Analisis Komparatif K-Means dan Hierarchical Clustering Menggunakan Validasi Internal Clustering

Andri Agustian*[1], Shafwan Awaludin[2], Muhammad Raihan Ilham[3], Epsilona Katiga Capricorna[4], Nur Ananda Rumi[5], Vitri Tundjungsari[6]
[1,2,3,4,5,6]Jurusan Teknik Komputer, Universitas Esa Unggul, Bekasi
Jl. Harapan Indah Boulevard No.2, Kota Harapan Indah, Bekasi
e-mail: *[1]andriagustian182@gmail.com, [2]Shafwan123awaludin@gmail.com,
[3]raihanilham413@gmail.com, [4]epsilonakatigacapricorna@gmail.com,
[5]araandraaaaaa@gmail.com, [6]vitri.tundjungsari@esaunggul.ac.id

Abstrak

Clustering termasuk metode teknik unsupervised learning yang diterapkan guna mengklasifikasi angka tanpa label kelas. Kajian ini mengontraskan metode K-Means dengan Hierarchical Clustering terhadap himpunan data penjualan daring retail sebanyak 581.586 entri. Tahapan pengelompokan dijalankan dengan memanfaatkan variabel Quantity, UnitPrice, dan variabel turunan TotalPrice pasca fase penyaringan serta penyeragaman data. Pengukuran mutu kelompok dijalankan dengan evaluasi dalam memanfaatkan Silhouette Coefficient, Davies-Bouldin Index, dan Calinski-Harabasz Index untuk situasi tiga kelompok. Output uji coba membuktikan K-Means menghasilkan skor Silhouette Coefficient sejumlah 0,45, Davies-Bouldin Index sejumah 1,20, dan Calinski-Harabasz Index sejumlah 3200, sementara itu, Hierarchical Clustering mencapai skor lebih kecil untuk ketiga parameter tersebut. Jarak skor metrik menggambarkan susunan kelompok output K-Means semakin padat dan juga mempunyai jarak antar kelompok yang semakin nyata. Manfaat kajian ini terdiri dari penyampaian komparasi kuantitatif berdasarkan evaluasi dalam sebagai acuan penentuan metode pengelompokkan untuk himpunan data penjualan kuantitaif berukuran luas.

Kata kunci*—clustering, k-means, hierarchical clustering, validasi internal, unsupervised learning*

Abstract

Clustering is an unsupervised learning technique applied to classify numbers without class labels. This study contrasts the K-Means method with Hierarchical Clustering on a retail online sales dataset of 581,586 entries. The clustering stage is carried out using the variables Quantity, UnitPrice, and the derived variable TotalPrice after the data filtering and standardization phase. Group quality measurement is carried out with an internal evaluation using the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index for a three-group situation. The trial output shows that K-Means produces a Silhouette Coefficient score of 0.45, Davies-Bouldin Index of 1.20, and Calinski-Harabasz Index of 3200, while Hierarchical Clustering achieves a lower score for all three parameters. The distance score metric illustrates that the K-Means output cluster arrangement is denser and also has a more significant distance between groups. The benefits of this study consist of providing quantitative comparisons based on in-depth evaluation as a reference for determining grouping methods for large-scale quantitative sales data sets.

Keywords*—clustering, k-means, hierarchical clustering, validasi internal, unsupervised learning*

1. PENDAHULUAN

Pertumbuhan teknologi data sudah menciptakan volume informasi transaksi yang masif, menuntut metode analisis yang efisien buat mengekstraksi pola tersembunyi tanpa label kelas. Clustering selaku tata cara *unsupervised learning* jadi pemecahan utama buat mengelompokkan informasi bersumber pada kemiripan ciri, dengan aplikasi luas dalam segmentasi pelanggan, analisis sikap, serta eksplorasi data[1]. Di antara bermacam algoritma, K- Means serta *Hierarchical Clustering* ialah 2 pendekatan sangat terkenal dalam riset serta praktik [2]

K- Means memaksimalkan pengelompokan bersumber pada jarak ke centroid, sesuai buat informasi numerik berskala besar, sedangkan *Hierarchical Clustering* membangun struktur bertingkat yang menggambarkan ikatan hierarkis antar informasi. Perbandingan prinsip ini menimbulkan keduanya mempunyai ciri performa yang berbeda, apalagi pada dataset yang sama [3]. Tetapi, dalam praktiknya, pemilihan algoritma kerapkali didasarkan pada preferensi subjektif ataupun Kerutinan, bukan penilaian objektif berbasis metrik kuantitatif.

Penilaian objektif dalam clustering tanpa label sangat tergantung pada validasi internal, semacam *Silhouette Coefficient*, Davies- Bouldin Index, serta Calinski- Harabasz Index. Tiap-tiap metrik memperhitungkan aspek berbeda: kohesi serta pembelahan cluster, rasio jarak intra- serta antar- cluster, dan rasio variansi. Pemakaian multi- metrik secara simultan bisa membagikan cerminan komprehensif tentang mutu clustering [4]

Beberapa riset sudah menyamakan K- Means serta *Hierarchical Clustering* pada bermacam domain. Tetapi, sebagian besar cuma memakai satu metrik penilaian ataupun analisis visual, sehingga hasilnya terbatas serta kurang komprehensif [5], [6]. Penemuan penelitian-penelitian terdahulu pula kerap tidak tidak berubah- ubah; sebagian menyebut K- Means lebih unggul pada informasi numerik [7], sedangkan yang lain memberi tahu keunggulan *Hierarchical Clustering* pada informasi berstruktur hierarkis [8]. Inkonsistensi ini menampilkan kalau performa kedua algoritma sangat dipengaruhi oleh ciri informasi serta metrik penilaian yang digunakan.

