IMPLEMENTASI METODE *CLUSTERING* HDBSCAN DALAM SEGMENTASI KONSUMEN *E-COMMERCE* BERDASARKAN PERILAKU BELANJA

SKRIPSI

OLEH
ROSSIMA EVA YULIANA
NIM. 220601110068

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

IMPLEMENTASI METODE *CLUSTERING* HDBSCAN DALAM SEGMENTASI KONSUMEN *E-COMMERCE* BERDASARKAN PERILAKU BELANJA

SKRIPSI

Diajukan Kepada
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
untuk Memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Matematika (S.Mat)

Oleh
Rossima Eva Yuliana
NIM. 220601110068

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026

IMPLEMENTASI METODE *CLUSTERING* HDBSCAN DALAM SEGMENTASI KONSUMEN *E-COMMERCE* BERDASARKAN PERILAKU BELANJA

SKRIPSI

Oleh
Rossima Eva Yuliana
NIM. 220601110068

Telah Disetujui untuk Diuji

Malang, 29 April 2026

Dosen Pembimbing I	Dosen Pembimbing II
Hisyam Fahmi, M.Kom	Mohammad Nafie Jauhari, M.Si
NIP. 19890727 201903 1 018	NIPPPK. 19870218 202321 1 018

Mengetahui,
Ketua Program Studi Matematika

Dr. Fachrur Rozi, M.Si.
NIP. 19800527 200801 1 012

IMPLEMENTASI METODE *CLUSTERING* HDBSCAN DALAM SEGMENTASI KONSUMEN *E-COMMERCE* BERDASARKAN PERILAKU BELANJA

SKRIPSI

Oleh
Rossima Eva Yuliana
NIM. 220601110068

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima sebagai Salah Satu Persyaratan
untuk Memperoleh Gelar Sarjana Matematika (S.Mat)

Tanggal 4 Juni 2026

Ketua Penguji : Prof. Dr. H. Turmudi, M.Si, Ph.D .....

Anggota Penguji 1 : Ria Dhea Layla Nur Karisma, M.Si .....

Anggota Penguji 2 : Hisyam Fahmi, M.Kom .....

Anggota Penguji 3 : Mohammad Nafie Jauhari, M.Si .....

Mengetahui,
Ketua Program Studi Matematika

Dr. Fachrur Rozi, M.Si.
NIP. 19800527 200801 1 012

PERNYATAAN KEASLIAN TULISAN

Saya bertanda tangan dibawah ini

Nama : Rossima Eva Yuliana

NIM : 220601110068

Program Studi : Matematika

Fakultas : Sains dan Teknologi

Judul Skripsi : Implementasi Metode *Clustering* HDBSCAN dalam Segmentasi Konsumen E-*commerce* Berdasarkan Perilaku Belanja

Menyatakan bahwa skripsi yang saya susun merupakan hasil karya sendiri dan tidak mengambil atau menyalin tulisan maupun pemikiran orang lain. Seluruh isi skripsi ini adalah hasil pemikiran saya, kecuali bagian yang telah disebutkan sumbernya dalam daftar rujukan pada halaman terakhir. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil plagiasi atau meniru karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 4 Juni 2026

Rossima Eva Yuliana
NIM.220601110068

MOTO

"Allah tidak mengubah nasib suatu kaum, hingga mereka mengubah diri mereka sendiri"
(Q.S Ar-Ra'd: 11)

PERSEMBAHAN

Dengan segenap rasa syukur kepada Allah SWT, karya ini penulis persembahkan kepada keluarga tercinta yang selalu menjadi sumber doa, semangat dan kekuatan.

Untuk bapaku tercinta, terima kasih atas setiap nasihat, doa, dan keyakinan yang selalu diberikan. Di saat penulis merasa ragu, bapak selalu mengingatkan bahwa penulis mampu melewati setiap ujian yang ada. Terima kasih karena selalu mengajarkan penulis untuk tetap berjuang, dan tidak mudah menyerah.

Untuk mamah tercinta, terima kasih atas kasih sayang, dukungan dan semangat yang tidak pernah putus. Doa dan dukungan mamah menjadi salah satu alasan penulis mampu bertahan dan menyelesaikan perjalanan ini.

Untuk adik perempuan penulis, terima kasih telah menjadi bagian dari perjalanan hidup penulis. Meskipun terkadangan menyebalkan, kehadiranmu selalu dirindukan dan menjadi warna tersendiri dalam hidup penulis.

Semoga karya ini menjadi bentuk kecil dari rasa terima kasih penulis kepada keluarga yang selalu mendukung, mendokan dan percaya kepada penulis.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Segala puji dan syukur penulis panjatkan ke hadirat Allah Swt. atas limpahan rahmat, karunia, serta hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Implementasi Metode *Clustering* HDBSCAN dalam Segmentasi Konsumen *E-Commerce* Berdasarkan Perilaku Belanja". Shalawat serta salam semoga senantiasa tercurah kepada junjungan Nabi Muhammad SAW, yang telah membimbing umat manusia dari zaman kegelapan menuju jalan kebenaran, yaitu Islam.

Skripsi ini disusun untuk memenuhi salah satu syarat dalam memperoleh gelar Sarjana Matematika pada Program Studi Matematika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pada kesempatan ini, dengan segala kerendahan hati penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Agus Mulyono, M.Kes, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrur Rozi, M.Si, selaku Ketua Program Studi Matematika, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Hisyam Fahmi, M.Kom, selaku Dosen Pembimbing I yang telah memberikan berbagai bimbingan, ilmu pengetahuan, arahan serta motovasi kepada penulis.
5. Mohammad Nafie Jauhari, M.Si, selaku Dosen Pembimbing II yang telah memberikan berbagai bimbingan, ilmu pengetahuan, arahan serta motivasi kepada penulis.
6. Prof. Dr. H. Turmudi, M.Si, Ph.D, selaku Ketua Penguji dalam Ujian Skripsi, yang telah berkenan memberikan arahan serta saran yang membangun bagi penyusun skripsi.
7. Ria Dhea Layla Nur Karisma, M.Si, selaku Penguji I dalam ujian Skripsi, yang telah memberikan arahan dan saran yang sangat mendukung dalam penyusunan skripsi ini.

8. Seluruh Dosen Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
9. Bapak tercinta sosok panutan yang telah menjadi sumber inspirasi, doa, serta semangat dalam setiap langkah perjalanan penulis, terima kasih segala pengorbanan, nasihat dan cinta tanpa batas yang menjadi kekuatan terbesar dalam meraih impian ini. Mamah tersayang terima kasih atas sayang, doa dan ketulusan yang tak pernah berhenti mengiringi setiap langkahku. Adiku yang kubanggakan, Rossima Kayla Alvina Ramadhani, terima kasih atas canda, semangat dan dukungan yang selalu membuat hari- hari penulis penuh warna.
10. Teman-teman kamar 25 Ummu Salamah (USA) serta pendamping kamar, terima kasih atas dukungan, kebersamaan, dan keceriaan yang telah dibagikan selama satu tahun di mahad. Kehadiran kalian memberikan semagat dan suasana positif yang membantu penulis dalam menjalani kegiatan sehari-hari.
11. Teman-teman Angakatan 2022 Program Studi Matematika yang selalu memberikan bantuan, dukungan dan kerja sama selama perkuliahan, penulis menyampaikan rasa terima kasih sebesar-besarnya.
12. Teman-teman seperjuangan, Amira Adelia Putri, Ayu Habibatun Wahyu Zainuri, Do'aul Isma Mufidah, Mutiara Alfi dan Nurus Syafi'ah, yang telah membantu, mendukung serta canda tawa yang sudah menemani penulis selama perkuliahan.
13. Teman-teman konsorsium Komputasi, Diah Mariatul Ulya, Farahnas Imaniyah Pranata, Fisal Ilham Ramadhani, Nicikya Bintang, Nurus Syafi'ah, Rojauna Zalfatthoriq, dan Yusuf Firmansyah, penulis menyampaikan terima kasih atas kebersamaan, dukungan, dan semangat yang telah diberikan selama perkuliahan konsorsium.
14. Kepada seseorang yang penulis tidak bisa menyebutkan namanya, yang senantiasa mendampingi, memberikan bantuan, serta mendukung penulis selama perkuliahan maupun berbagai kegiatan organisasi, penulis menyampaikan terima kasih yang sebesar-besarnya. Kehadiran dan dukungan tersebut memberikan semangat, serta motivasi yang turut

membantu penulis menyelesaikan proses perkuliahan hingga tahap skripsi ini.

15. Teruntuk diri sendiri, penulis mengucapkan terima kasih atas ketekunan, kesabaran, serta perjuangan yang telah dilalui hingga mencapai titik ini. Terima kasih atas usaha dan kerja keras yang tidak pernah sia-sia, serta maaf atas saat-saat penulis masih merasa ragu atau menyalahkan diri sendiri. Semoga pengalaman ini menjadi motivasi untuk terus berkembang di masa depan.

Penulis menyadari bahwa penyusunan skripsi ini masih jauh dari sempurna. Dengan penuh ketulusan, penulis mempersembahkan karya ini sebagai ungkapan terima kasih dan penghargaan kepada semua pihak yang telah memberikan dukungan, bimbingan, serta motivasi. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan menjadi landasan untuk penelitian yang lebih baik di masa depan.

Malang, 4 Juni 2026

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

ABSTRAK

Yuliana, Rossima Eva 2026. **Implementasi Metode *Clustering* HDBSCAN dalam Segmentasi Konsumen *E-commerce* Beradasarkan Perilaku Belanja**. Skripsi. Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
Pembimbing: (I) Hisyam Fahmi, M.Kom. (II) Mohammad Nafie Jauhari, M.Si

Kata Kunci: *Clustering*, *E-commerce*, HDBSCAN, Perilaku Belanja, Segmentasi Konsumen.

Clustering merupakan salah satu teknik data mining yang digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik yang dimiliki. Pada *e-commerce*, *clustering* dapat digunakan untuk melakukan segmentasi konsumen berdasarkan perilaku belanja sehingga membantu perusahaan dalam memahami karakteristik konsumen. Salah satu metode *clustering* yang dapat digunakan adalah *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) karena mampu mengelompokkan data dengan tingkat kepadatan yang berbeda serta mengidentifikasi data yang tidak termasuk ke dalam klaster sebagai *noise*. Penelitian ini bertujuan untuk mengimplementasikan metode HDBSCAN dalam segmentasi konsumen *e-commerce* berdasarkan perilaku belanja. Data yang digunakan merupakan data sekunder dari dataset *E-commerce Customer Behavior and Sales Analysis-TR* yang diperoleh dari Kaggle dengan jumlah 5.000 data transaksi konsumen *e-commerce* di Turki pada periode Januari 2023 hingga Maret 2024. Tahapan penelitian meliputi normalisasi data menggunakan *Min-Max Scaler*, perhitungan jarak *Euclidean*, penerapan algoritma HDBSCAN, serta evaluasi hasil klaster menggunakan *Silhouette Coefficient* dan *Davies Bouldin Index*. Hasil penelitian menunjukkan bahwa parameter terbaik diperoleh pada nilai min_samples 55 dan min_cluster_size 55 yang menghasilkan 6 klaster. Nilai *Silhouette Coefficien*t yang diperoleh sebesar 0,2155 dan *Davies Bouldin Index* sebesar 1,6047. Berdasarkan karakteristik masing-masing klaster, konsumen dapat dikelompokkan menjadi Pelanggan Terbaik, Pelanggan Berisiko, Pelanggan Potensial, Pelanggan Loyal, Pelanggan Sensitif Promosi, dan Pelanggan Hampir Hilang. Dengan demikian, metode HDBSCAN mampu melakukan segmentasi konsumen *e-commerce* berdasarkan kemiripan perilaku belanja serta mengidentifikasi data yang tidak tergabung dalam klaster sebagai *noise*.

ABSTRACT

Yuliana, Rossima Eva 2026. **Implementation of the HDBSCAN Clustering Method in E-commerce Consumer Segmentation Based on Shopping Behaviour**. Undergraduate thesis. Mathematics Study Program, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
Advisors: (I) Hisyam Fahmi, M.Kom. (II) Mohammad Nafie Jauhari, M.Si

Keywords: Clustering, E-commerce, HDBSCAN, Shopping Behaviour, Consumer Segmentation.

