KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research (JSSR)
Vol.5, No.1 Februari 2027
e-ISSN: 3025-9851, p-ISSN : 3025-986X, Hal 344-359
DOI : https://doi.org/10.61722/jssr.v5i1.13633

Pengelompokan Produk Vape Berdasarkan Pola Penjualan Menggunakan Algoritma HDBSCAN

Meizia Raihan[1*], Wina Witanti[2], Melina[3]

[1] Program Studi Teknik Informatika, Fakultas Sains Dan Informatika, Universitas Jenderal Achmad Yani
[2] Program Studi Teknik Informatika, Fakultas Sains Dan Informatika, Universitas Jenderal Achmad Yani
[3] Program Studi Teknik Informatika, Fakultas Sains Dan Informatika, Universitas Jenderal Achmad Yani
*Penulis Korespondensi: meiziaraihan21@if.unjani.ac.id

Abstract. *Transaction data are a key source for understanding customer behavior and market trends that can support effective business decision-making. However, a vapestore currently has a large volume of sales data that has not been optimally utilized, making it difficult to identify potential purchasing patterns manually. This study proposes the use of the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to cluster vape product sales patterns based on transaction characteristics. The dataset consists of vape sales transactions collected from January 2024 to January 2026, including product category, product name, variant, quantity, and transaction value attributes. The research process involved data preprocessing, feature transformation, standardization, parameter optimization, and cluster evaluation using the Silhouette Coefficient. Experimental results showed that the optimal parameters were obtained at min_samples = 3 and min_cluster_size = 5, producing 232 clusters with a Silhouette Score of 0.7436. The results indicate that HDBSCAN effectively identifies dense sales patterns and detects outlier products within high-dimensional retail transaction data. The generated clusters provide meaningful insights for inventory management, product segmentation, and sales strategy optimization in vapestore businesses.*

Keywords: *sales analysis; clustering; data mining; transaction data; HDBSCAN.*

Abstrak. Data transaksi merupakan sebuah kunci untuk memahami perilaku pelanggan dan tren pasar yang dapat membantu dalam pengambilan keputusan bisnis yang baik. Namun, sebuah vapestore saat ini memiliki volume data penjualan yang besar dan belum dimanfaatkan secara optimal, sehingga pola pembelian potensial sulit untuk diidentifikasi secara manual. Penelitian ini mengusulkan penggunaan algoritma Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) untuk mengelompokkan pola penjualan produk vape berdasarkan karakteristik transaksi. Dataset terdiri dari transaksi penjualan vape yang dikumpulkan dari Januari 2024 hingga Januari 2026, mencakup atribut kategori produk, nama produk, varian, sub varian, bulan_num, total_qty, dan jumlah_transaksi. Proses penelitian meliputi pra-pemrosesan data, transformasi fitur, standardisasi, optimasi parameter, dan evaluasi klaster menggunakan Silhouette Coefficient. Hasil eksperimen menunjukkan bahwa parameter optimal diperoleh pada min_samples = 3 dan min_cluster_size = 5, menghasilkan 232 klaster dengan Silhouette Score sebesar 0,7437. Hasil tersebut menunjukkan bahwa HDBSCAN efektif mengidentifikasi pola penjualan yang padat dan mendeteksi produk outlier dalam data transaksi ritel berdimensi tinggi. Klaster yang dihasilkan memberikan wawasan bermakna untuk manajemen inventaris, segmentasi produk, dan optimasi strategi penjualan pada bisnis vapestore.

Kata kunci: analisis penjualan; *clustering; data mining;* data transaksi; HDBSCAN.

1. LATAR BELAKANG

Pertumbuhan volume data yang dinamis secara global, yang dikenal sebagai big

Naskah Masuk:24 November 2026; Revisi: 25 November 2026; Diterima: 26 November 2026; Terbit: 27 November 2026

data, telah menciptakan kebutuhan mendesak akan metode analitis baru yang melampaui pendekatan statistik tradisional untuk menemukan pengetahuan baru dari data. Data Mining merupakan bidang interdisipliner yang muncul untuk secara otomatis atau semi-otomatis mempertimbangkan sejumlah besar prediktor gabungan untuk mengungkap pola tersembunyi yang valid dan signifikan dalam data (Shu & Ye, 2023). Kondisi persaingan global ini juga sangat relevan di Indonesia. Meskipun entitas ritel mengumpulkan data transaksi penjualan dalam jumlah besar, data tersebut pada umumnya masih jarang digunakan dan hanya menumpuk dalam basis data. Padahal, terdapat masalah yang dialami oleh pelaku usaha terkait produk terlaris dan upaya untuk meningkatkan penjualan berikutnya. Dampaknya, kesulitan dalam mengidentifikasi pola pembelian yang signifikan dapat memicu penurunan tingkat penjualan di tengah persaingan. Untuk mengatasi masalah tersebut, pemanfaatan data penjualan melalui teknik Data Mining dipandang perlu sebagai bentuk efisiensi guna membantu dalam meningkatkan strategi penjualan yang efektif (Nisa, 2021).