Dengan demikian, ada *research gap* berbentuk: belum terdapatnya kajian komparatif sistematis yang secara simultan menggunakan 3 metrik validasi internal utama(*Silhouette Coefficient*, Davies- Bouldin Index, serta Calinski- Harabasz Index) buat mengevaluasi serta menyamakan kinerja K- Means serta *Hierarchical Clustering* pada dataset transaksi ritel berskala besar.

Bersumber pada latar balik serta *research gap* di atas, riset ini diformulasikan buat menanggapi pertanyaan- pertanyaan berikut:

1. Gimana performa algoritma K- Means serta *Hierarchical Clustering* dalam mengelompokkan informasi transaksi ritel numerik bila dievaluasi memakai 3 metrik validasi internal(*Silhouette Coefficient*, Davies- Bouldin Index, serta Calinski- Harabasz Index) secara simultan?
2. Tata cara manakah yang menciptakan mutu clustering lebih baik bersumber pada konsistensi hasil ketiga metrik validasi internal tersebut?
3. Aspek algoritmik apa yang menimbulkan perbandingan performa antara kedua tata cara pada dataset transaksi ritel berskala besar?

Tujuan Riset ini merupakan menganalisis serta menyamakan performa K- Means serta *Hierarchical Clustering* memakai pendekatan multi- metrik validasi internal secara komprehensif.

Donasi riset merupakan penyediaan analisis komparatif berbasis fakta kuantitatif yang bisa jadi acuan dalam pemilihan algoritma clustering cocok ciri informasi numerik berskala besar, dan pengayaan literatur menimpa penilaian objektif dalam *unsupervised learning*.

2. METODE PENELITIAN

Dibagian ini menarangkan tentang langkah- langkah riset secara terstruktur buat menyamakan metode K- Means serta Hierarchical Clustering lewat validasi internal cluster. Metodologi riset dirancang bersumber pada eksperimen komputasional, yang berisi tentang pengumpulan data, pemprosesan informasi dini, pelaksanaan algoritma pengelompokan, penerapan eksperimen yang kesekian, serta evaluasi kinerja pengelompokan dengan memakai indeks validasi internal. dari Seluruh langkah analisis menggunakan bahasa pemrograman Python dengan dataset berbentuk spreadsheet selaku sumber utama informasi. pada Uraian ini bertujuan buat membagikan cerminan tentang alur riset dan prosedur teknis yang diterapkan supaya hasilnya bisa bertabiat objektif serta direplikasi [5].

2.1 Jenis dan Pendekatan Penelitian

Pendekatan yang digunakan dalam riset ini merupakan kuantitatif lewat tata cara eksperimen komputasional di bidang informasi mining, paling utama dalam konteks unsupervised learning. Pendekatan kuantitatif diseleksi sebab fokus riset ini merupakan pada pengolahan informasi numerik dan pengukuran kinerja algoritma pengelompokan memakai metrik validasi internal. Riset ini bertabiat komparatif sebab menyamakan 2 algoritma pengelompokan, ialah K-Means serta Hierarchical Clustering, bersumber pada mutu hasil pengelompokan pada dataset yang sama [6]. Alur totalitas riset disajikan dalam wujud flowchart pada Foto 1, yang menggambarkan tahapan pengambilan informasi, pemrosesan dini, pelaksanaan algoritma K-Means serta Hierarchical Clustering, penilaian validasi internal, serta analisis perbandingan dan penarikan kesimpulan.

2.2 Informasi serta Sumber Data

Dataset yang digunakan dalam riset ini diperoleh dari salah satu web yang sediakan dataset publik serta ada dalam format CSV(Comma Separated Values). Dataset ini mencakup informasi transaksi ritel online yang terdiri dari 581. 586 baris dan 8 atribut utama. Sebab dataset ini tidak mempunyai label kelas, hingga sangat sesuai buat pelaksanaan tata cara unsupervised learning.

Adapun delapan atribut yang terdapat dalam dataset adalah sebagai berikut.

1. InvoiceNo: Nomor transaksi.
2. StockCode: Kode produk.
3. Description: Deskripsi produk.
4. Quantity: Jumlah barang yang dibeli.
5. InvoiceDate: Tanggal transaksi.
6. UnitPrice: Harga satuan produk.
7. CustomerID: Identitas pelanggan.
8. Country: Negara asal pelanggan.

Dalam penelitian ini, fokus analisis dilakukan pada atribut numerik yang relevan untuk proses clustering, yaitu Quantity dan UnitPrice. Selain itu, dilakukan proses transformasi data untuk membentuk fitur turunan seperti TotalPrice (hasil perkalian Quantity dan UnitPrice) guna memperkaya informasi yang digunakan dalam pengelompokan. Pemilihan atribut ini bertujuan untuk merepresentasikan pola perilaku transaksi pelanggan berdasarkan kuantitas pembelian dan nilai transaksi.

Tabel 1 Statistik Deskriptif Atribut Numerik

Statistik	Quantity	UnitPrice	TotalPrice
Mean	9.55	4.61	44.02
Std Dev	218.08	96.76	85123A
Min	-80,995	-11,062.06	-168,469.60
25%	1	1.25	3.75

Median (50%)	3	2.55	9.75
75%	10	4.13	41.30
Max	80,995	38,970.00	168,469.60

Dataset ini dipilih karena merepresentasikan kondisi data nyata yang berskala besar, tidak memiliki label kelas, serta mengandung potensi pola tersembunyi yang dapat dieksplorasi melalui metode clustering [7].