Clustering is a data mining technique used to group data based on the degree of similarity in their characteristics. In e-commerce, clustering can be used to segment consumers based on shopping behavior, thereby helping companies understand consumer characteristics. One clustering method that can be used is Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) because it is capable of grouping data with different density levels and identifying data that does not belong to a cluster as noise. This study aims to implement the HDBSCAN method in e-commerce consumer segmentation based on shopping behavior. The data used is secondary data from the E-commerce Customer Behavior and Sales Analysis-TR dataset obtained from Kaggle, consisting of 5.000 e-commerce consumer transaction records in Turkey from January 2023 to March 2024. The research stages include data normalization using the Min-Max Scaler, calculation of Euclidean distance, application of the HDBSCAN algorithm, and evaluation of clustering results using the Silhouette Coefficient and Davies-Bouldin Index. The results show that the optimal parameters were obtained at min_samples = 55 and min_cluster_size = 55, yielding six clusters. The Silhouette Coefficient value obtained was 0,2155, and the Davies-Bouldin Index was 1,6047. Based on the characteristics of each cluster, consumers can be grouped into Top Customers, At-Risk Customers, Potential Customers, Loyal Customers, Promotion-Sensitive Customers, and Churn-Prone Customers. Thus, the HDBSCAN method is capable of segmenting e-commerce consumers based on similarities in shopping behavior and identifying data that does not belong to any cluster as noise.

مستخلص البحث

يوليانا، روسيما إيفا. ٢٠٢٦ .**تطبيق طريقة التجميع *HDBSCAN* في تصنيف مستهلكي التجارة الإلكترونية بناءً على سلوك الشراء**. الجت العلميقسم الرياضيات، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية في مالانج. المشرفان: (١) هشام فهمي، الماجستير فى علم الحاسوب. (٢) محمد نافع جوهري الماجستير فى العلوم.

الكلمات الأساسيه: التجميع، التجارة الإلكترونية, *HDBSCAN*, سلوك التسوق، تقسيم المستهلكين.

التجميع هو إحدى تقنيات استخراج البيانات التي تُستخدم لتصنيف البيانات بناءً على درجة تشابه خصائصها. في التجارة الإلكترونية، يمكن استخدام التجميع لإجراء تقسيم المستهلكين بناءً على سلوك التسوق، مما يساعد الشركات في فهم خصائص المستهلكين. إحدى طرق التجميع التي يمكن استخدامها هي التجميع المكاني الهرمي القائم على الكثافة للتطبيقات مع الضوضاء (*HDBSCAN*) لأنها هدفت على تجميع البيانات بمستويات كثافة مختلفة وكذلك تحديد البيانات التي لا تندرج ضمن المجموعات على أنها ضوضاء. هدفت هذه الدراسة إلى تطبيق طريقة *HDBSCAN* في تقسيم المستهلكين في التجارة الإلكترونية بناءً على سلوكهم الشرائي. البيانات المستخدمة هي بيانات ثانوية من مجموعة بيانات *E-commerce Customer Behavior and Sales Analysis-TR* التي تم الحصول عليها من *Kaggle* ، وتضم ٥.٠٠٠ معاملة للمستهلكين في التجارة الإلكترونية في تركيا خلال الفترة من يناير ٢٠٢٣ إلى مارس ٢٠٢٤. تشمل مراحل البحث تطبيع البيانات باستخدام *Min-Max* Scaler ، وحساب المسافة الأوقليدية، وتطبيق خوارزمية *HDBSCAN* ، وتقييم نتائج المجموعات باستخدام معامل *Silhouette Coefficient* ومؤشر *Davies Bouldin Index*. أظهرت نتائج البحث أن أفضل المعلمات تم الحصول عليها عند قيمة ٥٥ *min_samples* و ٥٥ *min_cluster_size* ، مما أدى إلى تكوين ٦ مجموعات. بلغت قيمة معامل ٠.٢١٥٥ *Silhouette Coefficient* ومؤشر ١.٦٠٤٧ *Davies Bouldin Index*. وبناءً على خصائص كل مجموعة، يمكن تصنيف المستهلكين إلى أفضل العملاء، والعملاء المعرضين للمخاطر، والعملاء المحتملين، والعملاء المخلصين، والعملاء الحساسين للعروض الترويجية، والعملاء الذين على وشك الانسحاب. وبالتالي، فإن طريقة *HDBSCAN* قادرة على إجراء تقسيم المستهلكين في التجارة الإلكترونية بناءً على تشابه سلوك الشراء، وكذلك تحديد البيانات غير المضمنة في المجموعات على أنها ضوضاء.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan industri *e-commerce* yang semakin pesat telah mendorong perubahan signifikan dalam perilaku belanja masyarakat di era digital (Amory dkk., 2025). Platform *e-commerce* tidak hanya menjadi sarana transaksi, tetapi juga membentuk pola konsumsi baru yang lebih dinamis dan berorientasi pada kemudahan serta kecepatan layanan. Di tengah meningkatnya volume transaksi dan keragaman konsumen, perusahaan menghadapi tantangan dalam memahami variasi perilaku belanja konsumen yang semakin kompleks (Larasati Mayang, 2024). Oleh karena itu, dibutuhkan metode analisis yang mampu mengelompokkan konsumen berdasarkan pola pembelian yang serupa secara lebih tepat. Segmentasi berbasis data transaksi menjadi salah satu metode yang efektif untuk memetakan karakteristik konsumen secara objektif dan sistematis. Dalam konteks ini, penerapan metode *clustering Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) memberikan peluang untuk menghasilkan segmentasi yang yang lebih fleksibel, terutama untuk data yang beragam dan mengandung *noise*, sehingga memberikan hasil perilaku belanja konsumen *e-commerce* yang lebih akurat.

Segmentasi konsumen memiliki keterkaitan erat dengan metode *clustering* karena keduanya bertujuan mengelompokkan individu berdasarkan kesamaan karakteristik tertentu (Febrianty dkk., 2023). Dalam konteks *e-commerce*, perilaku belanja menjadi indikator utama yang dapat digunakan untuk membedakan pola

tindakan konsumen, seperti frekuensi pembelian, jenis produk yang dipilih, maupun besaran transaksi. Melalui teknik *clustering*, data perilaku belanja yang kompleks dapat diolah secara matematis untuk membentuk kelompok konsumen yang memiliki kemiripan pola konsumsi (Rusvinasari & Annisa, 2025). Hasil segmentasi tersebut memungkinkan perusahaan memahami variasi kebutuhan pelanggan secara lebih detail, sehingga strategi pemasaran dapat disesuaikan dengan karakteristik tiap segmen. Dengan demikian, segmentasi berbasis perilaku belanja melalui *clustering* menjadi pendekatan yang efektif untuk menghasilkan keputusan bisnis yang lebih tepat sasaran dan berbasis data.

Metode *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) merupakan pengembangan dari algoritma *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) yang dirancang untuk mengatasi keterbatasan dalam menentukan parameter kepadatan yang berupa nilai radius (ε) dan minimum jumlah titik (*Minimum Points*) atau disingkat $MinPts$ yang bersifat tetap (Purnomo, 2024). Pendekatan ini bekerja dengan membangun struktur hierarki klaster berdasarkan variasi kepadatan data, kemudian mengekstraksi klaster yang paling stabil dari struktur tersebut (Ghiffary dkk., 2024). Keunggulan utama HDBSCAN terletak pada kemampuannya, untuk mengelompokkan data meskipun kepadatan titik data berbeda-beda di setiap wilayah. Artinya, HDBSCAN tetap bisa membentuk klatser yang baik walaupun ada area data yang sangat padat dan area yang lebih jarang, serta mendeteksi data yang bersifat *outlier* secara otomatis (Perdana & Santoso, 2025). Dalam konteks analisis konsumen *e-commerce*, metode ini memberikan fleksibilitas lebih tinggi dibandingkan algoritma *clustering* konvensional seperti K-Means atau DBSCAN

(Rizki dkk., 2023). HDBSCAN tidak memerlukan penentuan jumlah klaster di awal dan mampu menyesuaikan bentuk klaster sesuai distribusi alami data. Hal ini menjadikannya relevan untuk menganalisis perilaku konsumen yang bervariasi dan tidak mudah diprediksi, di mana data sering kali menunjukkan variasi pola pembelian, preferensi produk, dan frekuensi transaksi yang tidak homogen. Dengan demikian, penggunaan HDBSCAN memungkinkan peneliti untuk memperoleh segmentasi konsumen yang lebih jelas dan sesuai dengan kebutuhan bisnis.

Al-Qur'an juga mengajarkan pentingnya ihsan (berbuat baik) dan keseimbangan antara urusan dunia dan akhirat dalam aktivitas ekonomi, sebagaimana termaktub dalam surah Al-Qashash ayat 77:

وَابْتَغِ فِيْمَآ اٰتٰىكَ اللّٰهُ الدَّارَ الْاٰخِرَةَ وَلَا تَنْسَ نَصِيْبَكَ مِنَ الدُّنْيَا وَاَحْسِنْ كَمَآ اَحْسَنَ اللّٰهُ اِلَيْكَ وَلَا تَبْغِ الْفَسَادَ فِى الْاَرْضِۗ اِنَّ اللّٰهَ لَا يُحِبُّ الْمُفْسِدِيْنَ ٧٧

Terjemahan:
"Dan carilah pada apa yang telah dianugerahkan Allah kepadamu kebahagiaan negeri akhirat, dan janganlah kamu melupakan bagianmu dari (kenikmatan) duniawi; dan berbuat baiklah (kepada orang lain) sebagaimana Allah telah berbuat baik kepadamu, dan janganlah kamu berbuat kerusakan di muka bumi. Sesungguhnya Allah tidak menyukai orang-orang yang berbuat kerusakan." (QS. Al-Qashash [28]: 77)(Kementrian Agama, 2022)

Ayat tersebut mengajarkan bahwa manusia tidak boleh semata-mata mengejar keuntungan duniawi, melainkan juga harus menebarkan ihsan dalam kehidupan sosial. Dalam konteks *e-commerce*, penerapan segmentasi konsumen dan personalisasi berbasis prinsip keadilan dapat dipandang sebagai bentuk ihsan modern, di mana data konsumen dimanfaatkan untuk memberikan pengalaman belanja yang relevan, efisien, dan memuaskan tanpa melanggar privasi. Dengan demikian, penelitian ini tidak hanya meningkatkan efisiensi bisnis, tetapi juga mendukung pembentukan etika digital yang sejalan dengan nilai-nilai Islam, yakni memberikan manfaat tanpa menimbulkan mudarat.

Data transaksi *e-commerce* umumnya memiliki karakteristik heterogen, non-linear, dan mengandung *noise*, misalnya variasi besar dalam jumlah pembelian, frekuensi transaksi, atau total pengeluaran antar konsumen (Apriyanto & Sitio, 2025). Kondisi ini membuat metode klaster klasik seperti K-Means kurang efektif, karena metode tersebut mengasumsikan bahwa setiap klaster memiliki bentuk yang seragam, padahal data perilaku konsumen cenderung bervariasi. (Mutiah dkk., 2024). HDBSCAN lebih sesuai digunakan karena mampu menangani data yang menyimpang atau tahan terhadap *outlier* tanpa mengganggu hasil klaster, tidak memerlukan penentuan jumlah klaster di awal, serta mampu menemukan struktur alami dalam data dengan tingkat kepadatan yang berbeda. Dengan demikian, hasil segmentasi akan lebih representatif terhadap variasi nyata dalam perilaku konsumen digital.

Sebagian besar penelitian sebelumnya dalam bidang segmentasi konsumen pada *e-commerce* masih cenderung menerapkan metode *clustering* konvensional, seperti K-Means maupun DBSCAN (Alwi dkk., 2025). Namun, kedua metode tersebut memiliki keterbatasan dalam menangani data dengan tingkat kepadatan yang bervariasi serta dalam situasi di mana jumlah klaster belum diketahui secara pasti (Lestari, 2023). Metode K-Means cenderung tidak adaptif terhadap bentuk klaster yang tidak linear dan mengharuskan penentuan jumlah klaster di awal, sedangkan DBSCAN masih bergantung pada parameter *epsilon* yang sensitif terhadap perubahan skala data.

