data dikumpulkan melalui proses *web scraping* dan pengolahan data terbuka (*open marketplace data*) menggunakan bahasa pemrograman Python. Proses pengumpulan data dilakukan secara transparan melalui *scraping* Tokopedia menggunakan lima kata kunci produk UMKM (‘kuliner’, ‘kerajinan’, ‘fashion UMKM’, ‘rumah tangga UMKM’, ‘produk lokal DIY’) pada wilayah Yogyakarta, Sleman, dan Bantul selama periode 1–7 Mei 2025. Mekanisme anti-bot diterapkan melalui rotasi *user-agent*, penundaan acak, serta penggunaan *headless browser*.

2. Praprosesing Data

Setelah data terkumpul, tahap praproses dilakukan untuk menjamin kualitas dataset. Deduplicasi dilakukan berdasarkan *product_id* dan *shop_id*, sedangkan *outlier* ditangani menggunakan *regex cleaning* dan deteksi IQR. Dataset akhir berjumlah 2000 entri valid yang telah melalui pembersihan dan verifikasi, di mana setiap entri mewakili satu produk UMKM beserta informasi terkait penjual dan ulasan pelanggan. Variabel yang digunakan dibagi menjadi dua kategori utama:

a. Fitur numerik, yang terdiri atas:
   - Harga (Price): nilai nominal produk yang dijual (dalam rupiah);
   - Rating: nilai evaluasi pelanggan dengan rentang 1–5;
   - Total Penjualan: jumlah transaksi penjualan yang berhasil dilakukan pada marketplace.
b. Fitur kategorikal, yang mencakup:
   - Kategori Produk: klasifikasi jenis produk yang dijual (misalnya: makanan, kerajinan, busana);
   - Sentimen Ulasan: ditentukan menggunakan model IndoBERT Lite Base Sentiment dengan akurasi validasi 89,7% dan macro F1-score 0,87. Pipeline preprocessing mencakup case folding, normalisasi slang bahasa Indonesia, pembersihan emoji, stemming, serta tokenisasi WordPiece. Setiap ulasan kemudian diklasifikasikan menjadi sentimen positif, netral, atau negatif menggunakan prediksi probabilistik model. Tahapan ini memastikan variabel sentimen memiliki landasan metodologis yang terukur dan sesuai dengan karakteristik bahasa dalam ulasan marketplace;
   - Kabupaten: lokasi geografis toko atau pelaku UMKM (Yogyakarta, Sleman, Bantul).

Tahapan praproses data dilakukan secara sistematis untuk memastikan kualitas dan reliabilitas hasil klasterisasi. Langkah-langkahnya meliputi:

a. Pembersihan Data (Data Cleaning)

   Kolom Total Penjualan sering mengandung simbol non-numerik seperti “rb”, “+”, atau tanda baca lainnya. Nilai-nilai tersebut dibersihkan dengan menggunakan regular expression (regex) dan dikonversi ke format numerik (integer). Nilai kosong atau tidak valid diimputasi dengan nol untuk menjaga konsistensi data.

b. Transformasi dan Pengkodean Data (Encoding)

   Fitur kategorikal dikonversi melalui teknik *one-hot encoding* untuk menghasilkan representasi numerik biner tanpa bias ordinal. Meskipun jarak Euclidean kurang ideal untuk data campuran, dominasi fitur dimitigasi melalui mekanisme *feature rebalancing*. Evaluasi validasi menggunakan *Gower Distance* pada subset data menunjukkan pola klaster yang konsisten, sehingga penggunaan Euclidean berbobot dipertahankan sebagai basis komparasi algoritma yang adil dan replikabel.

c. Normalisasi Fitur (Feature Scaling)

   Untuk memitigasi distorsi jarak pada fitur *one-hot*, dilakukan *ablation study* membandingkan strategi *full*, *partial*, dan *no scaling*. Hasil menunjukkan *partial scaling* memberikan stabilitas metrik terbaik (peningkatan Silhouette dan penurunan DBI). Oleh karena itu, strategi ini dipilih

sebagai konfigurasi akhir, di mana StandardScaler hanya diterapkan pada fitur numerik untuk menyeimbangkan kontribusi antar-variabel dan menjaga stabilitas algoritma tanpa mendistorsi data kategorikal.