2.3 Pra-Pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk memastikan kualitas data sebelum proses clustering. Langkah-langkah pra-pemrosesan yang diterapkan meliputi:

1. Pembersihan data
   Data yang memiliki nilai kosong (*missing value*), khususnya pada atribut CustomerID, dihapus karena tidak dapat digunakan dalam analisis. Selain itu, data transaksi dengan nilai Quantity ≤ 0 dan UnitPrice ≤ 0 juga dihapus karena dianggap sebagai transaksi tidak valid (retur atau kesalahan input).
2. Seleksi Atribut
   Dari delapan atribut awal, dipilih atribut numerik yang relevan, yaitu Quantity, UnitPrice, dan TotalPrice. Atribut non-numerik tidak digunakan karena algoritma clustering berbasis perhitungan jarak.
3. Transformasi Data
   Dilakukan pembentukan fitur baru TotalPrice untuk merepresentasikan nilai transaksi setiap baris data.
4. Standarisasi Data
   Seluruh atribut numerik distandarisasi menggunakan metode *StandardScaler* agar memiliki rata-rata 0 dan standar deviasi 1. Proses ini bertujuan menghindari dominasi atribut dengan skala besar terhadap perhitungan jarak.

Tabel 2 Ringkasan Proses Pra-Pemrosesan Data

Tahap Preprocessing	Jumlah Baris	Jumlah Atribut	Persentase Data Tersisa	Keterangan
Data Awal	581,586	8	100%	Dataset mentah dari CSV
Setelah Pembersihan Missing Values	406,829	8	69.95%	Hapus baris dengan CustomerID kosong
Setelah Filter Quantity > 0	401,604	8	69.05%	Hapus transaksi retur/negatif
Setelah Filter UnitPrice > 0	397,884	8	68.42%	Hapus harga invalid
Setelah Transformasi Fitur	397,884	4	68.42%	Tambah kolom TotalPrice

Setelah Seleksi Atribut	397,884	3	68.42%	Hanya Quantity, UnitPrice, TotalPrice
Setelah Standarisasi	397,884	3	68.42%	Data siap untuk clustering

Tahapan pra-pemrosesan ini penting karena algoritma K-Means dan Hierarchical Clustering sangat sensitif terhadap perbedaan skala data. Standarisasi memastikan setiap atribut memiliki kontribusi yang seimbang dalam proses pengelompokan, sebagaimana diterapkan pada penelitian clustering data berbasis K-Means pada domain ketenagakerjaan[8].

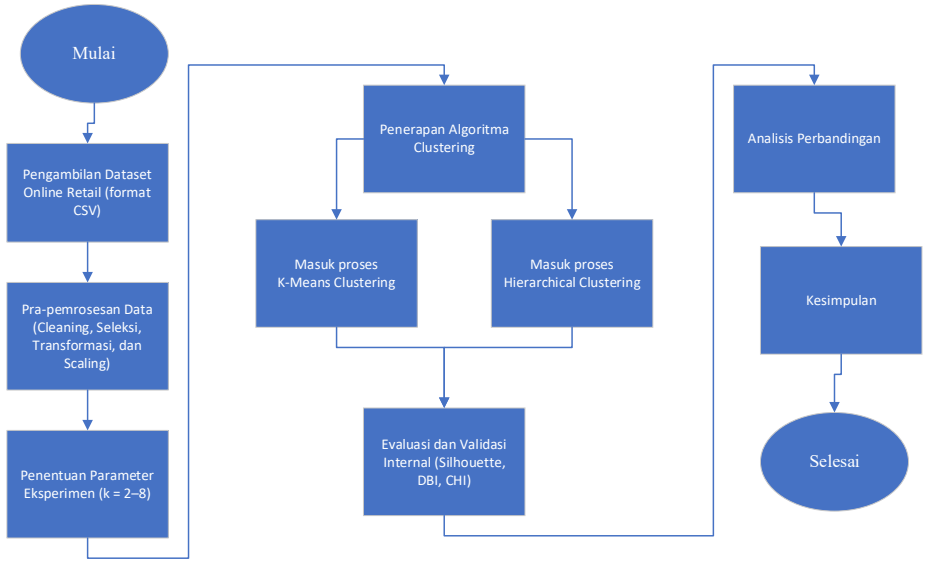
2.4 Prosedur Eksperimen

Penelitian ini menggunakan skema eksperimen berulang sebanyak tujuh kali untuk meningkatkan reliabilitas hasil. Eksperimen dilakukan dengan variasi jumlah cluster, yaitu k = 2 sampai k = 8. Setiap nilai k diuji pada kedua metode clustering.

Prosedur eksperimen dilakukan sebagai berikut:

1. Menentukan nilai k (jumlah cluster) dari 2 hingga 8.
2. Menerapkan algoritma K-Means dan Hierarchical Clustering untuk setiap nilai k.
3. Menghasilkan label cluster untuk setiap eksperimen.
4. Menghitung nilai indeks validasi internal untuk masing-masing metode.
5. Mencatat dan membandingkan hasil evaluasi setiap eksperimen.

Dengan demikian, total dilakukan tujuh skenario eksperimen untuk masing-masing metode, sehingga diperoleh gambaran performa algoritma pada berbagai jumlah cluster.



Gambar 1. Flowchart Metodologi Penelitian Clustering

2.5 Implementasi Algoritma

2.5.1 K-Means Clustering

Algoritma K- Means diimplementasikan memakai pustaka scikit- learn. Jumlah cluster didetetapkan cocok skenario eksperimen. Proses clustering dicoba secara iteratif sampai

konvergen dengan meminimalkan guna objektif berbentuk jumlah kuadrat jarak Euclidean antara informasi serta centroid cluster.