Penelitian-penelitian terdahulu menunjukkan bahwa metode HDBSCAN memiliki keunggulan signifikan dibandingkan DBSCAN, terutama dalam menangani data dengan variasi kepadatan dimana dalam satu dataset terdapat

kelompok data yang memiliki tingkat kepadatan berbeda, yaitu ada *cluster* yang sangat rapat dan ada yang lebih renggang. Metode DBSCAN sulit menangani kondisi ini karena menggunakan parameter tetap, sedangkan HDBSCAN mampu menyesuaikan dengan perbedaan kepadatan tersebut dan keberadaan *noise* yang tinggi. Bushra dkk., (2024) melalui studi berjudul AutoSCAN merupakan pendekatan yang bertujuan untuk menentukan parameter DBSCAN, yaitu (ε) dan minPts secara otomatis, sehingga mengurangi ketergantungan pada penentuan manual yang dapat menyebabkan hasil *cluster* tidak optimal, serta menjelaskan bahwa DBSCAN memiliki keterbatasan utama pada sensitivitas parameter *epsilon* dan minPts (*Minimum Points*), yang sering menghasilkan klaster tidak stabil, sedangkan HDBSCAN mampu mengatasi permasalahan ini dengan pendekatan hierarkis dan estimasi kepadatan adaptif. John dkk., (2023) dalam penelitian segmentasi pelanggan di pasar ritel Inggris juga menemukan bahwa HDBSCAN memberikan hasil klaster yang lebih representatif dibanding metode lain seperti K-Means dan DBSCAN, terutama ketika digunakan untuk data perilaku konsumen yang kompleks.

Temuan serupa diperkuat oleh studi yang dipublikasikan oleh (Chandrasekaran dkk., 2023) tentang segmentasi konsumen menggunakan kombinasi *Uniform Manifold Approximation and Projection* (UMAP), yaitu Teknik *dimensionality reduction* (reduksi dimensi) yang digunakan untuk memproyeksikan data berdimensi tinggi ke dalam ruang berdimensi lebih rendah (biasanya 2D atau 3D) dan HDBSCAN, di mana metode ini menghasilkan visualisasi klaster yang lebih jelas dan stabil dibandingkan DBSCAN. Selain itu (Alemán dkk., 2022) melalui penelitian pada analisis dinamika molekul

memperlihatkan bahwa HDBSCAN unggul dalam mengelompokkan data berdimensi tinggi dengan tingkat *noise* tinggi tanpa memerlukan penentuan parameter secara manual. Bahkan, penelitian, (Hunt & Reffert, 2021) dalam bidang astronomi menunjukkan bahwa metode HDBSCAN mampu mendeteksi pola dan struktur kompleks. Kemampuan ini diukur melalui keberhasilan metode dalam mengidentifikasi *cluster* dengan bentuk yang tidak beraturan serta memisahkan data berdasarkan variasi kepadatan, yang dievaluasi menggunakan metrik kualitas *cluster* seperti *Silhouette Coefficient* dan *Davies Bouldin Index*. Hasil evaluasi tersebut menunjukkan bahwa HDBSCAN menghasilkan nilai *Silhouette Coefficient* yang lebih tinggi dan nilai *Davies Bouldin Index* (DBI) yang lebih rendah, dibandingkan metode lain, sehingga menunjukkan kualitas *cluster* yang lebih baik. Selain itu, fleksibilitas HDBSCAN dibuktikan melalui kemampuannya dalam menangani berbagai karakteristik data tanpa bergantung pada satu parameter tetap seperti *epsilon*, melainkan menggunakan pendekatan berbasis kepadatan dan stabilitas *cluster*. Dengan demikian, metode ini dapat diterapkan secara efektif pada berbagai domain data, seperti astronomi, kesehatan, keuangan mapupun *e-commerce*. Berdasarkan hasil-hasil tersebut, dapat disimpulkan bahwa metode HDBSCAN menunjukkan performa yang lebih baik dibandingkan DBSCAN, ditinjau dari kemampuannya dalam menghasilkan *cluster* yang lebih sesuai dengan struktur data, mampu menyesuaikan diri terhadap variasi kepadatan data, serta menghasilkan *cluster* yang lebih konsisten. Dengan demikian, HDBSCAN berpotensi besar untuk diimplementasikan dalam konteks segmentasi konsumen *e-commerce*, terutama pada data yang bervariasi dan tidak sama antar konsumen.

Urgensi penelitian ini terletak pada kebutuhan industri *e-commerce* untuk memahami variasi perilaku belanja konsumen secara lebih mendalam melalui pendekatan analisis data yang akurat. Dalam lingkungan digital yang penuh saingan, perilaku belanja menjadi indikator penting untuk mengidentifikasi preferensi, pola pembelian, dan kecenderungan konsumen yang berbeda antarsegmen (Wardhana, 2024). Oleh karena itu, penerapan metode *clustering* HDBSCAN menjadi relevan karena mampu mengelompokkan konsumen berdasarkan pola belanja yang kompleks, heterogen, dan sering kali mengandung *noise*. Penelitian ini tidak hanya memberikan kontribusi pada aspek teknis melalui implementasi algoritma *clustering* yang lebih adaptif, tetapi juga menawarkan manfaat manajerial bagi perusahaan dalam merancang strategi pemasaran dan pelayanan yang lebih tepat sasaran. Dengan pemahaman segmen konsumen yang lebih akurat, perusahaan dapat meningkatkan efektivitas pengambilan keputusan dan memperkuat daya saing di pasar *e-commerce*.

Kebaruan dalam penelitian ini terletak pada penerapan metode HDBSCAN dengan evaluasi kinerja menggunakan *Davies Bouldin Index* (DBI). Penggunaan DBI sebagai alat evaluasi memberikan pendekatan objektif untuk menilai kualitas hasil segmentasi, karena DBI mengukur rasio antara jarak intra-klaster dan inter klaster. Nilai DBI yang lebih rendah menunjukkan segmentasi yang lebih baik, di mana anggota dalam satu klaster memiliki kemiripan tinggi dan berbeda signifikan dengan klaster lain (Syah dkk., 2025). Pendekatan ini memperkuat validitas hasil *clustering* dan menjadikan penelitian ini lebih komprehensif dibandingkan penelitian sejenis yang hanya berhenti pada tahap pembentukan klaster tanpa evaluasi kuantitatif.

Penelitian ini bertujuan untuk mengimplementasikan metode *clustering* HDBSCAN dalam proses segmentasi konsumen *e-commerce* berdasarkan perilaku belanja, sehingga mampu mengidentifikasi kelompok konsumen dengan pola pembelian yang serupa secara lebih akurat. Melalui segmentasi yang dihasilkan, penelitian ini diharapkan dapat mengungkap variasi perilaku belanja konsumen dan memberikan pemahaman yang lebih mendalam mengenai karakteristik setiap kelompok konsumen (Mukhlis Juliyanto dkk., 2024). Secara akademis, penelitian ini diharapkan memberikan kontribusi terhadap pengembangan ilmu data mining dan *machine learning*, khususnya dalam penerapan metode *clustering* berbasis kepadatan pada data transaksi konsumen yang bersifat heterogen. Secara praktis, hasil penelitian ini diharapkan menjadi dasar bagi perusahaan *e-commerce* dalam merancang strategi pemasaran, pelayanan, dan pengelolaan pelanggan yang lebih tepat sasaran, sehingga dapat meningkatkan efektivitas pengambilan keputusan dan mendukung keberlanjutan daya saing di pasar digital yang kompetitif.

1.2 Rumusan Masalah

Berdasarkan penjelasan pada bagian latar belakang sebelumnya, rumusan permasalahan dalam penelitian ini adalah bagaimana implementasi metode *clustering* HDBSCAN untuk melakukan segmentasi konsumen *e-commerce* berdasarkan perilaku belanja?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan dari penelitian ini adalah untuk mengetahui implementasi metode *clustering* HDBSCAN dalam proses segmentasi konsumen *e-commerce* berdasarkan perilaku belanja.

1.4 Manfaat Penelitian

Manfaat yang akan diperoleh dalam penelitian ini sebagai berikut:

1. Manfaat Teoritis

   Penelitian ini diharapkan dapat meningkatkan pemahaman dan pengetahuan tentang penerapan metode *clustering* berbasis kepadatan, khususnya HDBSCAN, dalam analisis segmentasi konsumen *e-commerce* serta hasil penelitian ini dapat dijadikan sebagai tambahan referensi bagi peneliti selanjutnya.

2. Manfaat Praktis

   Secara praktis hasil penelitian ini diharapkan mampu memberikan berbagai manfaat, antara lain:

   a. Bagi Penulis, penelitian ini juga menjadi sarana bagi penulis untuk memperdalam pemahaman dan keterampilan dalam penerapan metode *clustering* berbasis kepadatan, khususnya HDBSCAN, pada data *e-commerce*. Hasilnya diharapkan dapat memperluan kompetensi akademik dan profesional penulis dalam bidang data *science* serta analisis perilaku konsumen.

   b. Bagi Instansi (Universitan Islam Negeri Maulana Malik Ibrahim Malang), penelitian ini diharapkan dapat memberikan kontribusi akademik bagi UIN Maulana Malik Ibrahim Malang, khususnya dalam pengembangan kajian di bidang ekonomi, teknologi informasi, dan data mining. Selain itu, hasil penelitian ini dapat menjadi referensi tambahan bagi mahasiswa maupun peneliti lain dalam mengkaji penerapan metode HDBSCAN.

c. Bagi Pembaca, hasil penelitian ini dapat menjadi sumber referensi dalam memahami penerapan metode *clustering* HDBSCAN untuk segmentasi konsumen pada sektor *e-commerce*.

1.5 Batasan Masalah

Agar Penelitian ini terarah dan fokus, maka dilakukan beberapa batasan masalah sebagai berikut:

1. Penelitian ini difokuskan pada proses segmentasi konsumen berdasarkan data perilaku belanja *e-commerce* yang terdapat dalam dataset *E-commerce Customer Behavior and Sales Analysis*. Perilaku belanja tersebut direpretasikan melalui beberapa variabel, yaitu usia, harga perunit, jumlah perunit, potongan harga, total pembayaran, rata-rata durasi sesi, jumlah tayangan halaman, waktu pengiriman dan penilaian konsumen.
2. Penelitian ini berfokus pada hasil *clustering* dan evaluasi menggunakan *silhouette coefficient* serta *davies bouldin index*, yaitu mengukur kinerja hasil segmentasi konsumen *e-commerce* berdasarkan penilaian kualitas klaster yang terbentuk sehingga dapat melihat perilaku belanja konsumen dalam tiap klaster.

1.6 Definisi Istilah

Terdapat beberapa istilah yang digunakan dalam penelitian ini sebagai berikut:

1. *Clustering*

   Clustering adalah metode *data mining* yang mengelompokkan data ke dalam beberapa grup (*cluster)* berdasarkan kesamaan karakteristik menggunakan label kelas sebelumnya.

2. *HDBSCAN*

   Hierarchical Density-Based Spatial Clustering of Apllications with Noise (HDBSCAN) adalah algoritma *clustering* berbasis kepadatan yang membentuk klaster secara hierarkis dan mampu mengidentifikasi data *noise* atau *outlier* secara otomatis.

3. *E-commerce*

   E-commerce adalah kegiatan jual beli barang atau jasa yang dilakukan secara daring melalui jaringan internet.

4. Konsumen *E-commerce* Turki

   Konsumen *E-commerce* Turki adalah individu yang dibuat secara rekayasa untuk menggambarkan perilaku belanja secara daring melalui platform *e-commerce* di Turki, berdasarkan pola belanja nyata yang terjadi pada tahun 2023 hingga 2024.

5. Segmentasi Konsumen

   Segmentasi Konsumen adalah proses pengelompokan konsumen ke dalam beberapa kelompok berdasarkan perilaku belanja mereka.

6. Perilaku belanja

   Perilaku belanja adalah pola tindakan konsumen dalam mencari memilih, membeli, menggunakan, hingga mengevaluasi produk atau layanan berdasarkan kebutuhan, preferensi, dan faktor lingkungan yang memengaruhi keputusan pembelian.

7. *Silhouette Coefficient*

   Silhouette Coefficient adalah metrik evaluasi *clustering* yang mengukur seberapa baik suatu data berada dalam klasternya dibandingkan dengan klaster

lain berdasarkan jarak rata-rata antar titik.

8. *Davies-Bouldin Index* (DBI)

   Davies-Bouldin Index (DBI) adalah metrik evaluasi yang menilai kualitas klaster berdasarkan rasio antara jarak intra-klaster dan inter-klaster, di mana nilai yang lebih kecil menunjukkan hasil yang lebih baik.

9. *Core Point*

   Core point adalah titik data yang memiliki jumlah tetangga minimal dalam radius tertentu, sehingga dianggap mewakili kepadatan inti dari suatu klaster.