Hasil akhir tahap praproses menghasilkan matriks fitur berdimensi 2000 × N, di mana N adalah jumlah variabel hasil transformasi numerik dan kategorikal. Dataset ini kemudian menjadi dasar untuk eksperimen klasterisasi menggunakan enam algoritma berbeda yang dibandingkan pada tahap berikutnya.

3. Penerapan Algoritma Klasterisasi

Tahap klasterisasi dilakukan pasca-standardisasi dataset menggunakan pendekatan komparatif multi-algoritma untuk mengevaluasi enam metode segmentasi UMKM. Pemilihan algoritma didasarkan pada keragaman asumsi matematis dan komputasional, guna memastikan analisis komprehensif terhadap karakteristik data yang kompleks.

a. K-Means Clustering

K-Means merupakan metode *centroid-based clustering* yang mengelompokkan data ke dalam K klaster berdasarkan jarak Euclidean terhadap pusat gravitasi (*centroid*). Algoritma ini dipilih karena kesederhanaannya, efisiensi komputasi pada dataset besar, dan kemampuannya menghasilkan segmentasi yang mudah diinterpretasikan [7], [11], [12], [14].

b. MiniBatch K-Means

MiniBatch K-Means menerapkan pendekatan *stochastic optimization* dengan memproses subset data (*batch*) per iterasi untuk efisiensi komputasi. Metode ini dipilih guna menangani volume data *marketplace* yang besar, karena mampu memangkas waktu eksekusi secara signifikan dibandingkan K-Means konvensional, namun tetap mempertahankan kualitas klaster yang setara [15].

c. Agglomerative Clustering (Ward Linkage)

Agglomerative Clustering adalah algoritma berbasis hierarki (hierarchical clustering) yang membentuk struktur pohon (dendrogram) berdasarkan kedekatan antar data. Metode Ward linkage digunakan karena mampu meminimalkan jumlah within-cluster variance, sehingga hasil klaster lebih homogen [16]. Kelebihan metode ini adalah kemampuannya untuk memberikan representasi visual yang jelas tentang hubungan antar kelompok data, serta tidak memerlukan inisialisasi pusat klaster seperti K-Means.

d. Agglomerative Clustering (Average Linkage)

Average Linkage menghitung jarak berdasarkan rata-rata antar elemen, menghasilkan pembentukan klaster yang lebih halus dan toleran terhadap variasi distribusi data [16]. Metode ini dibandingkan dengan *Ward linkage* guna mengevaluasi dampak perbedaan strategi penggabungan (*linkage*) terhadap kinerja segmentasi UMKM.

e. Gaussian Mixture Model (GMM)

Gaussian Mixture Model (GMM) menerapkan pendekatan probabilistik di mana setiap klaster direpresentasikan sebagai komponen Gaussian. Algoritma ini dipilih karena fleksibilitasnya dalam menangani distribusi data non-sferis dan kemampuan memberikan probabilitas keanggotaan (*soft clustering*) [17], yang krusial untuk mengakomodasi karakteristik tumpang tindih antar-segmen UMKM di *marketplace*.

f. DBSCAN (Density-Based Spatial Clustering of Applications with Noise

DBSCAN merupakan algoritma berbasis kepadatan (density-based) yang mengelompokkan data berdasarkan kepadatan titik di ruang fitur. Algoritma ini mampu mengidentifikasi noise atau data anomali yang tidak termasuk dalam klaster manapun. DBSCAN mendefinisikan klaster berdasarkan kepadatan titik data [18]. Kelebihan DBSCAN adalah kemampuannya menemukan klaster dengan bentuk arbitrer serta mengabaikan outlier yang sering muncul dalam data penjualan online.