Implementasi menggunakan pustaka *scikit-learn* dengan parameter, *n_init* = 10, *max_iter* = 300, *algorithm* = '*lloyd*', dan *random_state* = 42 untuk memastikan reproducibility

2.5.2 Hierarchical Clustering

Hierarchical Clustering memakai tata cara agglomerative dengan ward linkage serta metrik jarak Euclidean. Tata cara ward diseleksi sebab meminimalkan variansi dalam cluster, berbeda dengan tata cara linkage lain semacam *single* (jarak minimum) ataupun complete (jarak maksimum) yang tidak digunakan dalam riset ini.

2.6 Evaluasi dan Validasi

Sebab dataset tidak mempunyai label kelas, penilaian dicoba memakai validasi internal cluster. Indeks yang digunakan meliputi:

1. Silhouette Coefficient
   Mengukur tingkatan kemiripan informasi dalam satu cluster dibanding cluster lain. Nilai berkisar antara- 1 sampai 1. Terus menjadi mendekati 1, mutu clustering terus menjadi baik.
2. Davies-Bouldin Index (DBI)
   Mengukur rasio keakraban intra- cluster terhadap jarak antar- cluster. Nilai yang lebih kecil menampilkan mutu clustering yang lebih baik.
3. Calinski-Harabasz Index (CHI)
   Mengukur rasio alterasi antar- cluster terhadap alterasi dalam cluster. Nilai yang lebih besar menampilkan hasil clustering yang lebih maksimal.

Penilaian dicoba pada tiap eksperimen (k= 2 hingga 8) buat kedua tata cara. Hasil penilaian setelah itu dibanding secara kuantitatif. Tata cara clustering dikatakan lebih unggul apabila menciptakan nilai Silhouette serta CHI yang lebih besar dan nilai DBI yang lebih rendah secara tidak berubah- ubah.

2.7 Teknik Analisis Data

Analisis informasi dicoba dengan menyamakan hasil validasi internal dari kedua tata cara pada tiap skenario eksperimen. Berikutnya dicoba analisis deskriptif terhadap ciri tiap cluster yang tercipta buat menguasai pola transaksi pelanggan. Kesimpulan ditarik bersumber pada kecenderungan performa tata cara pada segala skenario eksperimen.

2.8 Perangkat dan Lingkungan Penelitian

Riset ini memakai bahasa pemrograman Python dengan area Jupyter Notebook. Pustaka yang digunakan meliputi:

- pandas untuk manipulasi data
- numpy untuk komputasi numerik
- scikit-learn untuk implementasi clustering dan evaluasi
- scipy untuk perhitungan jarak dan analisis tambahan

Seluruh proses pengolahan data, eksperimen, serta evaluasi dilakukan menggunakan perangkat lunak tersebut.

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil eksperimen dan ulasan komprehensif menimpa perbandingan kinerja algoritma *K- Means* serta *Hierarchical Clustering* dalam mengelompokkan informasi transaksi online retail. Penilaian dicoba memakai 3 metrik validasi internal, ialah *Silhouette Coefficient, Davies- Bouldin Index, serta Calinski- Harabasz Index*. Tidak hanya menyajikan nilai

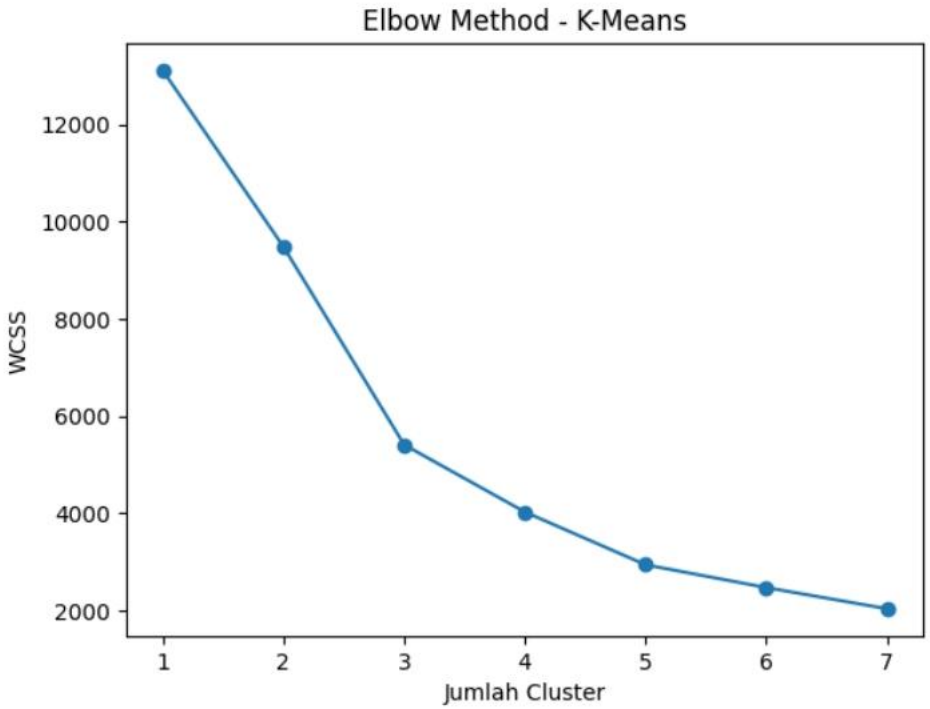
numerik, ulasan difokuskan pada interpretasi hasil eksperimen serta keterkaitannya dengan teori *clustering*[9].