10. *Core Distance*

    Core Distance adalah jarak minimum antara suatu titik dengan tetangga ke-MinPts-nya (*Minimum Points*), yang menunjukkan tingkat kepadatan di sekitar titik tersebut.

11. *Noise*

    Noise adalah titik data yang tidak termasuk dalam klaster mana pun karena tidak memiliki cukup tetangga atau tidak memenuhi tingkat kepadatan minimum yang ditentukan algoritma *clustering* seperti HDBSCAN.

12. *Outlier*

    Outlier adalah titik data yang memiliki karakteristik berbeda secara signifikan dari sebagian besar data lainnya, sehingga berada jauh dari klaster utama dan sering dianggap sebagai data *anomaly*.

13. *Mutual Reachability Distance*

    Mutual Reachablity Distance adalah jarak antara dua titik yang telah disesuaikan berdasarkan kepadatan di sekitar masing-masing titik, digunakan dalam HDBSCAN untuk menstabilkan perhitungan jarak antar titik.

14. *Minimum Spanning Tree*

    Minimum Spanning Tree adalah struktur graf yang menghubungkan semua titik data dengan total bobot jarak minimum tanpa membentuk siklus, digunakan dalam HDBSCAN untuk merepresentasikan hubungan terdekat antar titik berdasarkan *mutual reachability distance*.

15. Algoritma Prim

    Algoritma Prim adalah metode untuk membangun *Minimum Spanning Tree* (MST) dengan cara menambahkan simpul (*node*) baru yang memiliki bobot tepi (*edge*) terendah secara bertahap hingga semua titik terhubung tanpa membentuk siklus.

16. *Minimum_Cluster_Size*

    Minimum_Cluster_Size adalah parameter pada HDBSCAN yang menentukan jumlah minimum titik dalam satu klaster agar dianggap valid, berfungsi untuk membedakan antara klaster yang bermakna dan kumpulan titik acak atau *noise*.

17. *Min_Sample*

    Min_Sample adalah jumlah minimum tetangga yang harus dimiliki suatu titik agar dapat dianggap sebagai inti (*core point*) dalam algoritma HDBSCAN.

BAB II

KAJIAN TEORI

2.1 *Preprocessing Data*

Tahap *preprocessing data* merupakan langkah awal yang penting dalam proses analisis data untuk memastikan bahwa data siap digunakan pada tahap pengolahan selanjutnya. Pada tahap ini, data mentah diproses menggunakan normalisasi *Min-Max Scaler* agar setiap variabel berada pada rentang nilai yang sama, sehingga perbedaan skala tau nilai ekstrem tidak memengaruhi hasil analisis, khususnya pada metode *clustering* yang sensitif terhadap perbedaan skala antar atribut.

Dalam proses normalisasi, yang disesuaikan adalah skala nilai, bukan satuan dari atribut. Satuan menunjukkan jenis ukuran suatu variabel, seperti rupiah untuk harga atau kilogram untuk berat. Satuan tidak menjadi masalah utama dalam analisis data, karena yang lebih berpengaruh adalah seberapa besar rentang nilai (skala) antar atribut tersebut. Perbedaan skala dapat menyebabkan ketidakseimbangan pada metode analisis berbasis jarak, karena atribut dengan nilai numerik yang jauh lebih besar akan memberikan kontribusi lebih dominan dibanding atribut lain. Oleh karena itu, normalisasi dilakukan untuk menyeragamkan skala setiap variabel ke dalam rentang tertentu misalnya 0 hingga 1 pada metode *Min-Max*, agar setiap atribut memiliki pengaruh yang setara dalam proses analisis. Dengan demikian, normalisasi tidak mengubah satuannya, tetapi memastikan bahwa perbedaan skala antar variabel tidak menimbulkan bias dalam perhitungan jarak atau pemodelan lainnya (Hapsari, 2024). Rumus normalisasi *min-max scaler* adalah:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

Keterangan:

x' : Nilai hasil normalisasi.

x : Nilai persatuan data.

x_{min} : Nilai minimum dari seluruh data.

x_{max} : Nilai maksimum dari seluruh data.

2.2 *Clustering*

Clustering (pengelompokan) didefinisikan sebagai teknik untuk mencari struktur "pengelompokan alami" dalam data berdasarkan ukuran kemiripan (*similarity*) atau jarak (*distance*/*dissimilarity*), dan dibedakan secara tegas dari metode klasifikasi di mana klasifikasi mengasumsikan jumlah kelompok sudah diketahui, sedangkan *cluster analysis* tidak membuat asumsi apapun mengenai jumlah maupun struktur kelompok. Dalam menentukan kemiripan antar objek, *cluster analysis* menggunakan berbagai ukuran jarak, di antaranya *Euclidean distance* (jarak lurus antar dua titik), *statistical distance* atau jarak Mahalanobis (yang memperhitungkan korelasi dan variansi antar variabel), *Manhattan distance* (jumlah selisih absolut antar koordinat), *Canberra metric*, serta koefisien *Czekanowski* yang khusus digunakan untuk variabel bernilai non-negatif. Pemilihan ukuran jarak yang tepat sangat krusial karena definisi "kemiripan" secara langsung menentukan bagaimana objek-objek dikelompokkan dan hasil akhir dari analisis *clustering* itu sendiri (Johnson & Wichern, 2007)

Menghitung jarak antara data ke pusat *Cluster*, menggunakan metode perhitungan jarak. Ada beberapa persamaan kemiripan jarak di klaster, sebagai berikut:

1. *Euclidean Distance*

   Euclidean distance merupakan ukuran jarak antara dua titik dalam ruang berdimensi p yang menggambarkan seberapa dekat atau jauh kedua titik tersebut. Dalam analisis multivariat dan *clustering*, jarak ini digunakan untuk mengukur kedekatan antara dua *vector* data berdasarkan seuluruh atributnya. Secara matematis, jarak *Euclidean* antara dua titik $p_i = (p_{i1}, p_{i2}, \dots, p_{iz})$ dan $p_j = (p_{1j}, p_{2j}, \dots, p_{jz})$ dirumuskan sebagai:

$$d(p_i, p_j) = \sqrt{(p_{i1} - p_{j1})^2 + (p_{i2} - p_{j2})^2 + \cdots + (p_{iz} - p_{jz})^2}$$

   Atau dalam sigma:

$$d(p_i, p_j) = \sqrt{\sum_{l=1}^{z} (p_{il} - p_{jl})^2} \qquad (2.2)$$

   Keterangan:

   $d(p_i, p_j)$: Jarak *Euclidean* antara p_i dan p_j.

   p_i : Titik data ke-i.

   p_j : Titik data ke-j.

   l : Indeks atribut.

   z : Jumlah atribut/variabel.

   p_{il} : Nilai atribut ke-l pada titik p_i.

   p_{jl} : Nilai atribut ke-l pada titik p_j.

2.3 HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)merupakan metode analisis klaster yang dikembangkan sebagai penyempurnaan dalam menangani data pencilan. Sebagai modifikasi dari algoritma

Ordering Points to Identify the Clustering Structure (OPTICS), metode ini menggunakan satu parameter utama, yaitu MinPts (*Minimum Points*). Parameter MinPts atau *min_sample* merupakan jumlah minimum tetangga terdekat yang digunakan untuk mengukur kepadatan lokal suatu titik data dalam proses perhitungan *core distance*. Parameter ini berbeda dengan *min_cluster_size*, yang digunakan untuk menentukan jumlah minimum anggota dalam suatu *cluster* agar dapat dianggap sebagai *cluster* yang valid (Ghiffary dkk., 2024).

Algoritma HDBSCAN merupakan pengembangan dari metode DBSCAN yang memperkenalkan konsep kepadatan hierarkis untuk mengatasi keterbatasan parameter *epsilon* tunggal pada DBSCAN (Moulavi dkk., 2014). Berbeda dengan DBSCAN yang hanya menggunakan satu nilai ambang (ε) yang sama untuk seluruh titik data, HDBSCAN membangun struktur hierarki klaster dari berbagai tingkat kepadatan dengan menghitung *core distance* dan *mutual reachability distance* antar titik (Campello dkk., 2013). Algoritma HDBSCAN dikembangkan sebagai perluasan dari metode DBSCAN dengan memperkenalkan pendekatan hierarkis untuk mengatasi keterbatasan penentuan parameter *epsilon* tunggal (Alemán dkk., 2022). Adapun beberapa tahapan untuk menghitung algoritma HDBSCAN yaitu (Campello dkk., 2015):

1. Menghitung *core distance*

   Menghitung *core distance* digunakan untuk menilai tingkat kepadatan lokal di sekitar titik data. Secara umum, *core distance* untuk suatu titik data p_i dengan parameter jumlah tetangga terdekat k dituliskan sebagai berikut (Campello dkk., 2013):

$$core_k(p_i) = d(p_i, tetangga\ terdekat\ ke-k\ dari\ titik\ \ p_i) \tag{2.3}$$

Untuk memudahkah pemahaman dalam konteks matematis, berikut disajikan definisi formal dari setiap komponen yang terlibat

Misalkan terdapat himpunan data:

$$p = \{p_i, p_2, \dots, p_n\} \subset \mathbb{R}^z$$

Dengan setiap titik data

$$p_i = (p_{i1}, p_{i2}, \dots, p_{iz})$$

Merupakan vektor berdimensi z.

Parameter k menyatakan jumlah tetangga terdekat yang dipertimbangkan. Nilai k digunakan untuk menentukan tetangga ke-k dari titik p_i, yaitu titik yang memiliki jarak urutan ke-k terkecil terhadap p_i. Parameter ini sesuai dengan *minPts* atau *min_samples* dalam algoritma HDBSCAN.

Untuk mendapatkan tetangga ke-k, terlebih dahulu dihitung jarak antara titik p_i dan setiap titik lain $p_j \in P$, dengan i $\neq j$:

$$d_{ij} = d(p_i, p_j) \tag{2.4}$$

Pada penelitian ini digunakan jarak *Euclidean*, dapat dilihat di Persamaan (2.2). Seluruh nilai jarak d_{ij} kemudian diurutkan dari yang terkecil hingga yang terbesar:

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n-1)}.$$

Pengurutan ini menghasilkan urutan titik tetangga terdekat:

$$p_{(1)}, p_{(2)}, \dots, p_{(n-1)},$$

Dimana:

$p_{(1)}$: Tetangga terdekat pertama.

$p_{(2)}$: Tetangga terdekat kedua.

$p_{(k)}$: Tetangga terdekat ke-k dari titik p.

Dengan demikian, makna formal dari "k-*nearest neighbor of* p_i" adalah titik $p_{(k)}$.

Berdasarkan definisi tersebut, bentuk rinci dari Persamaan (2.3) dapat dituliskan sebagai:

$$core_k(p_i) = dist(p_i, p_{(k)}) = d_{(k)}.$$

Secara langsung, *core distance* dapat dituliskan sebagai:

$$core_k(p_i) = \sqrt{\sum_{l=1}^{z} \left(p_{il} - p_{(k)l}\right)^2} \tag{2.5}$$

Keterangan:

$core_k(p_i)$: *Core distance* titik p_i.

p_i : Titik data ke-i.

p_j : Ttitik data ke-j.

$p_{(k)}$: Tetangga terdekat ke-k dari titik p_i.

z : Jumlah atribut.

p_{il} : Atribut ke-l pada titik p_i.

p_{jl} : Atribut ke-l pada titik p_j.

$p_{(k)l}$: Atribut ke-l pada tetangga ke-k.

d_{ij} : Jarak antara titik p_i dan titik p_j.

2. Menghitung *Mutual Reachability Distance*

Dalam algoritma HDBSCAN, jarak antar titik tidak hanya dihitung menggunakan jarak *Euclidean* biasa, tetapi diperhalus melalui *mutual reachability distance* untuk menstabilkan perhitungan kepadatan lokal antar titik (McInnes dkk., 2017). Konsep ini digunakan untuk mengurangi pengaruh

noise dan membuat struktur *cluster* lebih stabil. Bentuk matematisnya diberikan sebagai berikut:

$$mreach_k(p_i, p_j) = \max\{core_k(p_i),\ core_k(p_j),\ d(p_i, p_j)\} \quad (2.6)$$

Agar persamaan tersebut dapat dipahami secara formal dalam konteks ruang vektor berdimensi z, berikut penjelasan matematis setiap komponennya.

Himpunan data dan titik yang dibandingkan.