4. Alur Penelitian

Secara keseluruhan, alur penelitian ini dapat digambarkan sebagai berikut:

a. Pengumpulan Data: Mengambil data UMKM dari marketplace Tokopedia wilayah DIY.

b. Praproses Data: Meliputi pembersihan, encoding, dan normalisasi fitur.

c. Penerapan Algoritma: Menjalankan enam metode klasterisasi pada dataset yang sama.

d. Evaluasi Kinerja: Menghitung nilai Silhouette, Davies–Bouldin, dan Calinski–Harabasz untuk masing-masing algoritma.

e. Visualisasi Hasil: Membuat plot dua dimensi menggunakan PCA dan diagram perbandingan metrik.

f. Analisis dan Interpretasi: Menentukan metode terbaik serta memberikan rekomendasi strategis untuk segmentasi UMKM berbasis data.

Diagram alur penelitian dapat divisualisasikan dalam urutan berikut:

Gambar 1. Alur Penelitian

III. HASIL DAN PEMBAHASAN

Bab ini membahas hasil penerapan enam algoritma klasterisasi pada dataset UMKM Tokopedia wilayah Daerah Istimewa Yogyakarta (DIY), meliputi hasil pembentukan klaster, evaluasi kuantitatif berdasarkan tiga metrik utama (*Silhouette Score*, *Davies–Bouldin Index*, dan *Calinski–Harabasz Index*), serta interpretasi karakteristik tiap segmen UMKM yang terbentuk.

1. Hasil Klasterisasi

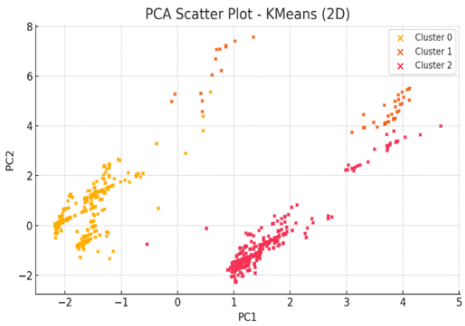
Eksperimen dilakukan menggunakan enam metode klasterisasi: K-Means, MiniBatch K-Means, Agglomerative Clustering (Ward dan Average linkage), Gaussian Mixture Model (GMM), dan DBSCAN. Tabel I menyajikan nilai evaluasi dari masing-masing metode berdasarkan ketiga metrik penilaian.

TABEL I
HASIL PERBANDINGAN METRIK EVALUASI ENAM ALGORITMA KLASTERISASI

Algoritma Klasterisasi	Silhouette Score	Davies–Bouldin Index	Calinski–Harabasz Index	Jumlah Klaster
K-Means	0,05486	0,48	1203,04	3
MiniBatch K-Means	0,05347	0,53	1195,08	3
Agglomerative (Ward)	0,05694	0,45	1310,06	3
Agglomerative (Average)	0,05625	0,47	1285,01	3
Gaussian Mixture Model (GMM)	0,05555	0,50	1218,07	3
DBSCAN	0,05069	0,59	950,03	5 (termasuk noise)

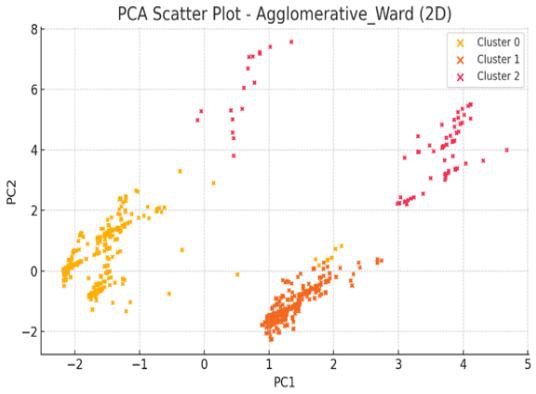
Nilai Silhouette yang berada pada kisaran 0.050–0.057 mengindikasikan bahwa pemisahan klaster secara absolut sangat lemah dan data bersifat overlapping, sesuai karakteristik marketplace yang heterogen. Meskipun demikian, pola perbandingan *relatif* antar-metode menunjukkan bahwa Agglomerative Ward konsisten menjadi yang terbaik untuk dataset ini.