3.1 Hasil Clustering dan Analisis Tata cara K-Means

Penentuan jumlah cluster maksimal ialah sesi krusial dalam clustering sebab sangat mempengaruhi mutu hasil pengelompokan. Pada tata cara *K- Means*, jumlah cluster didetetapkan memakai Elbow Method. Bersumber pada grafik Elbow yang ditunjukkan pada Foto Elbow Method, nampak penyusutan nilai WCSS(*Within- Cluster Sum of Squares*) yang signifikan dari k= 1 sampai k= 3. Sehabis k= 3, penyusutan WCSS cenderung melandai serta tidak lagi signifikan.

Fenomena ini menampilkan kalau akumulasi jumlah cluster di atas 3 tidak membagikan kenaikan mutu yang berarti. Secara teoritis, keadaan ini diucap selaku titik“ siku”(elbow point), yang menunjukkan jumlah cluster maksimal. Oleh sebab itu, jumlah cluster terbaik dalam riset ini diresmikan sebanyak 3 cluster. Hasil ini sejalan dengan riset terdahulu yang menyamakan proses klasterisasi memakai *K- Means* serta *Agglomerative Hierarchical Clustering*, di mana penentuan jumlah cluster maksimal sangat mempengaruhi terhadap mutu hasil pengelompokan[10].

Sedangkan itu, pada *Hierarchical Clustering*, penentuan jumlah cluster dicoba lewat interpretasi dendrogram. Bersumber pada dendrogram yang dihasilkan, terdapat jarak pembelahan yang lumayan besar pada rentang tertentu. Pemotongan dendrogram pada jarak tersebut menciptakan 3 kelompok utama. Perihal ini menguatkan keputusan kalau 3 cluster ialah jumlah yang sangat representatif buat struktur informasi *online retail*.



Gambar 2. Elbow Method untuk K-Means

Gambar 3.1 menampilkan grafik Elbow Method pada algoritma *K-Means*. Terlihat penurunan nilai WCSS yang signifikan hingga k = 3, kemudian melandai pada nilai k berikutnya. Hal ini menunjukkan bahwa jumlah cluster optimal berada pada k = 3.

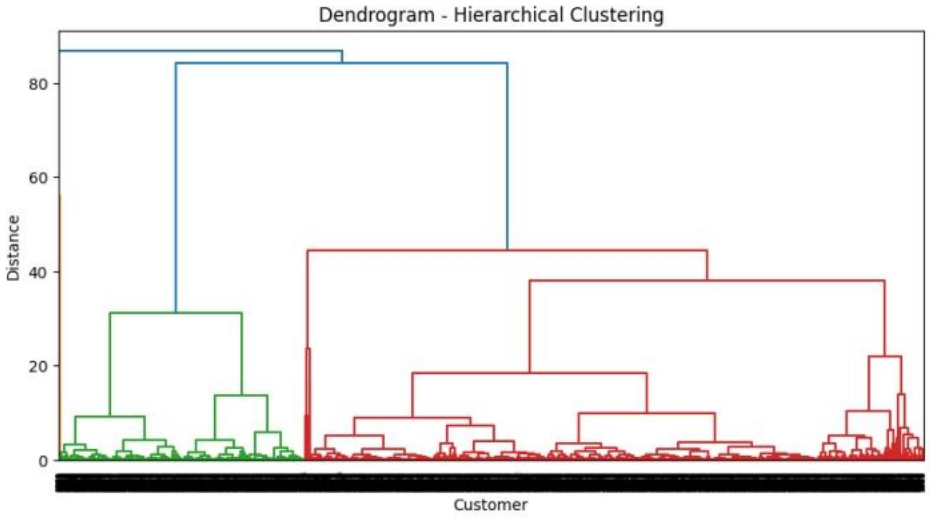
3.2 Hasil Clustering serta Analisis Tata cara Hierarchical Clustering

Sehabis jumlah cluster diresmikan sebanyak 3, algoritma *K- Means* diterapkan pada dataset yang sudah lewat sesi pra- pemrosesan. Hasil clustering menampilkan kalau informasi dibagi jadi 3 kelompok dengan ciri yang berbeda.

Secara konseptual, *K- Means* bekerja dengan meminimalkan jarak antara informasi serta *centroid* cluster. Oleh sebab itu, cluster yang tercipta cenderung mempunyai wujud yang kompak serta terpisah secara jelas. Perihal ini nampak dari hasil visualisasi serta nilai validasi internal yang menampilkan tingkatan kohesi yang lumayan besar di dalam cluster.

Cluster awal didominasi oleh pelanggan dengan nilai transaksi rendah, yang ditunjukkan oleh kuantitas pembelian serta harga transaksi yang relatif kecil. Cluster kedua merepresentasikan pelanggan dengan pola pembelian menengah, sebaliknya cluster ketiga menampilkan kelompok pelanggan dengan nilai transaksi besar. Pola ini mencerminkan segmentasi pelanggan bersumber pada sikap belanja, yang secara teoritis ialah salah satu aplikasi utama *clustering* dalam bidang analisis pelanggan.

Hasil ini cocok dengan teori *K- Means* yang efisien dalam mengelompokkan informasi numerik dengan distribusi relatif homogen serta berupa spherical clusters. Dataset *online retail* mempunyai ciri numerik yang jelas sehingga sesuai dengan anggapan bawah algoritma *K- Means*.



Gambar 3. Dendrogram buat Hierarchical Clustering

Gambar 3.2 memperlihatkan dendrogram hasil *Hierarchical Clustering* memakai tata cara *ward linkage*. Pemotongan dendrogram pada jarak tertentu menciptakan 3 cluster utama.