Pada sebuah kasus, untuk menemukan suatu pola yang dapat meningkatkan penjualan dan pemasaran barang menggunakan solusi penerapan teknik klasterisasi (Clustering). Teknik ini vital untuk mengidentifikasi pola-pola tersembunyi dalam data (Hendrastuty, 2024b), lalu proses klasterisasi dapat digunakan untuk mengidentifikasi kelompok pola serupa (Maulana et al., 2025), yang mana sangat berguna untuk memperkuat strategi penjualan dan segmentasi pasar. Dalam bidang klasterisasi berbasis kepadatan, salah satu algoritma yang berkembang adalah *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN). Algoritma ini merupakan pengembangan dari DBSCAN yang dirancang untuk mengatasi keterbatasan dalam penggunaan parameter epsilon tetap serta kesulitan dalam menangani variasi densitas data. HDBSCAN membangun struktur hierarki berdasarkan kepadatan lokal setiap titik data dan mengekstraksi klaster yang memiliki tingkat stabilitas tertinggi (Wahyuni, Hayati, & Rizki, 2021). Keunggulan metode ini adalah kemampuannya dalam mendeteksi noise secara otomatis dan tidak memerlukan penentuan jumlah klaster di awal proses (Sante, Font, Mistry, Ortega-Martorell, & Olier, 2025). Beberapa penelitian juga menunjukkan bahwa HDBSCAN efektif diterapkan pada data berdimensi tinggi dan data dengan distribusi tidak merata (Kaverinskiy et al., 2025).

Penelitian terdahulu yang menerapkan klasterisasi oleh Chandra Purnama, dkk

membahas penerapan algoritma K-Means untuk mengelompokkan data penjualan pakaian pada sebuah distro. Hasil penelitian menunjukkan bahwa data penjualan berhasil dikelompokkan ke dalam dua klaster utama berdasarkan kemiripan pola penjualan produk. Pengelompokan ini mampu mengidentifikasi produk dengan tingkat penjualan tinggi dan rendah, sehingga dapat digunakan sebagai dasar dalam penentuan strategi penjualan, pengelolaan stok, dan rekomendasi produk. Kualitas klaster yang dihasilkan dinilai cukup baik, ditunjukkan oleh nilai Davies Bouldin Index (DBI) sebesar 0,438, yang menandakan bahwa hasil klasterisasi memiliki tingkat pemisahan dan kekompakan yang baik (Purnama et al., 2022).

Penelitian oleh Nirwana Hendrastuty membahas penerapan algoritma K-Means untuk mengelompokkan data nilai hasil belajar siswa. Hasil penelitian menunjukkan bahwa proses clustering mampu membagi siswa ke dalam dua kelompok, yaitu kelompok siswa Rajin (C0) sebanyak 63 siswa dan kelompok siswa Sangat Rajin (C1) sebanyak 91 siswa. Evaluasi kualitas klaster menggunakan silhouette score menghasilkan nilai tertinggi sebesar 0,9168, yang menandakan bahwa hasil pengelompokan memiliki tingkat pemisahan yang sangat baik (Hendrastuty, 2024a).

Selanjutnya, penelitian oleh Nanda Anggun Wahyuni, dkk membahas penerapan algoritma HDBSCAN untuk mengelompokkan desa/kelurahan tertinggal berdasarkan data Potensi Desa (PODES) tahun 2018. Hasil penelitian menunjukkan bahwa metode HDBSCAN dengan parameter optimal Mpts = 4 mampu membentuk 2 cluster utama dari 84 desa/kelurahan tertinggal, yaitu cluster 1 yang terdiri dari 4 desa/kelurahan dan cluster 2 yang terdiri dari 61 desa/kelurahan, sementara sisanya teridentifikasi sebagai outlier. Hasil clustering memperlihatkan perbedaan karakteristik yang cukup jelas, khususnya pada variabel jumlah keluarga pengguna listrik PLN, di mana cluster 1 memiliki rata-rata pengguna listrik yang jauh lebih rendah dibandingkan cluster 2 (Wahyuni et al., 2021). Meskipun klasterisasi merupakan solusi yang menjanjikan, terdapat tantangan teknis dalam implementasinya. Keberhasilan pemisahan antar klaster sangat bergantung pada penentuan jumlah klaster yang optimal dari suatu korpus data (Prasetya, Yanti, Purnamasari, Dikananda, & Nurdiawan, 2022).

Berdasarkan permasalahan dan penelitian terdahulu, diperlukan metode klasterisasi yang mampu mengelompokkan data dengan kepadatan berbeda serta mengandung noise. Penelitian oleh Chandra, dkk menunjukkan K-Means mampu mengelompokkan data

penjualan dengan kualitas klaster baik (DBI 0,438) (Purnama et al., 2022), dan Nirwana memperoleh pemisahan klaster sangat baik dengan silhouette score 0,9168 (Hendrastuty, 2024a). Namun, K-Means memiliki kelemahan karena harus menentukan jumlah klaster di awal dan sensitif terhadap distribusi data serta outlier. Sementara itu, penelitian Nanda , dkk membuktikan HDBSCAN mampu membentuk klaster sekaligus mendeteksi outlier pada data berkepadatan bervariasi (Wahyuni et al., 2021). Di sisi lain, metode seperti *Agglomerative Hierarchical Clustering* juga masih memiliki keterbatasan dalam menentukan struktur klaster optimal pada data yang tidak merata (Fitrianto, 2024).

Oleh karena itu, fokus tujuan dari penelitian ini untuk menggunakan algoritma HDBSCAN dalam mengelompokkan produk VAPE. Hasil dari pertimbangan penelitian terdahulu, algoritma ini dipilih karena kinerjanya yang lebih baik dibanding beberapa algoritma pengelompokan lain yang telah dibandingkan. Lalu, karena kemampuannya dalam membentuk klaster secara adaptif berbasis kepadatan, mampu menangani kepadatan kelompok yang bervariatif, menemukan pola tersembunyi, serta menangani noise dengan lebih baik dibandingkan metode clustering konvensional.