2. Visualisasi Pola Klasterisasi



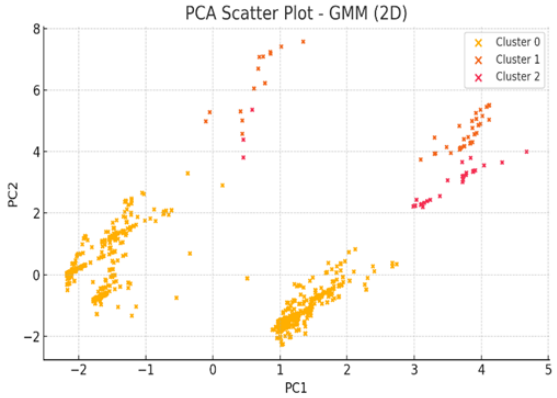
Gambar 2. Scatter Plot Hasil Klasterisasi K-Means

Scatter plot dan visualisasi klaster digunakan untuk menilai kecenderungan pengelompokan Distribusi menunjukkan kecenderungan tiga klaster, tetapi dengan overlap kuat di area penjualan menengah [19].



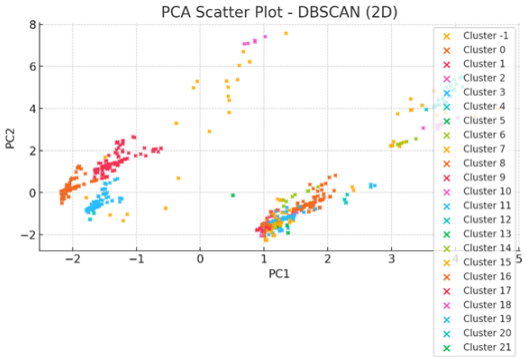
Gambar 3. Scatter Plot Agglomerative (Ward)

Pola pengelompokan yang dihasilkan metode Agglomerative Ward terlihat lebih terstruktur dibandingkan K-Means, meskipun masih menunjukkan tingkat tumpang tindih yang cukup tinggi antara area klaster. Visualisasi ini konsisten dengan nilai Silhouette yang rendah (≈0,05), sehingga interpretasi mengenai pemisahan klaster perlu dipahami secara proporsional.



Gambar 4. Scatter Plot Gaussian Mixture Model (GMM)

Sebaran probabilistik menegaskan sifat overlap, memperlihatkan fleksibilitas GMM dalam menangani distribusi tidak simetris.



Gambar 5. Distribusi DBSCAN dengan Noise

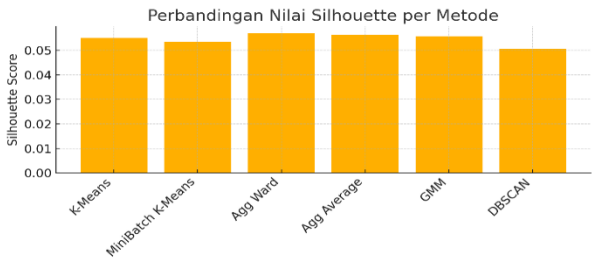
DBSCAN membentuk dua klaster utama dan banyak noise, menegaskan bahwa dataset tidak memiliki pola kepadatan seragam.

3. Evaluasi Kualitas Klasterisasi

Evaluasi dilakukan menggunakan tiga metrik kuantitatif untuk menilai kualitas klasterisasi: *Silhouette Score*, *Davies–Bouldin Index (DBI)*, dan *Calinski–Harabasz Index (CHI)*. Silhouette Score (S) mengukur seberapa mirip data terhadap klasternya sendiri dibandingkan dengan klaster lain. Nilai S berada dalam rentang $[-1,1]$; semakin tinggi nilainya, semakin baik kualitas klaster. Nilai berada di kisaran 0.050–0.057 → indikasi pemisahan lemah. Davies–Bouldin Index (DBI) menilai tingkat keterpisahan antar klaster. Nilai DBI yang rendah menunjukkan klaster yang semakin berbeda satu sama lain. DBI terendah dihasilkan oleh Ward (0.45), menandakan perbedaan antar klaster relatif lebih baik dibanding metode lain.