3.3 Perbandingan Kualitas Clustering Berdasarkan Validasi Internal

Pada tata cara *Hierarchical Clustering*, pendekatan agglomerative dengan *ward linkage* digunakan. Algoritma ini membangun struktur cluster secara bertahap dengan mencampurkan 2 cluster terdekat pada tiap iterasi.

Visualisasi dendrogram memperlihatkan proses penggabungan informasi secara hierarkis. Dari dendrogram, bisa diamati kalau sebagian informasi bergabung terlebih dulu membentuk sub- cluster kecil saat sebelum kesimpulannya membentuk cluster yang lebih besar. Perihal ini menampilkan terdapatnya struktur bertingkat dalam data[11].

Keunggulan utama tata cara ini terletak pada kemampuannya menunjukkan ikatan antar informasi secara visual lewat dendrogram. Periset bisa mengamati keakraban relatif antar cluster dan struktur internal informasi. Tetapi, bila dibanding dengan *K- Means*, hasil pengelompokan nampak kurang kompak. Perihal ini diakibatkan sebab *Hierarchical Clustering* tidak secara

eksplisit memaksimalkan guna objektif semacam *K- Means*, melainkan tergantung pada jarak antar pendamping informasi.

Secara teoritis, *Hierarchical Clustering* lebih sesuai buat informasi yang mempunyai struktur hierarkis yang kokoh. Tetapi, pada dataset online retail yang bertabiat transaksional serta numerik, struktur hierarkis tersebut tidak sangat dominan sehingga hasil clustering jadi kurang maksimal dibandingkan *K- Means*.

3.4 Perbandingan Kualitas Clustering Berdasarkan Validasi Internal

Nilai *Silhouette Coefficient* yang besar pada *K- Means* mencerminkan keakraban kokoh antar objek satu kelompok dan jarak pemisah yang jelas terhadap kelompok lain. Pola tersebut timbul sebab tiap informasi secara iteratif ditarik mengarah pusat representatif sampai jarak total antar titik serta centroid menggapai keadaan minimum. Proses ini memencet penyebaran internal kelompok sehingga tercipta struktur yang padat dan normal. Secara konseptual, hasil tersebut ialah konsekuensi langsung dari mekanisme optimasi kuadrat jarak Euclidean yang jadi bawah kerja *K- Means*.

Berbeda dengan pendekatan tersebut, *Hierarchical Clustering* membangun kelompok lewat penggabungan bertahap bersumber pada keakraban antar pendamping objek. Tiap sesi penggabungan tidak sediakan ruang koreksi terhadap keputusan lebih dahulu. Kepribadian permanen ini menimbulkan struktur akhir sangat dipengaruhi keadaan dini jarak informasi. Akibatnya nampak pada mutu pembelahan kelompok yang lebih lemah, sebab sebagian objek terletak pada zona transisi antar cluster tanpa representasi pusat yang jelas.

Nilai *Davies- Bouldin Index* yang lebih rendah pada *K- Means* merefleksikan rasio dispersi intra- cluster terhadap separasi antar- cluster yang lebih kecil. Secara konseptual, indeks ini mengukur seberapa jauh centroid satu cluster terhadap *centroid* lain dibandingkan tingkatan penyebaran internal cluster tersebut. Hasil DBI rendah mengindikasikan kalau tiap cluster mempunyai pusat yang silih berjauhan dan alterasi internal yang terkendali. Fenomena ini ialah konsekuensi langsung dari pendekatan *centroid- based* yang memforsir redistribusi informasi sampai tercapai konfigurasi jarak minimum *global*.

Pada *Hierarchical Clustering*, nilai DBI yang relatif lebih besar mengindikasikan terdapatnya tumpang tindih antar cluster. Perihal tersebut timbul akibat ketergantungan algoritma terhadap struktur jarak dini antar objek, bukan pada optimasi variansi global. Tata cara *ward linkage* memanglah berupaya meminimalkan kenaikan variansi, tetapi proses tersebut berlangsung secara lokal pada tiap sesi penggabungan. Ketidakhadiran proses penyesuaian ulang menimbulkan cluster akhir kurang maksimal secara geometris.

Perbandingan nilai *Calinski–Harabasz Index* menguatkan interpretasi tersebut. Nilai CHI besar pada *K- Means* menampilkan rasio alterasi antar cluster terhadap alterasi intra- cluster yang besar. Maksudnya, perbandingan antar kelompok informasi lebih dominan dibandingkan alterasi internal. Keadaan ini tidak berubah- ubah dengan anggapan spherical clusters pada *K- Means*, paling utama dikala atribut numerik sudah distandarisasi. Kebalikannya, struktur bertingkat hasil *Hierarchical Clustering* cenderung menciptakan alterasi internal yang lebih besar sebab proses penggabungan tidak memikirkan reposisi pusat cluster.

Secara totalitas, keunggulan metrik validasi internal pada *K- Means* bukan semata- mata fenomena numerik, melainkan refleksi langsung dari landasan teoritis algoritma tersebut. Optimasi berbasis centroid serta minimisasi variansi menciptakan cluster yang lebih terdefinisi secara matematis. Sedangkan itu, *Hierarchical Clustering* lebih unggul selaku perlengkapan eksplorasi struktur kedekatan informasi, tetapi kurang efisien buat menciptakan segmentasi maksimal pada dataset numerik berskala besar tanpa struktur hierarkis kokoh.

Dialog ini menyoroti berartinya menghubungkan interpretasi penemuan pengelompokan dengan tata cara algoritmik, bukan cuma menyamakan nilai metrik. Tata cara ini membolehkan penilaian mutu pengelompokan dengan metode yang valid secara teoritis serta ilmiah.