2. KAJIAN TEORITIS

A. Algoritma HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) merupakan pengembangan dari DBSCAN yang dirancang untuk mengatasi permasalahan variasi densitas dalam data. Berbeda dengan DBSCAN yang memerlukan parameter epsilon tetap, HDBSCAN mampu membangun struktur hierarki berdasarkan kepadatan data dan mengekstraksi cluster yang paling stabil (Wahyuni et al., 2021).

Beberapa studi baru menerapkan HDBSCAN pada berbagai domain (review aplikasi, kesehatan, astronomi, dan wilayah administratif). Kelebihan yang diulang-ulang adalah kemampuan menangani noise, variasi densitas, dan kemudahan dalam interpretasi klaster berdasar ketetapan densitas. Lalu studi (Juwita & Ali, 2024), (Handayani & Isnaini Rosyida, 2023), (Kaverinskiy et al., 2025) dapat dijadikan pedoman dalam Pemilihan fitur (aggregat transaksi: frekuensi, total, jam, kategori produk), Skala/normalisasi data, Penggunaan e*mbedding/dimensionality reduction* (UMAP/PCA) sebelum HDBSCAN jika data berdimensi tinggi, dan Validasi klaster secara statistik dan bisnis.

Proses HDBSCAN diawali dengan perhitungan jarak antar titik menggunakan rumus jarak Euclidean pada Persamaan (2.1) (Wahyuni et al., 2021).

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (2.1)$$

dimana

$d(A,B)$= jarak Euclidean antara titik data $x$dan y

A_i = nilai atribut ke-i pada titik data A

B_i = nilai atribut ke-i pada titik data B

i = indeks atribut yang dihitung

n = jumlah atribut atau dimensi data

A dan B = dua titik data yang akan dihitung jaraknya

Selanjutnya dihitung *core distance* untuk setiap titik, yaitu jarak ke tetangga terdekat ke-k berdasarkan parameter min_samples. *Core distance* menyatakan tingkat kepadatan lokal suatu titik. Setelah itu dihitung *mutual reachability distance* antara dua titik a dan b menggunakan Persamaan (2.2) (Sante et al., 2025).

$$\text{mreach}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\} \quad (2.2)$$

$mreach(a,b)$ = mutual reachability distance antara titik $a$dan b

$core_k(a)$ = core distance dari titik a terhadap tetangga ke-k

$core_k(b)$ = core distance dari titik b terhadap tetangga ke-k

$d(a,b)$ = jarak Euclidean antara titik $a$dan b

a dan b = dua titik data yang dibandingkan

k = jumlah tetangga terdekat yang ditentukan berdasarkan parameter *min_samples*

Nilai mutual *reachability distance* digunakan untuk membangun *minimum spanning tree* (MST). Dari MST tersebut dibentuk struktur hierarki cluster berdasarkan perubahan kepadatan. Cluster yang memiliki stabilitas tertinggi akan dipilih sebagai hasil akhir. Parameter utama dalam HDBSCAN adalah min_cluster_size dan min_samples. Parameter min_cluster_size menentukan jumlah minimum anggota dalam satu cluster, sedangkan min_samples menentukan sensitivitas terhadap kepadatan lokal. Pemilihan parameter yang tidak tepat dapat menyebabkan cluster terfragmentasi atau terlalu menyatu (Sante et al., 2025). Keunggulan utama HDBSCAN adalah kemampuannya dalam menangani variasi densitas, mendeteksi noise secara otomatis, serta tidak memerlukan penentuan jumlah cluster di awal proses (Kaverinskiy et al., 2025).

B. Silhouette Coefficient

Silhouette Coefficient merupakan salah satu metode evaluasi internal yang

digunakan untuk mengukur kualitas hasil klasterisasi. Metode ini mengukur seberapa baik suatu objek berada dalam cluster-nya dibandingkan dengan cluster lain (Piao Tan & A. Floudas, 2008). Untuk setiap titik data i, dihitung dua nilai utama, yaitu:

$a(i)$: rata-rata jarak antara titik $i$dengan seluruh titik lain dalam cluster yang sama.

$b(i)$: rata-rata jarak terkecil antara titik $i$dengan titik-titik pada cluster lain yang terdekat.

Silhouette Coefficient memiliki nilai rentang -1 hingga 1. Interpretasi nilai tersebut adalah sebagai berikut:

- Nilai mendekati 1 menunjukkan bahwa titik terklaster dengan sangat baik.
- Nilai mendekati 0 menunjukkan bahwa titik berada di batas antara dua cluster.
- Nilai mendekati -1 menunjukkan titik tidak terklaster dengan baik.

Rata-rata nilai *silhouette* dari seluruh data digunakan sebagai indikator kualitas keseluruhan cluster. Semakin tinggi nilai rata-rata *silhouette*, semakin baik kualitas pemisahan cluster.

3. METODE PENELITIAN

Pendekatan penelitian yang digunakan dalam penelitian ini adalah pendekatan kuantitatif dengan metode eksperimen berbasis data mining. Proses eksperimen dilakukan dengan melakukan processing data, transformasi data ke bentuk numerik, standarisasi, hingga penerapan algoritma HDBSCAN menggunakan parameter *min_cluster_size* dan *min_samples* yang berbeda-beda untuk melihat pengaruhnya terhadap hasil pengelompokan. Pendekatan data mining dalam penelitian ini meliputi data selection, data cleaning, transformasi data, pemodelan (*clustering*), serta evaluasi hasil menggunakan metode validasi berupa *silhouette score*. Algoritma HDBSCAN digunakan dalam desain penelitian ini untuk mengolah data penjualan produk vape. Dataset yang digunakan dalam penelitian ini adalah data penjualan vapestore.