Misalkan himpunan data didefinisikan sebagai:

$$p = \{p_1, \mathrm{p}_2, \dots, p_n\} \subset \mathbb{R}^d,$$

dengan setiap elemen berupa vektor:

$$p_i = (p_{i1}, p_{i2}, \dots, p_{iz}).$$

Dua titik data yang sedang dibandingkan dinotasikan sebagai:

$$p_i = (p_{i1}, p_{i2}, \dots, p_{iz}).$$

Dan,

$$p_j = (p_{j1}, p_{j2}, \dots, p_{jz}).$$

Jarak *Euclidean* Antar Titik

Jarak *Euclidean* antara titik $p_i$dan p_j dapat dilihat di Persamaan (2.2).

Core Distance Tiap Titik

Nilai $core_k(p_i)$dan $core_k(p_j)$diperoleh dari langkah sebelumnya:

$$core_k(p_i) = d(p_i, p_{i(k)}),$$

$$core_k(p_j) = d(p_j, p_{j(k)})$$

Dimana

$p_{i(k)}$: Tetangga terdekat ke-k dari titik p_i,

$p_{j(k)}$: Tetangga terdekat ke-k dari titik p_j.

Pemilihan k mengikuti parameter minPts atau *min_samples*. Dengan

menggunakan komponen-komponen yang digunakan dalam perhitungan *mutual reachbility distance* terdiri dari *core distance* dari masing-masing titik yang dibandingkan serta jarak langsung antar kedua titik tersebut. *Core distance* merepresentasikan tingkat kepadatan lokal suatu titik berdasarkan parameter *min_sample*, sedangkan jarak antar titik dihitung menggunakan jarak *Euclidean*. Ketiga komponen ini kemudian digunakan dalam fungsi maksimum untuk menghasilkan nilai *mutual reahbility distance*. Definisi *mutual reachability distance* (Persamaan 2.6) dapat dituliskan secara eksplisit:

$$mreach_k(p_i, p_j) = \max\{d(p_i, p_{i(k)}),\ d(p_j, p_{j(k)}),\ d(p_i, p_j)\}. \quad (2.7)$$

Artinya, jarak antara dua titik a dan b pada HDBSCAN tidak hanya dilihat dari jarak langsung $\text{dist}\,(p_i, p_j)$, tetapi juga tingkat kepadatan lokal (*core distance*) masing-masing titik melalui nilai *core distance*. Kepadatan lokal menunjukkan seberapa dekat suatu titik dengan tetangga-tetangga terdekatnya. Jika suatu titik memiliki tetangga yang jaraknya dekat, maka titik tersebut berada pada area yang padat. Sebaliknya, jika jarak titik tersebut dengan tetangga terdekatnya relatif jauh, maka titik tersebut berada pada area yang kurang padat. Oleh karena itu, *mutual reachbility distance* membantu HDBSCAN membentuk klaster berdasarkan kedekatan jarak sekaligus tingkat kerapatan titik data.

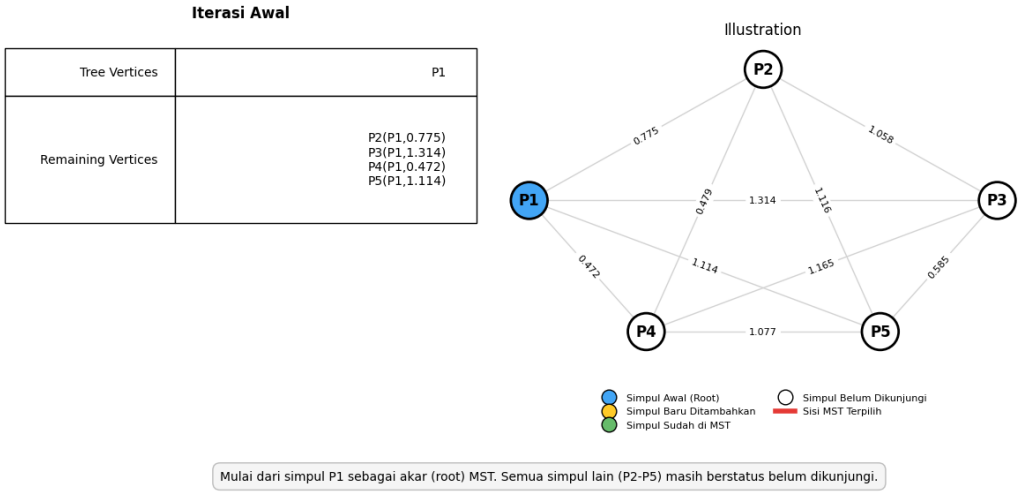
3. Membangun *Minimum Spanning Tree* (MST)

   Setelah memperoleh nilai *mutual reachability distance* dari setiap pasangan titik data pada langkah sebelumnya, algoritma HDBSCAN membangun suatu graf berbobot yang memuat seluruh titik data sebagai simpul, dan bobot antar simpul didefinisikan sebagai (Campello dkk., 2013):

$$w(p_i, p_j) = mreach_k(p_i, p_j). \quad (2.8)$$

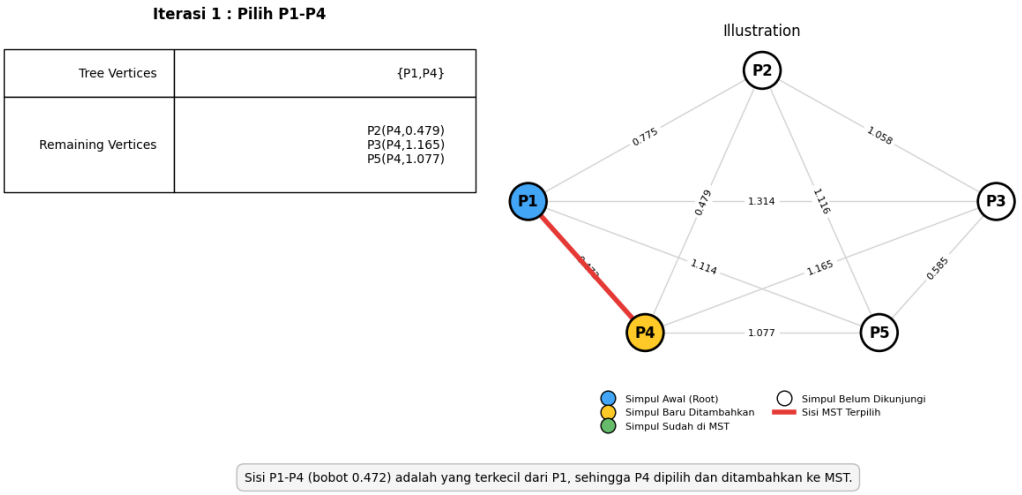
Graf berbobot ini kemudian digunakan untuk membentuk *Minimum Spanning Tree* (MST), yaitu himpunan sisi yang menghubungkan seluruh titik data dengan total bobot minimum tanpa membentuk siklus. Dalam HDBSCAN, MST penting untuk mengidentifikasi struktur hierarki klaster melalui pemodelan hubungan antar titik berbasis kepadatan (McInnes dkk., 2017). Proses pembentukan MST dilakukan menggunakan algoritma Prim, yang efisien untuk graf berbobot padat seperti yang dibentuk oleh *mutual reachability distance*.

Algoritma Prim merupakan algoritma greedy yang digunakan untuk membentuk *Minimum Spanning Tree* (MST) pada graf berbobot terhubung dengan cara menambahkan satu simpul ke dalam pohon secara bertahap melalui sisi yang memiliki bobot minimum. Proses dimulai dari sebuah simpul awal yang dipilih secara bebas, kemudian pada setiap iterasi dipilih sisi dengan bobot yang terkecil yang menghubungkan simpul yang sudah berada dalam MST dengan simpul yang belum tergabung. Sisi dan simpul terpilih selanjutnya ditambahkan ke dalam MST, dan proses ini diulangin hingga seluruh simpul pada graf terhubung tanpa membentuk siklus. Oleh karena itu, algoritma Prim mampu menghasilkan pohon rentang minimum yang memiliki total bobot paling kecil sehingga terbentuk struktur pohon yang menghubungkan seluruh simpul dengan jumlah bobot sisi minimum (Levitin, 2012). Adapun visualisasi langkah-langkah pembentukan MST menggunakan algoritma Prim:



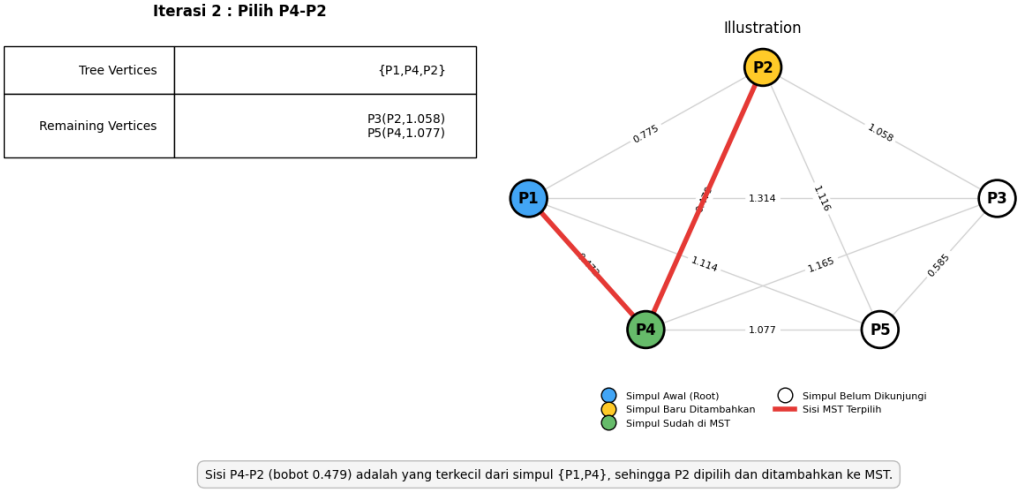
Gambar 2.1 Visualisasi iterasi awal pembentukan MST menggunakan algoritma Prim

Berdasarkan gambar 2.1 visualisasi pada iterasi awal algoritma Prim, simpul p_1 dipilih sebagai titik awal (*root*) dalam pembentukan *Minimum Spanning Tree* (MST). Pada tahap ini, hanya simpul p_1 yang masuk ke dalam MST, sedangkan simpul lainnya (p_2, p_3, p_4 dan p_5) masih berada di luar MST. Seluruh sisi yang terhubung dengan p_1 beserta bobotnya ditampilkan sebagai kandidat untuk dipilih pada iterasi berikutnya.



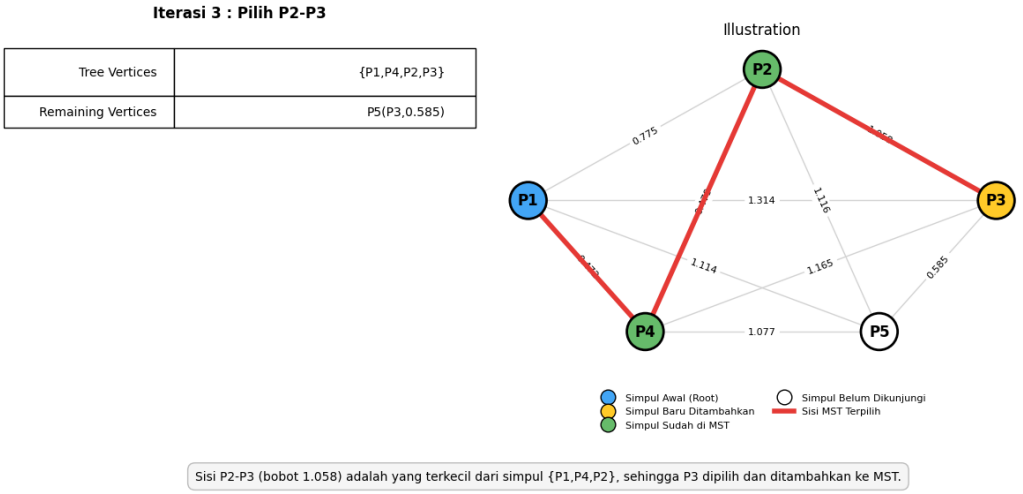
Gambar 2.2 Visualisasi iterasi 1 pembentukan MST menggunakan algoritma Prim

Berdasarkan gambar 2.2 pada iterasi pertama algoritma Prim, sisi $p_1 - p_4$ dipilih karena memiliki bobot paling kecil yaitu, 0,472, dibandingkan seluruh sisi yang menghubungkan simpul dalam MST dengan simpul di luar MST. Dengan terpilihnya sisi tersebut, simpul p_4 berhasil ditambahkan ke dalam MST sehinggan himpunan simpul yang telah tergabung menjadi $\{p_1, p_4\}$.