3.5 Implikasi Hasil pada Pemilihan Metode

Perbandingan performa antara *K- Means* serta *Hierarchical Clustering* bisa dipaparkan bersumber pada ciri algoritma serta watak dataset. *K- Means* memakai pendekatan berbasis *centroid* dengan tujuan meminimalkan variansi dalam cluster. Oleh sebab itu, algoritma ini cenderung menciptakan cluster yang kompak serta berbatas jelas.

Sedangkan itu, *Hierarchical Clustering* mengandalkan penggabungan bertahap bersumber pada jarak terdekat. Pendekatan ini tidak senantiasa menciptakan cluster yang maksimal secara *global* sebab keputusan penggabungan bertabiat permanen serta tidak bisa diperbaiki kembali. Perihal ini menimbulkan tata cara ini lebih sensitif terhadap *noise* serta *outlier*.

Pada dataset *online retail* yang mempunyai jumlah informasi sangat besar dan ciri numerik yang dominan, *K- Means* jadi lebih unggul sebab efisiensi komputasi serta kemampuannya membentuk cluster yang normal. Hasil ini memantapkan teori kalau tidak terdapat algoritma clustering yang senantiasa terbaik buat seluruh tipe informasi, melainkan sangat tergantung pada struktur serta ciri dataset.

3.6 Implikasi Hasil Penelitian

Hasil riset ini membagikan implikasi instan kalau buat dataset transaksi pelanggan berskala besar, tata cara *K- Means* lebih direkomendasikan sebab menciptakan cluster yang lebih kompak serta terpisah dengan jelas. Cluster yang tercipta bisa dimanfaatkan buat keperluan segmentasi pelanggan, strategi pemasaran, dan pengambilan keputusan bisnis.

Di sisi lain, *Hierarchical Clustering* senantiasa mempunyai nilai tambah dalam sesi eksplorasi informasi dini, paling utama kala periset mau menguasai struktur ikatan antar informasi tanpa wajib memastikan jumlah cluster semenjak dini.

Penemuan ini pula menegaskan berartinya pemakaian validasi internal selaku bawah objektif dalam pemilihan tata cara *clustering*. Pendekatan ini menolong periset menjauhi keputusan subjektif serta membenarkan kalau tata cara yang diseleksi betul- betul cocok dengan ciri informasi.

4. KESIMPULAN

Output uji coba mengindikasikan metode K-Means memperoleh keunggulan pengelompokkan semakin tinggi daripada Hierarchical Clustering untuk himpunan data penjualan elektronik retail sebanyak 581.586 entri. Penemuan tersebut dikonfirmasi dengan skor Silhouette Coefficient sejumlah 0,45, Davies-Bouldin Index sejumlah 1,20, dan Calinski-Harabasz Index sejumlah 3200 untuk situasi tiga kelompok. Rangkaian skor tersebut menunjukkan level kepadatan kelompok besar dan jarak antar kelompok yang semakin nyata[13].

Kekuatan K-Means berhubungan erat pada ciri himpunan dta kuantitatif berukuran luas yang dipakai. Metode yang menggunakan centroid berhasil mengurangi keragaman intra-cluster yang akibatnya susunan kelompok tercipta semakin padat serta konsisten. Di sisi lain, Hierarchical Clustering memperoleh skor evaluasi dalam jauh lebih kecil, hal ini mengindikasikan keterpisahan kelompok tidak cukup baik untuk susunan himpunan data non-hierarkis[14].

Menurut output berbasis angka tersebut, K-Means semakin tepat diterapan dalam penelaahan pengelompokkan himpunan data penjualan berukuran luas dengan memanfaatkan evaluasi dalam sebagai suatu acuan penilaian netral. Seleksi teknik pengelompokkan harus diselaraskan terhadap susunan himpunan data dan maksud studi sehingga keunggulan pengelompokan bisa terhitung dengan cara melalui data[15].

5. SARAN

Berdasarkan temuan serta keterbatasan yang teridentifikasi dalam penelitian ini, disarankan agar penelitian selanjutnya dapat mengembangkan beberapa aspek guna memperdalam analisis dan memperluas cakupan penerapan metode clustering. Penelitian mendatang dapat difokuskan pada pengujian kedua algoritma menggunakan dataset yang lebih beragam, tidak hanya dari segi ukuran, tetapi juga dari jenis dan karakteristik data, seperti data kategorikal, data deret waktu, maupun data berdimensi sangat besar. Hal ini bertujuan untuk mengkaji sejauh mana kinerja K-Means dan Hierarchical Clustering dapat digeneralisasikan pada berbagai konteks data nyata.

Selain itu, eksplorasi terhadap variasi parameter algoritma perlu dilakukan secara lebih mendalam. Sebagai contoh, penelitian lanjutan dapat menguji berbagai metode *linkage* selain *ward* pada Hierarchical Clustering atau mengevaluasi pengaruh perbedaan inisialisasi centroid pada K-Means terhadap hasil pengelompokan. Penambahan metrik validasi internal maupun eksternal yang lebih bervariasi juga dapat memberikan gambaran yang lebih komprehensif mengenai kualitas clustering yang dihasilkan.

Saran berikutnya adalah mengintegrasikan teknik pra-pemrosesan data yang lebih lanjut, seperti penanganan outlier, penerapan metode normalisasi yang sesuai dengan distribusi data, serta reduksi dimensi sebelum proses clustering. Pendekatan ini diharapkan mampu meningkatkan kualitas hasil pengelompokan, terutama pada dataset yang kompleks dan mengandung noise.