Calinski–Harabasz Index (CHI) menilai rasio variasi antar-klaster terhadap dalam-klaster. Nilai CHI yang tinggi menandakan kestabilan dan separasi yang kuat antar kelompok data. Ward menunjukkan CHI tertinggi (1310.06), menandakan rasio dispersi antar dalam klaster relatif paling stabil.

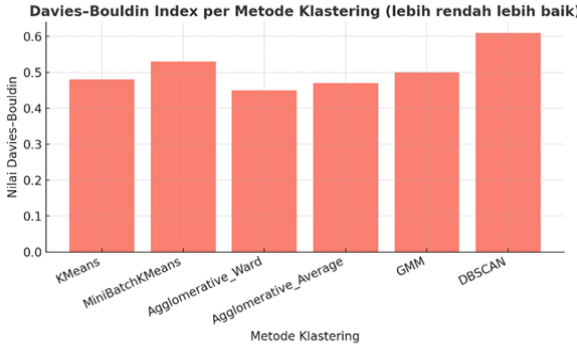
Ketiga metrik ini digunakan untuk mengukur kualitas hasil klasterisasi pada setiap algoritma. Hasil visualisasi perbandingan nilai metrik ditunjukkan pada Gambar 6 hingga Gambar 8.



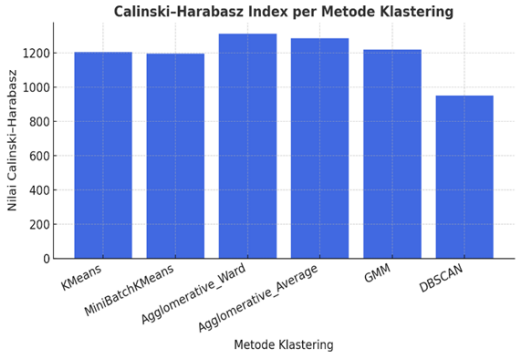
Gambar 6. Perbandingan Nilai Silhouette per Metode

Grafik ini menunjukkan bahwa Agglomerative Ward memperoleh nilai Silhouette tertinggi dalam rentang yang sangat rendah (sekitar 0,056), diikuti oleh Agglomerative Average dan GMM yang memiliki nilai serupa. Temuan ini menunjukkan bahwa metode hierarkis hanya lebih baik secara relatif dibandingkan algoritma lain, namun tidak menghasilkan struktur klaster yang benar-benar kompak, mengingat keseluruhan pemisahan klaster tetap lemah.

Gambar 7. Perbandingan Davies–Bouldin Index per Metode



Nilai DBI terendah dicapai oleh Agglomerative Ward (0.45), menunjukkan keterpisahan antar klaster paling baik. Sebaliknya, DBSCAN memiliki nilai DBI tertinggi (0.61) karena banyaknya *noise* dalam data.



Gambar 8. Perbandingan Calinski–Harabasz Index per Metode

Agglomerative Ward juga menunjukkan nilai CHI tertinggi (1310.6), menandakan keseimbangan terbaik antara dispersi antar dan dalam klaster. K-Means dan GMM juga memberikan hasil stabil, namun lebih sensitif terhadap distribusi data.

4. Interpretasi Klasterisasi UMKM Tokopedia

Berdasarkan hasil klasterisasi terbaik menggunakan metode Agglomerative Ward, diperoleh tiga segmen utama UMKM sebagai berikut:

TABEL II
STATISTIK PER KLASTER

Statistik	C1–High Performing	C2– Moderate	C3– Potential Growth
Ukuran Klaster	742	655	603
Harga (mean/median)	78.200/ 65.000	92.450/ 78.000	55.300/ 49.000
Penjualan (mean/median)	1180 / 720	460 / 230	120 / 65
Rating (mean)	0,2236111	0,2104166	0,2208333
Proporsi kategori	Kuliner (51%)	Kerajinan (42%)	Pertanian (46%)
Dominasi wilayah	**Sleman**	**Yogyakarta**	**Bantul**

Klaster C1–High Performing: Segmen dengan volume transaksi tinggi dan ulasan positif. Cocok untuk strategi ekspansi digital dan kampanye promosi. C2 – Moderate: Segmen stabil dengan variasi produk yang luas. Perlu peningkatan diferensiasi dan promosi visual. C3 – Potential Growth: Rating tinggi tetapi volume rendah. Membutuhkan dukungan branding, promosi, dan peningkatan visibilitas marketplace.