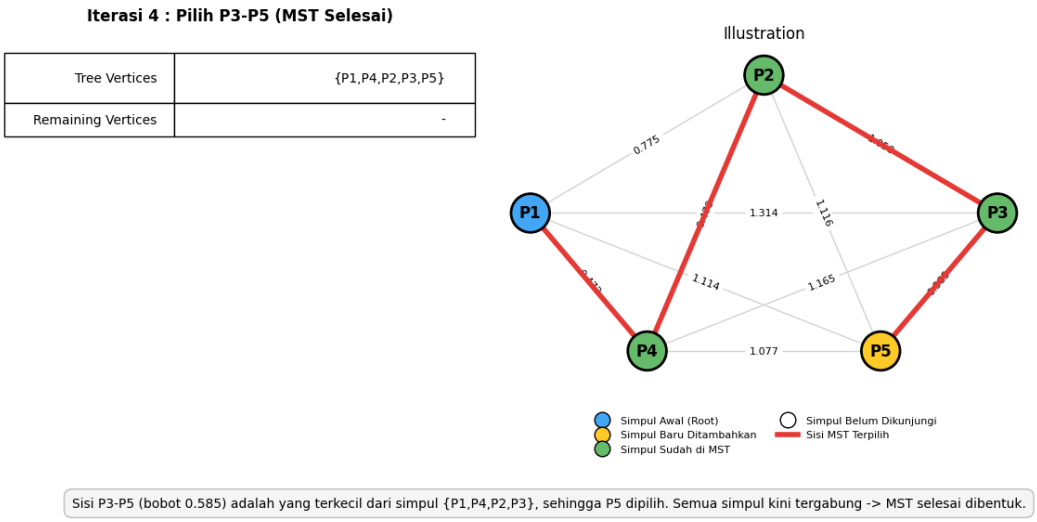
Gambar 2.3 Visualisasi iterasi 2 pembentukan MST menggunakan algoritma Prim

Pada iterasi kedua, algoritma Prim mengevaluasi seluruh sisi yang menghubungkan himpunan simpul dalam MST, yaitu $\{p_1, pp_4\}$, dengan simpul yang masih berasa di luar MST. Dari kandidat yang tersedia, sisi $p_4 - p_2$ memiliki bobot terkecil sebesar 0,479, sehingga dipilih untuk memperluas MST. Pemilihan ini dilakukan untuk memastikan penambahan simpul baru memberikan bobot minimum pada struktur pohon yang sedang dibangun.



Gambar 2.4 Visualisasi iterasi 3 pembentukan MST menggunakan algoritma Prim

Pada iterasi ketiga, algoritma Prim memilih sisi $p_2 - p_3$ dengan bobot 1,058 karena merupakan bobot terkecil yang menghubungkan himpunan simpul dalam MST, yaitu $\{p_1, p_4, p_2\}$ dengan simpul yang masih berada di luar MST. Pemilihan sisi ini dilakukan untuk menjaga agar total bobot MST tetap minimum pada setiap tahap pembentukannya.



Gambar 2.5 Visualisasi iterasi 4 pembentukan MST menggunakan algoritma Prim

Pada iterasi keempat, algoritma Prim memilih sisi $p_3 - p_5$ dengan bobot 0,585 untuk menghubungkan simpul p_5 ke dalam *Minimum Spanning Tree* (MST). Sisi ini merupakan pilihan dengan bobot minimum yang menghubungkan himpunan simpul dalam MST, yaitu $\{p_1, p_4, p_2, p_3\}$, dengan satu-satunya simpul yang masih berada di luar MST, yaitu p_5. Setelah sisi $p_3 - p_5$ ditambahkan, seluruh simpul pada graf telah terhubung sehingga himpunan simpul dalam MST menjadi $\{p_1, p_4, p_2, p_3, p_5\}$. Pada tahap ini menandakan bahwa proses pembentukan MST telah selesai.

Secara matematis, MST didefinisikan sebagai:

$$MST = \underset{T \subseteq G}{\arg min} \sum_{(p_i, p_j) \in T} mreach_k(p_i, p_j) \tag{2.9}$$

Dengan penjelasan komponen sebagai berikut.

Misalkan himpunan titik data adalah:

$$p = \{p_1, p_2, \dots, \mathrm{p_z}\}$$

Graf berbobot didefinisikan sebagai:

$$G = (V, E),$$

Dengan,

Simpul (*vertices*):

$$V = p$$

Yaitu seluruh titik data.

Sisi (*edges*):

$$E = \{(p_i, p_j) \mid p_i, p_j \in V,\ i \neq j\}$$

Yang menyatakan hubungan antar pasangan titik data.

Bobot sisi:

Bobot setiap sisi ditentukan menggunakan *mutual reacbility distance*.

$$w(p_i, p_j) = mreach_k(p_i, p_j).$$

Dengan:

$$mreach_k(p_i, p_j) = \max\{core_k(p_i),\ core_k(p_j),\ d(p_i, p_j)\} \quad (2.10)$$

Bobot ini konsisten dengan hasil langkah sebelumnya.

$G = (V, E)$: Graf berbobot yang mewakili seluruh titik data dan bobot *mutual reachability distance* antar titik.

T : Himpunan sisi yang membentuk *spanning tree* tanpa siklus.

V : Himpunan simpul (*vertices*).

E : Himpunan sisi (*edges*).

$w(p_i, p_j)$: Bobot sisi antara titik p_i dan p_j.

$mreach_k(p_i, p_j)$: *Mutual reachbility distance* antara titik p_i dan p_j.

MST : *Minimum Spanning Tree* dengan total bobot minimum.

4. Membentuk Hierarki Klaster (*Cluster hierarchy*)

   Berdasarkan tingkat kepadatan, menunjukkan pembagian klaster dari padat ke jarang. Proses ini secara otomatis memisahkan *core cluster* dari *noise* dengan mengeliminasi area berisi kepadatan rendah. Kepadatan direpresentasikan oleh parameter λ, yang berbanding terbalik dengan nilai jarak w (Malzer & Baum, 2021)

$$\lambda = \frac{1}{w} \quad (2.11)$$

   Keterangan

   w : Nilai jarak antar titik (*threshold*), biasanya berupa *mutual reachability distance* dari langkah sebelumnya (2.7).

λ : Tingkat kepadatan (*density level*) yang digunakan dalam pembentukan hierarki klaster.

Ketika nilai w menurun atau λ meningkat , maka klaster dengan kepadatan lebih tinggi akan terbentuk dari klaster yang lebih besar.

Untuk membentuk hierarki klaster, setiap tingkat kepadatan λ berhubungan dengan komponen terhubung dari graf *mutual reachability*. Misalkan bobot antar titik diberikan oleh:

$$w(p_i, p_j) = mreach_k(p_i, p_j). \tag{2.12}$$

Dua titik $p_i$dan p_j dianggap terhubung pada level kepadatan λ apabila memenuhi:

$$w(p_i, p_j) \leq \frac{1}{\lambda} \tag{2.13}$$

Dengan demikian, *cluster* pada tingkat kepadatan λ didefinisikan sebagai himpunan *connected components* dari graf:

$$G(\lambda) = \left\{(p_i, p_j) \middle| mreach_k(p_i, p_j) \leq \frac{1}{\lambda}\right\} \tag{2.14}$$

Ketika λ semakin besar atau $w = \frac{1}{\lambda}$ semakin kecil, maka *cluster* akan terpecah menjadi *cluster-cluster* yang lebih kecil namun memiliki tingkat kepadatan lebih tinggi.

Penjelasan ini menunjukkan secara matematis bahwa ketika nilai w diturunkan atau λ dinaikkan, terbentuk klaster yang memiliki kepadatan lebih tinggi. Kepadatan (*density*) dalam konteks *clustering* merujuk pada tingkat kedekatan antar titik data dalam suatu *cluster*. *Cluster* dikatakan lebih padat apabila jarak

antar titik dalam *cluster* tersebut relatif kecil, sehingga titik-titik data terkonsentrasi dalam suatu area yang sempit.

5. Penyederhanaan Hierarki Berdasarkan *Minimum Cluster Size*

Setelah struktur hierarki di hasilkan oleh proses HDBSCAN, tahap selanjutnya adalah melakukan penyaringan (*condensation*) untuk mempertahankan klaster yang berukuran cukup besar. Parameter yang digunakan pada tahap ini adalah *minimum cluster size* yang dinotasikan dengan m_{clsize}. (Campello dkk., 2015)

Setiap klaster C_i memiliki himpunan anggota:

$$C_i = \{p_{i_1}, p_{i_2}, \dots, p_{i_{N(C_i)}}\}$$

Dengan,

$$N(C_i) = \mid C_i \mid$$

yang menyatakan banyaknya titik dalam klaster C_i.

Pada tahap *condensation*, HDBSCAN mempertahankan *cluster* apabila ukuran *cluster* memenuhi batas minimum yang ditentukan, yaitu:

$$C_i \text{ valid} \iff N(C_i) \geq m_{\text{clsize}} \tag{2.15}$$

Sebaliknya, klaster akan dieliminasi apabila jumlah anggotanya lebih kecil daripada parameter *minimum_cluster_size*

$$C_i \text{ tidak valid} \iff N(C_i) < m_{\text{clsize}} \tag{2.16}$$

Ketika kondisi (2.12) terpenuhi, maka klaster dikeluarkan dari struktur hierarki:

$$N(C_i) < m_{\text{clsize}} \Longrightarrow C_i \notin \mathcal{H} \tag{2.17}$$

dengan $\mathcal{H}$ adalah himpunan klaster pada struktur hierarki yang dipertahankan setelah proses penyaringan.

m_{clsize} : Parameter *minimum cluster size* (ukuran minimal klaster).

C_i : Klaster ke-i pada hierarki.

$N(C_i) =\mid C_i \mid$: Jumlah anggota klaster C_i

$\mathcal{H}$: Struktur hierarki klaster yang tersisa setelah tahap penyaringan.

p_{ij} : Titik data anggota *cluster* C_i

6. Ekstraksi Klaster Berdasarkan Stabilitas

Setelah proses penyederhanaan hierarki, langkah berikutnya adalah mengekstraksi klaster berdasarkan stabilitas (*stability*). Stabilitas menggambarkan seberapa lama sebuah klaster bertahan pada berbagai tingkat kepadatan (nilai λ). Semakin lama sebuah klaster bertahan (semakin "hidup" dalam ruang kepadatan), semakin layak klaster tersebut dipilih sebagai klaster akhir (Moulavi dkk., 2014).

Dalam konteks HDBSCAN, setiap klaster $C_i$memiliki:

$\lambda_{\min}(C_i)$: Tingkat kepadatan saat klaster C_i mulai terbentuk.

$\lambda_{\max}(p_j, C_i)$: Tingkat kepadatan saat titik p_j"keluar" dari klaster (tidak lagi menjadi anggota).

Untuk setiap titik $p \in C_i$, rentang hidupnya di dalam klaster adalah:

$$\Delta\lambda(p_j \mid C_i) = \lambda_{\max}(p_j, C_i) - \lambda_{\min}(C_i) \quad (2.18)$$

Definisi Matematis Stabilitas Klaster

Total stabilitas klaster C_i didefinisikan sebagai jumlah seluruh rentang hidup semua anggotanya:

$$S(C_i) = \sum_{p_j \in C_i} (\lambda_{\max}(p_j, C_i) - \lambda_{\min}(C_i)) \quad (2.19)$$

Definisi ini sejalan dengan konsep *cluster persistence*:

Klaster yang mempertahankan banyak titik untuk rentang kepadatan yang luas dianggap lebih stabil dan lebih representatif.

Struktur dan Penjelasan Secara Formal

a. Himpunan Anggota Klaster

$$C_i = \{p_1, p_2, \dots, p_{N(C_i)}\} \tag{2.20}$$

b. Rentang hidup tiap anggota

$$\Delta\lambda(p_j \mid C_i) = \lambda_{\max}(p_j, C_i) - \lambda_{\min}(C_i) \tag{2.21}$$

c. Stabilitas total klaster

$$(C_i) = \sum_{j=1}^{N(C_i)} \Delta\lambda(p_j \mid C_i) \tag{2.22}$$

d. Kriteria pemilihan klaster akhir

Klaster akhir dipilih sebagai klaster dengan nilai stabilitas tertinggi dalam struktur hierarki:

$$C_{\text{final}} = \underset{C_i}{\text{argmax}}\ S(C_i) \tag{2.23}$$

Keterangan

$S(C_i)$: Stabilitas Klaster C_i, menggambarkan total masa hidup klaster pada rentang kepadatan.