Penelitian selanjutnya juga dapat mempertimbangkan pengembangan atau penerapan pendekatan clustering berbasis *hybrid* maupun *ensemble* yang mengombinasikan keunggulan dari masing-masing algoritma. Dengan menggabungkan hasil dari beberapa metode, diharapkan dapat diperoleh cluster yang lebih stabil, robust, dan mudah diinterpretasikan.

Terakhir, apabila memungkinkan, penggunaan dataset yang memiliki label kelas dapat dimanfaatkan untuk melakukan validasi eksternal. Pendekatan ini bertujuan untuk memberikan konfirmasi tambahan terhadap hasil evaluasi berbasis validasi internal, sekaligus memperkaya analisis komparatif antara metode clustering yang diteliti.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Esa Unggul yang sudah membagikan sokongan sarana serta sumber energi sepanjang penerapan riset ini. Perkataan terima kasih pula di informasikan kepada seluruh pihak yang sudah berkontribusi secara langsung ataupun tidak langsung dalam penyelesaian riset ini.

DAFTAR PUSTAKA

[1] L. Trapanese, G. Bifulco, M. Santinello, N. Pasquino, G. Campanile, and A. Salzano, “Comparison of K-Means and Hierarchical Clustering Methods for Buffalo Milk Production Data,” *Animals*, vol. 15, no. 22, Nov. 2025, doi: 10.3390/ani15223246.

[2] Srilekha S and Adhilakshmi, “Comparative Evaluation of K-Means, Hierarchical Clustering, and DBSCAN in Blood Donor Segmentation,” 2024. [Online]. Available: www.ijfmr.com

[3] F. Rusman and V. Tundjungsari, “IMPLEMENTASI ALGORTIMA CLUSTERING PADA DATA PENJUALAN MOBIL DI AUTI2000 SALEMBA,” 2025.

[4] A. Rima Hariyanto, B. Salsabiilah, S. Khodijah, W. Septi Aulia, and M. Fansyuri, “Klasterisasi Mahasiswa Berdasarkan Performa Akademik Menggunakan Algoritma K-

Means pada RapidMiner: Studi Kasus dengan Dataset Student Academic Performance," 2025.

[5] S. Wijayanti and A. Kesumawati, "Perbandingan Analisis K-means dan Hierarchical Clustering dalam Mengelompokkan Kecamatan di Kabupaten Grobogan Berdasarkan Jumlah Titik Kejadian Bencana Alam," *Emerging Statistics and Data Science Journal*, vol. 3, no. 2, 2025.

[6] A. Rahmah, N. Faoziatun Khusna, S. A. Sanmas, S. Aulia, S. Amaria, and F. Fauzi, "Comparison Analysis of Hierarchical Clustering and K-Means Methods in Grouping Provinces in Indonesia Based on Dengue Hemorrhagic Fever (DHF) Cases," 2025.

[7] W. I. Permatasari and V. Tundjungsari, "Implementasi Algoritma K-Means untuk Mengetahui Minat Siswa SMA Terhadap Mata Pelajaran Teknologi Informasi dan Komunikasi," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 168–174, Dec. 2024, doi: 10.57152/malcom.v5i1.1639.

[8] R. Mubarok, A. Angga Syahputra, A. Teguh Permana, L. Sholiah, and P. Studi Sistem Informasi, "Implementasi Data Mining untuk Clustering Lowongan Pekerjaan Menggunakan Metode Algoritma K-Means," 2025.

[9] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "DATA MINING CLUSTERING FOOD EXPENDITURE IN INDONESIA," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 6, pp. 1491–1500, Dec. 2022, doi: 10.20884/1.jutif.2022.3.6.331.

[10] B. Hartono, V. Lusiana, and H. Al Amin, "Perbandingan Proses Klasterisasi Data Menggunakan K-Means Clustering dan Agglomerative Hierarchical Clustering," *Jurnal Riset Komputer)*, vol. 12, no. 4, pp. 2407–389, 2025, doi: 10.30865/jurikom.v12i4.8766.

[11] A. Gupta, H. Sharma, and A. Akhtar, "A COMPARATIVE ANALYSIS OF K-MEANS AND HIERARCHICAL CLUSTERING," *EPRA International Journal of Multidisciplinary Research (IJMR)-Peer Reviewed Journal*, no. 8, 2021, doi: 10.36713/epra2013.

[12] M. Dani Yusuf, T. Ai Munandar, K. Fadhilla Ramdhania, and C. Autor, "Comparative Analysis of K-Means and Hierarchical Clustering for Regional Welfare Disparity Identification in West Java Province," *International Journal of Information Technology and Computer Science Applications*, 2025.

[13] A. Azzahra and A. W. Wijayanto, "Perbandingan Agglomerative Hierarchical dan K-Means dalam Pengelompokkan Provinsi Berdasarkan Pelayanan Kesehatan Maternal," 2022.

[14] A. R. R. Alzahrani, "Impact of Dataset Scaling on Hierarchical Clustering: A Comparative Analysis of Distance Based and Ratio-Based Methods," *International Journal of Analysis and Applications*, vol. 22, 2024, doi: 10.28924/2291-8639-22-2024-36.

[15] J. Novaldi and A. W. Wijayanto, "Analisis Cluster Kualitas Pemuda di Indonesia pada Tahun 2022 dengan Agglomerative Hierarchical dan K-Means," *Komputika : Jurnal Sistem Komputer*, vol. 12, no. 2, pp. 91–99, Sep. 2023, doi: 10.34010/komputika.v12i2.10348.