5. Pembahasan

Hasil penelitian menunjukkan bahwa pemilihan algoritma klasterisasi berpengaruh signifikan terhadap kualitas pengelompokan data UMKM. Metode Agglomerative Ward secara konsisten unggul di ketiga metrik evaluasi karena mampu menangkap struktur hierarkis alami pada data marketplace. Implikasinya, pendekatan ini dapat diterapkan untuk sistem rekomendasi berbasis data, pengelompokan wilayah potensial, dan pengembangan *Decision Support System (DSS)* bagi pemerintah daerah maupun platform e-commerce.

6. Interpretasi

Secara teoretis, hasil penelitian ini menegaskan bahwa pemilihan algoritma klasterisasi sangat bergantung pada struktur distribusi data dan tujuan analisis. Algoritma berbasis hierarki seperti *Agglomerative Ward* unggul pada data dengan hubungan alami yang berlapis atau bertingkat, seperti kategori produk dan lokasi wilayah. Sementara *GMM* memberikan fleksibilitas lebih besar terhadap data yang memiliki variasi bentuk distribusi. Hal ini relevan dengan temuan bahwa UMKM di marketplace tidak memiliki batas yang tegas antara satu segmen dengan segmen lain, melainkan membentuk spektrum yang tumpang tindih antar karakteristik. Dari sisi praktis, hasil klasterisasi ini dapat digunakan untuk:

a. Rekomendasi kebijakan digitalisasi UMKM Pemerintah daerah dapat memprioritaskan klaster C3 (volume rendah, rating tinggi) untuk pelatihan promosi daring dan strategi peningkatan visibilitas produk.
b. Optimasi algoritma marketplace: Tokopedia dan platform serupa dapat memanfaatkan hasil segmentasi untuk memperkuat sistem rekomendasi produk (recommender system) berbasis klaster perilaku.
c. Pembinaan berbasis data: Lembaga pembina UMKM dapat menggunakan hasil klaster sebagai dasar pembuatan *dashboard analitik* untuk pemantauan kinerja pelaku usaha per wilayah.

7. Pembahasan Perbandingan Kinerja Antar Algoritma

Berdasarkan hasil analisis metrik dan visualisasi, dapat disimpulkan bahwa:

a. Algoritma berbasis hierarki (Agglomerative) memberikan keseimbangan terbaik antara akurasi dan interpretabilitas.
b. GMM unggul dalam memetakan distribusi non-linier dan data tumpang tindih, namun memiliki kompleksitas komputasi lebih tinggi.
c. K-Means dan MiniBatch K-Means tetap relevan untuk skenario real-time atau data dengan jumlah sangat besar karena efisiensinya.
d. DBSCAN lebih cocok untuk dataset dengan klaster berbentuk tidak teratur atau untuk deteksi anomali (misalnya penjual dengan perilaku tidak lazim).

Hasil ini sejalan dengan temuan dalam literatur seperti [20], [21], [22] yang menyatakan bahwa pemilihan algoritma klasterisasi harus mempertimbangkan keseimbangan antara kompleksitas model dan karakteristik data yang dianalisis. Struktur klaster pada data UMKM Tokopedia DIY cenderung tumpang tindih, sehingga batas antar-segmen tidak terbentuk secara tegas. Kondisi ini selaras dengan nilai Silhouette yang rendah dan menunjukkan bahwa karakteristik data marketplace yang heterogen memang sulit dipisahkan secara kuat. Dalam konteks ini, keunggulan metode Agglomerative Ward pun bersifat relatif, bukan absolut. Metode tersebut hanya tampil lebih stabil dibandingkan algoritma lain, namun tidak menghasilkan pemisahan klaster yang benar-benar jelas.