$\underset{C_i}{\text{argmax}}$: Argumen *maximal cluster* ke-i.

C_i : *Cluster* ke-i.

p_j : Titik data anggota klaster C_i.

$\lambda_{\min}(C_i)$: Tingkat kepadatan saat klaster mulai terbentuk (*birth*).

$\lambda_{\max}(p_j, C_i)$: Tingkat kepadatan saat titik $x_j$keluar dari klaster (*death*).

$\Delta\lambda(p_j \mid C_i)$: Rentang hidup titik p_j dalam klaster.

C_{final} : *Cluster* akhir yang dipilih berdasarkan stabilitas maksimum.

$N(C_i)$: Jumlah anggota pada *cluster* C_i

2.4 Evaluasi Hasil *Cluster*

1. *Silhouette Coefficient*

   Silhouette Coefficent merupakan metode evaluasi *clustering* secara intrinsik yang digunakan untuk mengukur kualitas hasil *clustering* berdasarkan tingkat kekompakan (*compactness*) dan keterpisahan (*separation*) antar *cluster*. Nilai ini dihitung menggunakan rata-rata jarak suatu objek terhadap anggota dalam *cluster* yang sama serta jarak rata-rata minimum terhadap *cluster* lain. Semakin mendekati nilai 1, maka *cluster* yang terbentuk semakin kompak dan semakin terpisah dari *cluster* lainnya (Han dkk., 2012). Sebaliknya, nilai negatif menunjukkan bahwa suatu objek cenderung lebih dekat dengan *cluster* lain dibandingkan *cluster* tempat objek tersebut berada. Lebih jauh, rata-rata lebar *silhouette* (*average silhouette width*) dari keseluruhan dataset dapat digunakan sebagai kriteria untuk memilih jumlah klaster k yang paling optimal, yaitu dengan memilih nilai k yang menghasilkan rata-rata $s(i)$ terbesar, sehingga metode ini tidak hanya berguna untuk menilai kualitas tiap klaster secara individual, tetapi juga membantu peneliti menentukan berapa jumlah klaster yang paling alami dan bermakna dalam data(Rousseeuw, 1987).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2.24)$$

Keterangan:

$a(i)$: Jarak rata- rata antar anggota kelompok.

$b(i)$: Nilai terkecil antar jarak rata-rata objek i dengan objek pada kelompok lainnya.

2. *Davies Bouldin Index* (DBI)

Davies Bouldin Index (DBI) adalah ukuran evaluasi *cluster* yang bekerja dengan cara mengukur rata-rata kemiripan setiap klaster dengan klaster lain yang paling mirip dengannya, di mana kemiripan antar klaster dihitung berdasarkan perbandingan antara tingkat penyebaran (*dispersion*) data di dalam klaster dengan jarak antar pusat klaster, sehingga klaster yang kompak dan berjauhan satu sama lain akan menghasilkan nilai DBI yang kecil. Prinsip utamanya adalah semakin kecil nilai DBI, semakin baik kualitas hasil *clustering*, karena nilai yang rendah mencerminkan klaster-klaster yang rapat di dalam (*tight*) sekaligus terpisah jauh satu sama lain (*well-separated*). Keunggulan DBI terletak pada sifatnya yang tidak bergantung pada algoritma *clustering* tertentu maupun jumlah klaster yang dianalisis, sehingga dapat digunakan secara fleksibel untuk membandingkan berbagai hasil partisi data dan membantu peneliti menentukan jumlah klaster *k* yang paling optimal, yaitu nilai *k* yang menghasilkan DBI paling minimum(Fu et al., 1977).Nilai *Davies Bouldin Index* rendah menunjukkan *cluster* terpisah satu sama lain dengan baik, sedangkan nilai tinggi menunjukkan adanya tumpang tindih antar *cluster* (Wororomi dkk., 2024). Adapun persamaan yang digunakan adalah sebagai berikut:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_{ij} \tag{2.25}$$

dengan

$$R_{ij} = maxR_{ij}$$

dan

$$R_{ij} = \frac{var(C_i) + var(C_j)}{|c_i - c_j|}$$

Dimana C_i merupakan *centroid* dari *cluster* i

Keterangan:

k : Jumlah *cluster* yang digunakan,

R_{ij} : Rasio perbandingan antara *cluster* ke i dan cluster ke j.

Nilainya yang didapatkan dari komponen kohesi dan separasi. *Cluster* yang baik adalah yang mempunyai kohesi sekecil dan separasi sebesar mungkin.

2.5 Segementasi dan *E-Commerce*

E-Commerce seperangkat media pembelian dan penjualan jasa atau produk yang dilakukan antara dua pihak melalui internet dan seperti mekanisme bisnis yang pada transaksi bisnis antara individu dengan internet sebagai media komunikasi dalam melakukan penjualan tersebut (Soetiyani dkk., 2024). Saat ini, *e-commerce* banyak diterapkan oleh perusahaan karena memiliki berbagai keunggulan (Novy dkk., 2023). Platform ini berfungsi sebagai media transaksi dengan tingkat efisiensi yang tinggi serta dapat diakses dengan mudah kapan pun dan di mana pun (Natania & Dwijayanti, 2024). Salah satu manfaat utama penggunaan *e-commerce* bagi pelaku usaha adalah kemampuannya dalam memperluas jangkauan pasar hingga ke tingkat internasional (Nurjannah & Ratnah S, 2024). Selain itu, teknologi *e-commerce* memungkinkan operasional bisnis berjalan lintas negara dengan biaya

yang lebih rendah, proses penjualan yang lebih cepat, pengurangan potensi kesalahan manusia (*human error*), serta penghematan biaya pencetakan dan administrasi.

Segmentasi merupakan proses pemisahan pasar ke dalam kelompok konsumen yang bersifat homogen, di mana setiap kelompok dapat dijadikan target pasar sesuai strategi perusahaan. Segmentasi pelanggan dilakukan dengan mengelompokkan konsumen berdasarkan perbedaan karakteristik, perilaku, dan kebutuhan mereka. Penerapan strategi ini berperan penting dalam mengidentifikasi tingkat loyalitas pelanggan serta membantu perusahaan merancang strategi pemasaran yang lebih tepat sasaran, efektif, dan efisien (Perdana dkk., 2022). Dalam penelitian ini, setiap klaster diinterpretasikan berdasarkan nilai rata-rata variabel pada masing-masing kelompok, seperti total pembayaran, jumlah produk, potongan harga, durasi sesi, jumlah tayangan halaman, waktu pengiriman, dan penilaian konsumen. Nilai rata-rata tersebut digunakan untuk melihat ciri utama dari setiap klaster, sehingga penamaan klaster dapat disesuaikan dengan pola perilaku belanja konsumen. Dengan demikian, hasil segmentasi tidak hanya menunjukkan jumlah klaster, tetapi juga menjelaskan perbedaan perilaku konsumen pada masing-masing kelompok (Ramkumar dkk.,2025).

2.6 Perilaku Belanja

Perilaku belanja merupakan pola tindakan dan keputusan konsumen dalam mencari, memilih, membeli, menggunakan, hingga mengevaluasi produk atau layanan yang dipengaruhi oleh kebutuhan, preferensi, motivasi, serta faktor lingkungan (Fadilah Aswar, 2025). Menurut teori perilaku konsumen, proses belanja mencerminkan respons individu terhadap stimulus internal dan eksternal

yang kemudian membentuk kebiasaan dan kecenderungan tertentu dalam aktivitas pembelian (Ismail dkk., 2020). Faktor-faktor seperti persepsi, sikap, pengalaman sebelumnya, serta pengaruh sosial dan teknologi turut menentukan variasi perilaku belanja pada setiap konsumen. Dalam konteks *e-commerce*, perilaku belanja menjadi semakin dinamis karena kemudahan akses informasi, variasi produk, dan kecepatan layanan digital yang memengaruhi pola pengambilan keputusan. Dengan demikian, pemahaman terhadap perilaku belanja penting untuk menganalisis perbedaan karakteristik konsumen dan menjadi dasar dalam proses segmentasi yang tepat sasaran.

2.7 Kajian Integrasi Ayat Alqur'an dan Hadist

Dalam perspektif Islam, kegiatan ekonomi dan transaksi digital harus dijalankan dengan menjunjung tinggi nilai-nilai kejujuran, keadilan, dan tanggung jawab. Setiap proses jual beli hendaknya dilakukan secara transparan dan berdasarkan kerelaan kedua belah pihak agar tercipta kepercayaan serta keberkahan dalam transaksi. Prinsip ini menjadi dasar bagi terciptanya ekosistem *e-commerce* yang sehat, di mana pelaku usaha tidak hanya berorientasi pada keuntungan, tetapi juga memperhatikan etika dan keseimbangan moral. Kejujuran dalam penyampaian informasi produk, keadilan dalam penetapan harga, dan tanggung jawab terhadap kepuasan konsumen merupakan wujud nyata penerapan nilai-nilai Islam dalam ekonomi modern. Dengan mengedepankan integritas dan keadilan, dunia bisnis digital dapat menjadi sarana untuk mencapai kemaslahatan bersama sekaligus menjaga keberkahan dalam setiap aktivitas ekonomi.

Al-Qur'an juga menekankan pentingnya kejujuran, keadilan, dan prinsip suka sama suka dalam setiap transaksi ekonomi, sebagaimana termaktub dalam surah

An-Nisa ayat 29.

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا لَا تَأْكُلُوْٓا اَمْوَالَكُمْ بَيْنَكُمْ بِالْبَاطِلِ اِلَّآ اَنْ تَكُوْنَ تِجَارَةً عَنْ تَرَاضٍ مِّنْكُمْۗ وَلَا تَقْتُلُوْٓا اَنْفُسَكُمْۗ اِنَّ اللّٰهَ كَانَ بِكُمْ رَحِيْمًا ٢٩

Terjemahan:
"Wahai orang-orang yang beriman, janganlah kamu memakan harta sesamamu dengan cara yang batil (tidak benar), kecuali berupa perniagaan atas dasar suka sama suka di antara kamu. Janganlah kamu membunuh dirimu. Sesungguhnya Allah adalah Maha Penyayang kepadamu" (QS. Al-An-Nisa' [4]: 29)(Kementrian Agama, 2022)

Ayat ini menegaskan prinsip kejujuran, keadilan, dan kerelaan dalam setiap bentuk transaksi ekonomi. Dalam konteks *e-commerce*, ayat tersebut menjadi landasan moral agar setiap aktivitas jual beli dilakukan dengan transparansi dan tanpa adanya kecurangan, seperti manipulasi data, penipuan produk, atau eksploitasi pelanggan. Nilai ini sejalan dengan tujuan penelitian yang menggunakan metode HDBSCAN untuk memahami perilaku konsumen secara objektif melalui analisis data. Dengan segmentasi pelanggan yang akurat dan adil, perusahaan dapat merancang strategi yang tidak hanya berorientasi pada keuntungan, tetapi juga menjunjung nilai kejujuran dan kepuasan pelanggan sesuai prinsip 'an tarāḍin minkum (saling ridha). Hal ini menunjukkan bahwa penerapan analisis data dalam *e-commerce* tidak sekadar berfungsi bisnis, tetapi juga sebagai bentuk penerapan etika Islam dalam menjaga keadilan dan keberkahan transaksi digital.