4. HASIL DAN PEMBAHASAN

A. Analisis Eksplorasi Data

Distribusi Kategori Produk

Dari total 11.979 transaksi yang dianalisis, distribusi kategori menunjukkan pola yang jelas, yang biasanya terlihat pada bisnis ritel vape. Kategori Liquid mendominasi dengan 5.024 transaksi (41,9%), diikuti oleh Wire (yang mencakup coil dan cartridge)

dengan 4.909 transaksi (41,0%). Kedua kategori ini menyumbang lebih dari 82% dari seluruh aktivitas penjualan. Kategori lainnya memiliki porsi yang jauh lebih kecil: Isi Kulkas (613 transaksi), Device (517), Non Kategori (456), Accessories (261), Cotton (170), dan RDA/RTA (29).

Dominasi Liquid dan Wire ini mengindikasikan bahwa bisnis vapestore ini sangat bergantung pada produk habis pakai (consumables). Dalam hal clustering, ini berarti bahwa klaster-klaster besar dan sangat padat akan terbentuk di daerah Liquid dan Wire, sementara klaster yang lebih kecil dan renggang akan terbentuk di kategori seperti Device. Kondisi yang tidak merata inilah yang menjadi alasan utama pemilihan HDBSCAN, karena algoritma ini secara natural dapat menangani variasi kepadatan tanpa harus membentuk klaster dengan ukuran yang sama.

Analisis Dominasi Nama Produk

Pada tingkat Nama Produk, ditemukan konsentrasi yang cukup tinggi pada beberapa produk tertentu. Sepuluh produk teratas menyumbang lebih dari 60% dari total transaksi. Tiga produk teratas adalah Cat Oxva (1.523 transaksi), Foom Salt (1.384 transaksi), dan Cat Foom X (1.328 transaksi). Ketiga produk ini sendiri sudah mencakup 35% dari seluruh penjualan.

Hasil tersebut menunjukkan adanya produk-produk dominan yang berpotensi menjadi pusat kepadatan (*core density*). Sebaliknya, terdapat 328 nama produk lainnya yang tersebar pada sisa 40% transaksi, dengan banyak di antaranya hanya muncul kurang dari 10 kali selama tiga tahun. Produk-produk dengan frekuensi sangat rendah ini diprediksi akan terdeteksi sebagai noise oleh HDBSCAN, yang secara bisnis dapat diinterpretasikan sebagai produk slow-moving atau produk percobaan.

Tabel 1 Produk dengan Frekuensi Tertinggi

Peringkat	Nama Produk	Jumlah Transaksi	Persentase
1	Cat Oxva	1.523	12,7%
2	Foom Salt	1.384	11,6%
3	Cat Foom X	1.328	11,1%
4	Cat Ursa	416	3,5%
5	Min-Ral	410	3,4%

Preferensi Varian dan Sub Varian

Analisis pada atribut Varian mengungkap preferensi konsumen yang sangat jelas terhadap produk liquid dengan kadar nikotin tinggi. Varian Saltnic 30mg menjadi yang

paling dominan dengan 3.214 transaksi, jauh melampaui varian kedua yaitu "0.6 Ohm" (1.650 transaksi) yang merupakan spesifikasi coil. Varian "Freebase 3mg" berada di posisi keempat dengan 1.360 transaksi. Pola ini mengkonfirmasi tren pasar saat ini yang bergeser dari freebase ke saltnic. Pada atribut Sub varian, dominasi "Single Pack" (4.940 transaksi) dan "30ml" (3.664 transaksi) menunjukkan bahwa konsumen lebih memilih pembelian dalam kemasan kecil dan satuan.

Hasilnya, data penjualan vape ini memiliki karakteristik ideal untuk HDBSCAN: distribusi sangat tidak merata, terdapat beberapa inti kepadatan yang sangat kuat (Liquid Saltnic 30mg), dan terdapat ekor panjang (long tail) dari produk-produk yang jarang terjual yang kemungkinan akan menjadi noise. Penjelasan ini menjadi landasan penting untuk menggambarkan mengapa parameter min_cluster_size yang lebih besar akan menghasilkan klaster yang lebih bermakna pada tahap selanjutnya.

B. Implementasi HDBSCAN

Skenario Pengujian Parameter

Pada Keunggulan utama HDBSCAN adalah tidak diperlukannya parameter epsilon seperti pada DBSCAN. Namun, algoritma ini tetap memerlukan dua parameter utama yang sangat memengaruhi hasil, yaitu min_cluster_size dan min_samples. Secara konseptual, min_cluster_size berfungsi sebagai ambang batas minimal jumlah titik yang dapat dianggap sebagai sebuah klaster yang stabil, sementara min_samples mengontrol seberapa konservatif algoritma dalam menilai sebuah titik sebagai core point. Nilai min_samples yang rendah akan membuat algoritma lebih toleran terhadap noise, sedangkan nilai yang tinggi akan menghasilkan klaster yang lebih konservatif.

Rentang min_cluster_size diuji pada nilai 5, 10, 15, 20, dan 30, yang merepresentasikan spektrum dari klaster yang sangat spesifik hingga klaster yang sangat umum. Untuk min_samples, digunakan nilai 5, 10, dan 15. Kombinasi ini menghasilkan 15 skenario pengujian yang berbeda.

Proses Eksekusi Algoritma Secara Internal

Secara internal, algoritma melakukan tiga tahapan utama. Pertama, transformasi ruang berdasarkan jarak mutual reachability, di mana jarak antar titik dihitung tidak hanya berdasarkan jarak Euclidean murni, tetapi juga mempertimbangkan kepadatan lokal dari masing-masing titik. Kedua, pembangunan hierarki klaster menggunakan algoritma single linkage pada Minimum Spanning Tree yang telah ditransformasi. Ketiga, ekstraksi

klaster stabil berdasarkan parameter min_cluster_size yang diberikan, di mana cabang-cabang pada pohon hierarki yang memiliki stabilitas di bawah ambang batas akan dipangkas. Titik-titik yang tidak berhasil masuk ke dalam klaster stabil manapun secara otomatis diberi label -1 sebagai noise.