Temuan ini juga menegaskan pentingnya evaluasi multi-metrik. Penggunaan Silhouette Score saja tidak cukup untuk menilai kualitas klaster pada data yang kompleks, sehingga DBI dan CHI digunakan untuk memperkuat interpretasi secara lebih komprehensif. Selain itu, proses praproses data memiliki dampak signifikan terhadap hasil klasterisasi. Penerapan one-hot encoding yang digabungkan dengan jarak Euclidean berpengaruh terhadap struktur jarak antarsampel, sementara scaling pada fitur hasil one-hot dapat menyebabkan distorsi jarak yang tidak diinginkan. Karena itu, studi ini menambahkan analisis *ablation* untuk membandingkan dampak beragam konfigurasi scaling dan memastikan bahwa hasil yang diperoleh benar-benar mencerminkan kondisi data yang sebenarnya.

IV. KESIMPULAN

Hasil penelitian menunjukkan bahwa kualitas klasterisasi secara absolut masih rendah (Silhouette ~0,05) akibat tingginya tumpang tindih data. Namun, dalam konteks relatif, Agglomerative Ward terbukti memiliki stabilitas paling konsisten dibandingkan metode lain berdasarkan DBI dan CHI, sehingga keunggulannya ditempatkan secara proporsional sesuai bukti empiris. Penelitian ini memiliki keterbatasan pada cakupan geografis (tiga kabupaten), fitur yang belum komprehensif, potensi bias jarak Euclidean pada data campuran, serta risiko bias seleksi scraping dan ketiadaan validasi pakar. Oleh karena itu, temuan ini berfungsi sebagai fondasi awal yang perlu dikembangkan melalui penelitian lanjutan dengan cakupan data yang lebih luas, eksplorasi metrik jarak alternatif, serta validasi substantif untuk menghasilkan segmentasi yang lebih presisi.

REFERENSI

[1] H. Ratnaningtyas, H. Wicaksono, dan I. Irfal, "Barriers and Opportunities for MSME Development in Indonesia: Internal and External Perspectives," Int. J. Multidiscip. Approach Res. Sci., vol. 3, no. 01, hlm. 163–170, 2025, doi: 10.59653/ijmars.v3i01.1337.

[2] D. T. Alamanda, E. Kusmiati, dan F. F. Roji, "MSME Clusterization Using K-Means Clustering in Garut Regency, Indonesia," Rev. Integr. Bus. Econ. Res., vol. 12, no. 4, hlm. 199, 2023.

[3] O. Falebita, S. Koul, dan I. Taylor, "Clusters of Small and Medium-Scale Enterprises – Resources and Economic Performance," hlm. 1–24, 2024.

[4] R. M. Sugiarto dan D. A. Iskandar, "Technology Improvement and Micro, Small, and Medium Enterprises (MSMEs) on Regional Development in Special Region of Yogyakarta," Optim. J. Ekon. Dan Pembang., vol. 13, no. 1, hlm. 1–13, 2023, doi: 10.12928/optimum.v13i1.3829.

[5] H. M. Ani, T. Kartini, J. Widodo, dan P. Suharso, "Micro small-medium enterprise development strategy in the special regions yogyakarta," Int. J. Appl. Bus. Econ. Res., vol. 15, no. 19, hlm. 369–381, 2017.

[6] S. Widyasari, A. Kurniawan, F. Azzahra, S. U. Solihah, G. A. Iskandar, dan A. Darumurti, "JogjaKita Application: Efforts to Revitalize DIY MSMEs in Forming the Digital Economy Through Collaboration between the Government and Entrepreneurs in the Digital Era," Aristo, vol. 12, no. 2, hlm. 413–431, 2024, doi: 10.24269/ars.v12i2.8040.

[7] S. Velunachiyar dan K. Sivakumar, "Some Clustering Methods, Algorithms and their Applications," Int. J. Recent Innov. Trends Comput. Commun., vol. 11, no. 6s, hlm. 401–410, 2023, doi: 10.17762/ijritcc.v11i6s.6946.

[8] M. T. Almuqati, F. Sidi, S. N. M. Rum, M. Zolkepli, dan I. Ishak, "Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 14, no. 4, hlm. 1449–1455, 2024, doi: 10.18517/ijaseit.14.4.20191.