Mekanisme Evaluasi dan Pemilihan Parameter Optimal

Hasil pengujian menunjukkan pola yang konsisten dan dapat dijelaskan. Peningkatan min_cluster_size secara konsisten menurunkan jumlah klaster yang terbentuk, sementara nilai Silhouette Score naik-turun tergantung kombinasi parameter. Pada kombinasi dengan min_cluster_size besar (25–30), klaster yang tersisa hanya berjumlah 7, dengan noise sangat rendah (<1%) namun kurang baik karena beberapa klaster yang terbentuk pada rentang ini menggabungkan produk dengan karakteristik penjualan yang cukup berbeda, sehingga kurang bermanfaat secara bisnis meskipun nilai Silhouette-nya tampak tinggi.

Tabel 2 Pengujian Eksperimen

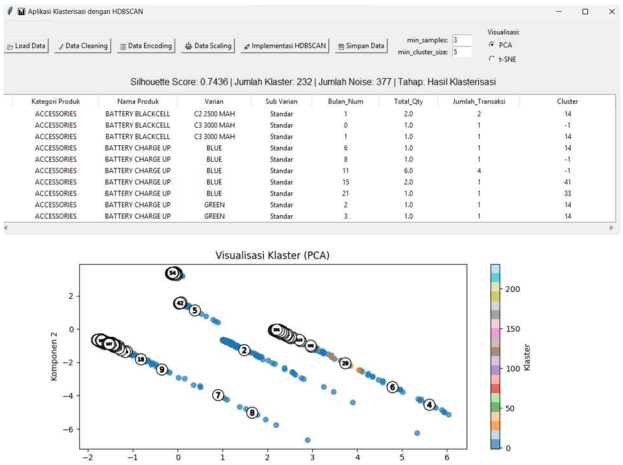
min_sample	min_cluster_size	Jumlah Cluster	Silhouette Score
2	5	230	0,7091
3	5	232	0,7436
5	10	77	0,5502
10	15	33	0,5029
15	20	19	0,4895
5	25	7	0,6553
10	5	82	0,5108

Berdasarkan Tabel di atas, terlihat bahwa kombinasi min_sample = 3 dan min_cluster_size = 5 menghasilkan parameter terbaik dengan nilai Silhouette Coefficient sebesar 0,7436 dan jumlah klaster sebanyak 232 klaster yang valid dengan noise bejumlah 377 produk. Nilai Silhouette yang mendekati 1 ini menunjukkan bahwa klaster yang terbentuk memiliki kualitas yang sangat tinggi sekaligus terpisah dengan sangat jelas dari klaster lainnya. Parameter ini kemudian ditetapkan sebagai parameter optimal dan digunakan untuk analisis hasil.

C. Hasil Pengelompokan Produk Vape

Tahap pengelompokan dilakukan dengan menggunakan parameter optimal yang diperoleh dari proses tuning, yaitu min_samples = 3 dan min_cluster_size = 5. Proses clustering dijalankan terhadap dataset hasil agregasi Produk × Bulan yang berjumlah 2.700 baris dengan 16 fitur. Hasil dari algoritma HDBSCAN pada aplikasi menunjukkan

kinerja yang baik dengan nilai Silhouette Score sebesar 0,7436.

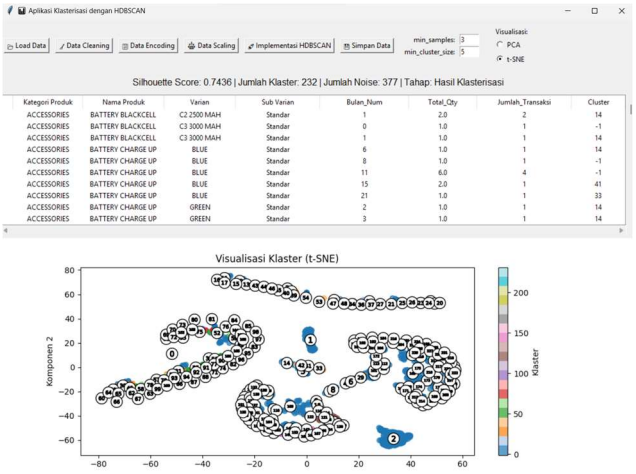


Gambar 1 Tampilan Visualisasi Klasterisasi HDBSCAN Menggunakan PCA

Secara keseluruhan, algoritma berhasil membentuk sebanyak 232 klaster dengan total 2.323 baris (86%) berhasil terkelompokkan. Sisanya, sebanyak 377 baris (14%), diklasifikasikan sebagai noise dengan label -1.

Interpretasi Visualisasi t-SNE

Visualisasi menggunakan t-SNE pada Gambar 2 memperlihatkan struktur pengelompokan yang lebih jelas terpisah dibandingkan visualisasi PCA. Sebagian besar klaster kecil (ditandai lingkaran putih bernomor) mengelompok rapat pada area kiri dan kanan atas merepresentasikan kombinasi produk-bulan dari kategori Liquid dan Wire yang mendominasi transaksi. Berbeda dengan hal tersebut, terdapat tiga area yang terpisah jauh dari kerumunan utama, ditandai dengan lingkaran besar bernomor 0, 1, dan 2 yang seluruhnya adalah klaster besar dengan karakteristik kategori produk yang khas dan konsisten.



Gambar 2 Visualisasi Cluster Menggunakan t-SNE

Pada **Klaster 2** (area kanan bawah, 119 anggota) seluruhnya berasal dari kombinasi Kategori Produk "ISI KULKAS" dengan Sub Varian "DINGIN", tersebar dari bulan_num=0 hingga bulan_num=18 dengan total_qty bervariasi antara 4 hingga 23 unit per bulan yang diartikan produk minuman dingin dengan pola penjualan reguler sepanjang periode data. Lalu, **Klaster 1** (85 anggota) didominasi kombinasi "COTTON"/"STANDAR" (produk kapas vape), sementara **Klaster 0** (16 anggota, area kiri) didominasi kombinasi "RDA/RTA"/"STANDAR" — kategori aksesoris niche dengan frekuensi transaksi rendah namun konsisten. Terpisahnya ketiga klaster ini secara jauh dari kerumunan utama pada bidang t-SNE mengonfirmasi bahwa kategori produk berhasil menjadi pembeda utama antar klaster.

Hasil klasterisasi juga memperlihatkan bagaimana atribut bulan_num memengaruhi klaster: produk "BATTERY CHARGE UP" varian BLUE misalnya, masuk ke Klaster 14 pada bulan_num=6, namun terdeteksi sebagai noise (Cluster -1) pada bulan_num=8 dan bulan_num=11 menjelaskan bahwa penjualan produk tersebut pada bulan-bulan tersebut bersifat tidak reguler dibandingkan pola penjualan produk sejenis lainnya pada bulan yang sama.

Analisis Produk Noise

Sebanyak 377 baris kombinasi produk-bulan yang terdeteksi sebagai noise umumnya merupakan kombinasi dengan jumlah_transaksi sangat rendah (1–2 transaksi dalam sebulan) menjelaskan bahwa penjualan yang jarang atau produk yang baru muncul/hampir tidak lagi dijual pada bulan tersebut. Berbeda dari pendekatan sebelumnya yang mengidentifikasi noise pada transaksi individual, pendekatan agregasi Produk × Bulan ini memungkinkan identifikasi noise yang lebih jelas: sebuah produk dapat "normal" (masuk klaster) pada bulan-bulan tertentu, namun menjadi "noise" pada bulan lain ketika penjualannya tidak reguler.

D. Evaluasi Hasil Clustering

Evaluasi dengan Silhouette Coefficient

Silhouette Coefficient dipilih karena mengukur dua prinsip dasar clustering, yaitu seberapa mirip sebuah objek dengan anggota dalam klasternya sendiri dibandingkan dengan klaster terdekat lainnya. Nilai Silhouette berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa objek ditempatkan pada klaster yang sangat tepat.

Pada skenario parameter optimal (min_samples = 3 dan min_cluster_size = 5),

diperoleh nilai Silhouette Coefficient sebesar 0,7436. Berdasarkan kategori interpretasi Kaufman & Rousseeuw (Kaufman & Rousseeuw, 1990). Nilai di atas 0,71 tergolong dalam kategori "struktur kuat" (*strong structure*), sehingga hasil ini menunjukkan pemisahan antar klaster yang baik.

Distribusi nilai Silhouette per klaster juga menunjukkan pola yang lebih sejenis. Klaster-klaster besar yang didominasi kategori Liquid dan Wire memiliki nilai rata-rata sekitar 0,71–0,72, sementara klaster-klaster kecil berisi kategori niche seperti RDA/RTA, Cotton, dan Accessories memiliki nilai rata-rata yang relatif setara, yaitu sekitar 0,73–0,74. Kemiripan rentang nilai ini mengindikasikan bahwa kualitas pemisahan klaster relatif konsisten di seluruh kategori produk, tidak hanya terkonsentrasi pada kategori dengan volume transaksi besar.

Analisis Kemampuan Penanganan Noise

Salah satu keunggulan utama HDBSCAN adalah kemampuannya untuk tidak memaksakan setiap titik masuk ke dalam klaster. Pada hasil optimal (min_samples=3, min_cluster_size=5), dari total 2.700 baris data kombinasi produk-bulan, sebanyak 377 baris (14,0%) dikategorikan sebagai noise.

Data yang terdeteksi sebagai noise memiliki rata-rata jumlah_transaksi sebesar 9,35 dan rata-rata total_qty sebesar 10,67 per bulan keduanya lebih tinggi dibandingkan rata-rata data yang berhasil masuk ke dalam klaster (jumlah_transaksi=3,63; total_qty=4,09). Distribusi kategori pada data noise didominasi oleh Wire (165 baris) dan Liquid (159 baris), yaitu dua kategori dengan volume transaksi terbesar dalam dataset.

Temuan ini diketahui bahwa noise pada hasil klasterisasi ini bukan merepresentasikan produk yang tidak laku, melainkan kombinasi produk-bulan dengan pola penjualan yang tidak biasa dibandingkan pola umum pada kategori dan periode yang sama baik berupa lonjakan penjualan pada bulan tertentu (misalnya akibat promosi atau musim tertentu) maupun kombinasi bulan-produk yang baru muncul dengan volume tidak konsisten dengan bulan-bulan sebelumnya. Sebagai contoh, kombinasi "CAT OXVA" pada bulan_num=16 tercatat memiliki total_qty=50 dari 49 transaksi, jauh di atas rata-rata bulanan produk tersebut pada periode lain, sehingga terdeteksi sebagai anomali oleh algoritma.