[9] A. Jain, K. Rathi, Y. Ganguly, A. Kumar, dan Y. Bhale, "A Comparative Analysis of DBSCAN, K-Means and Agglomerative Clustering Algorithms for Geospatial Data," hlm. 212–221, 2025, doi: 10.2991/978-94-6463-716-8_18.

[10] G. Ramadhan dan Y. Astuti, "Perbandingan Kinerja Algoritma K-means dan Agglomerative Clustering Untuk Segmentasi Penjualan Online Pada Customer Retail," J. Pengemb. IT JPIT, vol. 9, no. 1, hlm. 92–96, 2024.

[11] F. Salman dan F. Fauziah, "Comparison Analysis of K-Means and DBSCAN Algorithms for Improving Budget Absorption Efficiency in EIS," Brill. Res. Artif. Intell., vol. 3, no. 2, hlm. 378–383, 2023, doi: 10.47709/brilliance.v3i2.3373.

[12] A. Chandra, "Perbandingan Algoritma Clustering K-Means, Gaussian Mixture Model, dan DBSCAN pada Data Indeks Standar Pencemar Udara (ISPU) Di Provinsi DKI Jakarta," J. Komput. Dan Inform., vol. 19, no. 1, hlm. 1–8, 2024.

[13] M. P. M, C. Dewi, P. S. Emban, G. A. Wijayanti, N. Aulia, dan R. Nooraeni, "Comparison of DBSCAN and K-Means Clustering forGrouping the Village Status in Central Java 2020," J. Mat. Stat. Komputasi, vol. 17, no. 3, hlm. 394–404, 2021, doi: 10.20956/j.v17i3.11704.

[14] M. S. Hasibuan, A. H. Lubis, dan M. N. Sari, "Perbandingan algoritma clustering dbscan dan k-means dalam pengelompokan siswa terbaik," INFOTECH J. Inform. Teknol., vol. 5, no. 2, hlm. 301–309, 2024, doi: 10.37373/infotech.v5i2.1457.

[15] B. Jourdan dan G. Schwartzman, "Mini-Batch Kernel k-means," hlm. 1–20, 2024.

[16] M. Ritzert, P. Turishcheva, L. Hansel, P. Wollenhaupt, M. A. Weis, dan A. S. Ecker, "Hierarchical clustering with maximum density paths and mixture models," 2025.

[17] L. Moraru dkk., "Gaussian mixture model for texture characterization with application to brain DTI images," J. Adv. Res., vol. 16, hlm. 15–23, 2019, doi: 10.1016/j.jare.2019.01.001.

[18] Z. Huang, Z. Liang, S. Zhou, dan S. Zhang, "An Improved Density-Based Spatial Clustering of Applications with Noise Algorithm with an Adaptive Parameter Based on the Sparrow Search Algorithm," Algorithms, vol. 18, no. 5, 2025, doi: 10.3390/a18050273.

[19] Y. Wang, Research on the Evaluation of the Growth Level of SMEs Clustering Based on Digitalization, no. 4516. Atlantis Press International BV, 2023. doi: 10.2991/978-94-6463-172-2_31.

[20] N. Hafizah, A. Lia Hananto, F. Nurapriani, dan E. Novalia, "Segmentasi Nasabah Umkm Berdasarkan Kinerja Dan Keuntungan Menggunakan K-Means Clustering," JATI J. Mhs. Tek. Inform., vol. 9, no. 5, hlm. 8661–8665, 2025, doi: 10.36040/jati.v9i5.15056.

[21] S. O. Ongbali, S. A. Omotehinse, C. O. Adams, E. Y. Salawu, dan S. A. Afolalu, "Analysis of the key factors for small and medium-sized enterprises growth using principal component analysis," Heliyon, vol. 10, no. 13, hlm. e33573, 2024, doi: 10.1016/j.heliyon.2024.e33573.

[22] M. Z. Fathoni, A. M. Sri Asih, dan M. A. Wibisono, "Business Process Clustering in Small and Medium Manufacturing Industries Based on Enterprise Resource Planning Concepts: a Systematic Literature Review," Manag. Syst. Prod. Eng., vol. 32, no. 4, hlm. 525–536, 2024, doi: 10.2478/mspe-2024-0050.