E. Pembahasan

Penjelasan Hasil terhadap Tujuan Penelitian

Penelitian ini memiliki tiga tujuan utama, dan seluruhnya terjawab melalui hasil implementasi. Pertama, terkait pengelompokan produk berdasarkan karakteristik penjualan, HDBSCAN berhasil membuktikan kemampuannya dalam membentuk klaster tanpa perlu menentukan jumlah klaster di awal. Terbentuknya 232 klaster secara otomatis menunjukkan bahwa data penjualan Permata Vapestore secara alami memiliki setidaknya 232 pola pembelian yang berbeda.

Kedua, terkait evaluasi kualitas, nilai Silhouette Score sebesar 0,7436 secara statistik mengkonfirmasi bahwa klaster yang terbentuk memiliki kohesi internal yang sangat tinggi. Artinya, kombinasi produk-bulan di dalam satu klaster, seperti Klaster 2 yang seluruhnya berisi "ISI KULKAS" dengan Sub Varian "DINGIN" pada berbagai bulan, atau Klaster 1 yang berisi "COTTON" dengan Sub Varian "STANDAR", memang memiliki kemiripan karakteristik yang nyata, bukan hasil pengelompokan acak. Nilai ini bahkan melampaui penelitian terdahulu oleh Wahyuni dkk. (2020) yang memperoleh Silhouette 0,72 menggunakan K-Means pada data retail.

Ketiga, terkait identifikasi noise, ditemukannya 377 kombinasi produk-bulan (14,0% dari total data) sebagai noise memberikan kontribusi praktis yang berbeda dari dugaan awal penelitian. Noise didominasi oleh kombinasi produk-bulan dengan volume transaksi yang tidak biasa dibandingkan pola umum pada kategori yang sama, baik berupa lonjakan penjualan tak terduga (seperti "CAT OXVA" pada bulan tertentu dengan 49 transaksi) maupun pola yang menyimpang dari kebiasaan bulan-bulan sebelumnya. Kemampuan HDBSCAN memisahkan anomali semacam ini secara otomatis tetap menjadi kontribusi pada penelitian, meskipun penjelasan bisnisnya perlu disesuaikan: bukan sekadar "daftar produk untuk didiskon", melainkan indikator bulan-bulan yang memerlukan penelusuran lebih lanjut oleh manajemen (misalnya untuk mengevaluasi efektivitas promosi atau mendeteksi anomali pencatatan).

Keunggulan HDBSCAN Dibandingkan Metode Berbasis Centroid

Hasil visualisasi t-SNE pada Gambar 4.2 memberikan visual yang kuat mengenai mengapa HDBSCAN lebih unggul untuk kasus ini. Struktur data seperti ini, yang memiliki variasi kepadatan ekstrem, merupakan kelemahan bagi algoritma berbasis centroid seperti K-Means. K-Means akan cenderung memecah mega-klaster pusat menjadi beberapa bagian yang tidak natural, atau sebaliknya, memaksa klaster-klaster kecil di pinggir untuk bergabung.

Pendekatan berbasis kepadatan dengan HDBSCAN dapat mempertahankan klaster sekaligus tetap mengakui klaster-klaster kecil yang valid. Kemampuan untuk tidak memaksakan setiap titik masuk ke dalam klaster (*noise handling*) juga menjadi pembeda utama, di mana K-Means akan tetap memasukkan 377 transaksi anomali tersebut ke dalam klaster terdekat, sehingga mengotori kemurnian klaster.

Manajerial bagi Permata Vapestore

Secara praktis, hasil 232 klaster kombinasi produk-bulan ini dapat diterjemahkan menjadi beberapa rekomendasi strategi bisnis yang lebih presisi dibandingkan pengelompokan berdasarkan jenis produk semata.

Jaminan stok untuk kategori dengan pola stabil. Klaster besar seperti Klaster 2 (kombinasi "ISI KULKAS"/"DINGIN", 119 anggota) dan Klaster 1 (kombinasi "COTTON"/"STANDAR", 85 anggota) menunjukkan pola penjualan yang konsisten dan berulang di hampir seluruh periode 25 bulan data. Kategori-kategori ini dapat dijadikan prioritas dalam perencanaan stok rutin, karena polanya yang stabil memudahkan kebutuhan bulanan tanpa risiko besar.

Data noise justru didominasi kombinasi produk-bulan dengan volume transaksi yang tidak biasa baik lonjakan (seperti "CAT OXVA" yang mencatat 49 transaksi pada satu bulan tertentu, jauh di atas rata-rata bulan lain) maupun penurunan mendadak. Sebanyak 165 dari 377 baris noise (43,8%) berasal dari kategori Wire dan 159 baris (42,2%) dari kategori Liquid dua kategori dengan volume transaksi terbesar. Oleh karena itu, rekomendasi yang lebih tepat adalah memprioritaskan penelusuran bulan-bulan anomali pada kategori Liquid dan Wire, untuk mengevaluasi apakah lonjakan disebabkan oleh promosi yang berhasil, atau justru kemungkinan kesalahan pencatatan transaksi yang perlu dikoreksi.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma HDBSCAN berhasil diterapkan untuk mengelompokkan data penjualan produk vape berdasarkan kombinasi karakter pada produk (Kategori Produk, Sub Varian) dan pola penjualan bulanan (volume dan frekuensi transaksi). Penelitian ini melakukan pendekatan dengan mengagregasi data transaksi ke tingkat kombinasi produk-bulan, sehingga dihasilkan 2.700 baris data dari 11.979 transaksi valid yang telah melalui tahap *data cleaning*. Dari 7 kali pengujian kombinasi parameter, parameter paling optimal yang

dipilih berdasarkan pertimbangan gabungan antara jumlah klaster, proporsi noise, dan nilai Silhouette Score adalah min_samples=3 dan min_cluster_size=5, dengan Silhouette Score sebesar 0,7436 tergolong kategori struktur klaster yang kuat menurut Kaufman & Rousseeuw. Hasil klasterisasi membentuk 232 klaster yang mencerminkan variasi pola penjualan bulanan yang kompleks, dengan 377 kombinasi produk-bulan (14,0%) terklasifikasi sebagai noise.

Berdasarkan analisis eksplorasi data pada tingkat transaksi, produk dengan frekuensi penjualan tertinggi selama periode penelitian adalah Cat Oxva (1.523 transaksi), Foom Salt (1.384 transaksi), dan Cat Foom X (1.328 transaksi), yang bersama-sama menyumbang lebih dari 35% dari seluruh transaksi. Klaster dengan jumlah anggota terbanyak dari hasil pengelompokan didominasi oleh kombinasi Kategori Produk "Isi Kulkas" dengan Sub Varian "Dingin" (Klaster 2, 119 anggota) serta kombinasi "Cotton" dengan Sub Varian "Standar" (Klaster 1, 85 anggota), yang menunjukkan pola penjualan bulanan yang konsisten sepanjang periode data. Analisis terhadap data noise menemukan bahwa kombinasi produk-bulan yang terklasifikasi sebagai noise justru memiliki rata-rata volume dan frekuensi transaksi yang lebih tinggi dibandingkan data yang berhasil dikelompokkan. Temuan ini diindikasikan bahwa noise dalam penelitian ini merepresentasikan pola penjualan yang tidak biasa (anomali, baik berupa lonjakan maupun penurunan) dibandingkan pola umum pada kategori dan periode yang sama.

DAFTAR REFERENSI

Fitrianto, N. (2024). *PERBANDINGAN ALGORITMA HDBSCAN DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING DALAM KLASTERISASI PADA DATA YANG MENGANDUNG*. *08*(02), 122–135.

Handayani, F. D., & Isnaini Rosyida. (2023). Clustering Review Pengguna Aplikasi Zenius pada Layanan Google Play Store Menggunakan Metode DBSCAN dan HDBSCAN. *Emerging Statistics and Data Science Journal*, *1*(2), 178–191. https://doi.org/10.20885/esds.vol1.iss.2.art19

Hendrastuty, N. (2024a). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika Dan Ilmu Komputer (JIMA-ILKOM)*, *3*(1), 46–56. https://doi.org/10.58602/jima-ilkom.v3i1.26

Hendrastuty, N. (2024b). Pengertian Analisis Data. *Jurnal Ilmiah Informatika Dan Ilmu Komputer (Jima-Ilkom)*, *3*(1), 46–56.

Juwita, I., & Ali, I. (2024). Penerapan Pola Penjualan Dengan Menggunakan Metode Algoritma Asosiasi Fp-Growth Bertujuan Untuk Meningkatkan Penjualan Kopi Di Point Coffee. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), 1600–1607.

https://doi.org/10.36040/jati.v8i2.9025

Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data: An Introduction To Cluster Analysis. In *Wiley, New York. ISBN 0-471-87876-6.* https://doi.org/10.2307/2532178

Kaverinskiy, V., Chaikovsky, I., Mnevets, A., Ryzhenko, T., Bocharov, M., & Malakhov, K. (2025). Scalable Clustering of Complex ECG Health Data: Big Data Clustering Analysis with UMAP and HDBSCAN. *Computation*, *13*(6), 1–21. https://doi.org/10.3390/computation13060144

Maulana, I., Witanti, W., Melina, Informatika, J. T., Jenderal, U., Yani, A., & Artikel, I. (2025). *Klasterisasi Gaya Belajar Mahasiswa Berbasis VARK dengan Algoritma DBSCAN untuk Personalisasi E-Learning*. 1619–1629. https://doi.org/10.33364/algoritma/v.22-2.2980

Nisa, K. (2021). *Penerapan Data Mining terhadap Data Transaksi sebagai Pendukung Informasi Strategi Penjualan menggunakan Algoritma Apriori*. *06*, 306–315.

Piao Tan, M., & A. Floudas, C. (2008). Determining the Optimal Number of Clusters. *Encyclopedia of Optimization*, *1*, 687–694. https://doi.org/10.1007/978-0-387-74759-0_123

Prasetya, T., Yanti, J. E., Purnamasari, A. I., Dikananda, A. R., & Nurdiawan, O. (2022). Analisis Data Transaksi Terhadap Pola Pembelian Konsumen Menggunakan Metode Algoritma Apriori. *INFORMATICS FOR EDUCATORS AND PROFESSIONAL : Journal of Informatics*, *6*(1), 43. https://doi.org/10.51211/itbi.v6i1.1688

Purnama, C., Sabrina, P. N., Studi, P., Informatika, T., Jenderal, U., & Yani, A. (2022). *MENGGUNAKAN K-MEANS*. *04*(01), 35–38.

Sante, A., Font, A. S., Mistry, D., Ortega-Martorell, S., & Olier, I. (2025). *Optimized HDBSCAN clustering for reconstructing the merger history of the Milky Way: applications and limitations*. *15*(September), 1–15.

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, *110*. https://doi.org/10.1016/j.ssresearch.2022.102817

Wahyuni, N. A., Hayati, M. N., & Rizki, N. A. (2021). Metode Hierarchical Density-Based Spatial Clustering of Application with Noise (HDBSCAN) Pada Wilayah Desa / Kelurahan Tertinggal di Kabupaten Kutai Kartanegara (Studi Kasus : Data Hasil Pendataan Potensi Desa Tahun 2018). *Eksponensial*, *12*(1), 47–52.