Adapun menurut tafsir Riwayat Hadist Bukhari dan Muslim menyebutkan:

عَنْ عَبْدِ اللَّهِ بْنِ عُمَرَ رَضِيَ اللَّهُ عَنْهُمَا :قَالَ رَسُولُ اللَّهِ ﷺ
إِذَا تَبَايَعَ الرَّجُلَانِ فَكُلُّ وَاحِدٍ مِنْهُمَا بِالْخِيَارِ مَا لَمْ يَتَفَرَّقَا، وَكَانَا جَمِيعًا، أَوْ يُخَيِّرُ أَحَدُهُمَا الْآخَرَ، فَإِنْ»
خَيَّرَ أَحَدُهُمَا الْآخَرَ فَتَبَايَعَا عَلَى ذَلِكَ، فَقَدْ وَجَبَ الْبَيْعُ، وَإِنْ تَفَرَّقَا بَعْدَ أَنْ تَبَايَعَا وَلَمْ يَتْرُكْ
أَحَدُهُمَا الْبَيْعَ فَقَدْ.وَجَبَ الْبَيْعُ
(رواه البخاري ومسلم)

Terjemahan:
"Dari Ibnu ʿUmar ra., dari Rasulullah saw. bersabda, "Apabila dua orang melakukan jual beli, maka masing-masing dari keduanya mempunyai hak khiar (memilih antara membatalkan atau meneruskan jual beli) selama mereka belum berpisah atau masih bersama; atau jika salah seorang di antara keduanya menentukan khiar kepada yang lainnya. Jika salah seorang menentukan khiar pada yang lain, lalu mereka berjual beli atas dasar itu, maka jadilah jual beli itu. Jika mereka berpisah setelah melakukan jual beli dan masing-masing dari keduanya tidak mengurungkan jual beli, maka jadilah jual beli itu." (Muttafaq A'laih, dan lafaz hadis ini menurut riwayat Muslim) (Shahih & Muslim (1531), 2020)

Hadis riwayat Bukhari & Muslim (2017), tersebut menekankan pentingnya kejujuran, keterbukaan, dan tanggung jawab dalam setiap transaksi jual beli. Prinsip ini menjadi dasar moral dalam dunia *e-commerce* yang kini didominasi oleh interaksi digital tanpa tatap muka, di mana kepercayaan antara penjual dan pembeli harus dijaga melalui transparansi data dan pelayanan yang jujur. Sejalan dengan hal tersebut, penelitian ini menggunakan metode HDBSCAN untuk melakukan segmentasi konsumen *e-commerce* secara objektif dan ilmiah, dengan tujuan memahami perilaku pelanggan berdasarkan pola transaksi yang nyata. Melalui hasil *clustering* yang akurat dan representatif, perusahaan dapat mengembangkan strategi retensi pelanggan yang adil, personal, dan sesuai kebutuhan tiap segmen. Pendekatan ini tidak hanya mendukung efisiensi bisnis, tetapi juga mencerminkan penerapan nilai-nilai etika Islam dalam membangun hubungan dagang yang jujur, saling menguntungkan, dan membawa keberkahan dalam ekosistem digital.

Secara keseluruhan, integrasi nilai-nilai islam dalam penelitian ini menunjukkan bahwa pengembangan teknologi, termasuk *data mining* dan *clustering*, perlu disertai dengan pertimbangan moral dan spiritual. Penerapan metode HDBSCAN dalam segmentasi pelanggan seharusnya diarahkan untuk mencapai kemaslahatan bersama, bukan semata-mata untuk memperoleh keuntungan finansial. Berlandaskan prinsip keadilan, kejujuran, ihsan, serta

tanggung jawab sosial, penelitian ini merepresentasikan penerapan nyata nilai-nilai islam dalam pengelolaan data dan pengambilan keputusan bisnis berbasis teknologi, guna mewujudkan ekosistem ekonomi digital yang beretika, berkeadilan, dan membawa keberkahan.

2.8 Kajian Topik dengan Teori Pendukung

Penelitian ini berfokus pada penerapan metode HDBSCAN sebagai pendekatan utama dalam segmentasi konsumen pada platform *e-commerce*. Dalam beberapa tahun terakhir, pertumbuhan data transaksi digital meningkat secara eksponensial seiring dengan pesatnya perkembangan teknologi dan perubahan pola konsumsi masyarakat. Kondisi ini menuntut adanya metode analisis data yang tidak hanya mampu menangani volume data yang besar, tetapi juga dapat mengungkap pola-pola laten yang sulit ditangkap oleh metode segmentasi konvensional. HDBSCAN hadir sebagai salah satu teknik *clustering* berbasis kepadatan yang semakin relevan dalam konteks *big data*, khususnya untuk data transaksi *e-commerce* yang cenderung tidak berdistribusi normal, memiliki tingkat keragaman tinggi, serta mengandung *noise* dan *outlier* dalam jumlah besar.

Secara teknis, HDBSCAN merupakan penyempurnaan dari DBSCAN dengan memperkenalkan konsep *hierarchical density clustering*. Tidak seperti DBSCAN yang mengandalkan satu nilai *epsilon* untuk seluruh ruang data, HDBSCAN membangun hierarki klaster berdasarkan variasi kepadatan, sehingga dapat mengidentifikasi struktur klaster yang lebih kompleks dan fleksibel. Pendekatan ini memungkinkan pembentukan klaster secara otomatis tanpa menentukan jumlah klaster di awal, menjadikannya jauh lebih adaptif terutama ketika bekerja dengan *data real-world* yang tidak memiliki batas klaster yang jelas. Selain itu, HDBSCAN

mampu memberikan estimasi tingkat keanggotaan (*cluster membership probability*), sehingga hasil segmentasi tidak hanya berupa pembagian kelompok, tetapi juga memberikan ukuran seberapa kuat suatu data menjadi bagian dari klaster tertentu. Hal ini memberikan nilai tambah dalam interpretasi, karena perusahaan dapat melihat tingkat keyakinan model terhadap setiap pengelompokan.

Dalam konteks *e-commerce*, segmentasi konsumen merupakan elemen fundamental dalam memahami perilaku pembelian, preferensi produk, nilai konsumen, serta potensi profitabilitas. Penerapan HDBSCAN memungkinkan pengungkapan kelompok konsumen dengan pola transaksi yang serupa secara lebih mendalam dan akurat dibandingkan metode seperti K-Means, yang mensyaratkan variansi homogen dan sensitivitas terhadap inisialisasi centroid. Pada dataset transaksi nyata, konsumen dapat menunjukkan frekuensi belanja yang tidak merata, rentang nilai pembelian yang sangat bervariasi, serta kecenderungan perilaku yang tidak linear, sehingga pendekatan berbasis kepadatan seperti HDBSCAN menjadi jauh lebih ideal. Dengan kemampuannya menangani *noise*, HDBSCAN juga dapat memisahkan segmen pelanggan yang benar-benar tidak konsisten atau berperilaku tidak umum, sehingga analisis segmen inti dapat lebih fokus dan berkualitas.

Hasil segmentasi menggunakan HDBSCAN membawa implikasi strategis bagi berbagai aspek operasional dan pengambilan keputusan bisnis digital. Misalnya, identifikasi kelompok konsumen dengan pola pembelian tinggi dan konsisten dapat digunakan untuk menyusun program loyalitas yang lebih tepat sasaran. Sementara itu, klaster yang menunjukkan kecenderungan impulsif atau sangat sensitif terhadap diskon dapat menjadi target ideal untuk promosi musiman. Lebih jauh lagi, hasil klaster juga dapat terintegrasi dengan sistem rekomendasi, sehingga konsumen

menerima rekomendasi produk yang lebih relevan dan personal. Strategi personalisasi ini terbukti meningkatkan *conversion rate*, mendorong *engagement* konsumen, serta memperkuat retensi dalam jangka panjang.

Selain manfaat bisnis, penerapan HDBSCAN juga memberikan kontribusi signifikan dalam aspek metodologis penelitian data mining. Fleksibilitas adaptifnya memungkinkan eksplorasi struktur data yang lebih mendalam, sehingga peneliti dapat memahami dinamika pasar digital secara lebih komprehensif. Dengan kemampuannya dalam menangani berbagai bentuk distribusi data, metode ini dapat dengan mudah diperluas ke berbagai jenis analisis selain transaksi, seperti perilaku browsing, durasi interaksi, atau respons terhadap kampanye pemasaran. Dengan demikian, penelitian ini tidak hanya relevan untuk industri *e-commerce*, tetapi juga memberikan gambaran mengenai bagaimana teknik *clustering* modern dapat diterapkan secara luas dalam ekosistem digital yang terus berkembang.

Secara keseluruhan, penggunaan HDBSCAN dalam segmentasi konsumen *e-commerce* menawarkan kombinasi antara ketepatan analisis, fleksibilitas model, dan ketahanan terhadap data berkualitas rendah. Dengan meningkatnya kebutuhan akan strategi pemasaran berbasis data, metode ini menjadi solusi yang menjanjikan bagi perusahaan yang ingin mengoptimalkan pendekatan personalisasi dan memahami karakteristik konsumennya secara lebih mendalam. Penelitian ini diharapkan dapat memberikan kontribusi ilmiah sekaligus praktis, baik bagi akademisi yang mengeksplorasi teknik *clustering* lanjutan, maupun bagi pelaku industri digital yang ingin memperkuat daya saing melalui pemanfaatan data secara lebih strategis.

BAB III

METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini yang digunakan adalah penelitian kuantitatif. Penelitian ini menggunakan pendekatan kuantitatif karena seluruh analisis difokuskan pada pengolahan data numerik yang berkaitan dengan variabel perilaku konsumen serta hasil penerapan algoritma *clustering* HDBSCAN. Data yang digunakan berupa data sekunder konsumen *e-commerce* yang diolah menggunakan metode *clustering* HDBSCAN untuk melakukan segmentasi konsumen.

3.2 Data dan Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diambil dari platform *Kaggle*, dataset *E-commerce Customer Behavior and Sales Analysis-TR* yang dibuat oleh Umut Tuygur (2023-2024). Dataset ini bukan berasal dari perusahaan atau toko online yang sesungguhnya, melainkan dibuat secara buatan menggunakan pola angka yang mendekati kondisi belanja online di Turki pada tahun 2023 hingga 2024, tanpa menyebut merek atau platform tertentu. Dataset ini berisi 5.000 catatan transaksi yang mencakup informasi seperti usia, aktivitas pelanggan saat berbelanja, serta tingkat kepuasan setelah melakukan pembelian. Periode waktu yang dicakup dalam dataset ini adalah Januari 2023 hingga Maret 2024, yang menggambarkan kebiasaan belanja konsumen dari berbagai kalangan usia dan kota besar di Turki. Adapun variabel-variabel yang digunakan dalam penelitian ini disajikan pada Tabel 3.1.

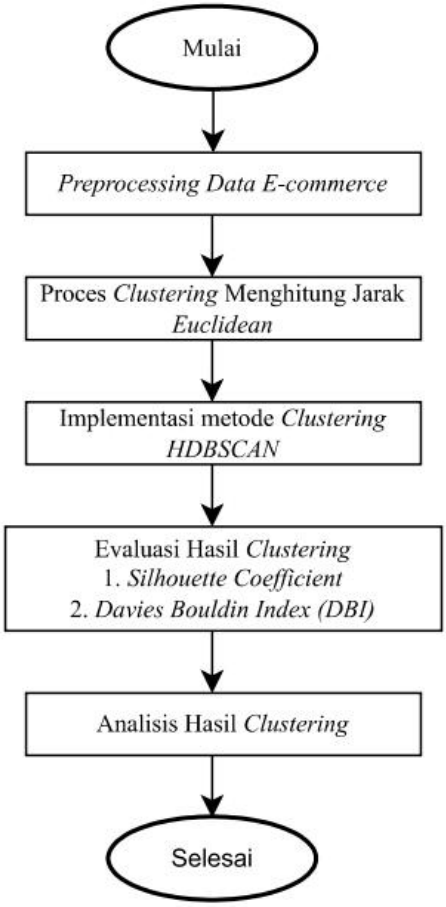
Tabel 3.1 Data *E-commerce*

Nama Variabel	Jenis Data	Keterangan Variabel
Umur	Numerik	Umur pelanggan dalam tahun
Harga perunit	Numerik	Harga perunit barang
Jumlah perunit	Numerik	Jumlah unit dibeli dala transaksi
Potongan Harga	Numerik	Besarnya diskon dalam transaksi
Total Pembayaran	Numerik	Nilai total transaksi setelah diskon
Rata-rata durasi sesi	Numerik	Durasi sesi pengguna di aplikasi atau website
Jumlah tayangan halaman	Numerik	Jumlah halaman yang dibuka
Waktu pengiriman	Numerik	Waktu pengiriman dalam hari
Penilaian Konsumen	Numerik	Penilaian konsumen (1-5)

Dataset ini dimanfaatkan sebagai dasar dalam penerapan metode *clustering* HDBSCAN, di mana setiap variabel digunakan untuk menggambarkan perilaku dan karakteristik konsumen dalam platform *e-commerce*. Variabel harga perunit, jumlah perunit, potongan harga, dan total pembayaran merepresentasikan aspek transaksi serta tingkat pembelian konsumen. Sementara itu, variabel umur, rata-rata durasi sesi pengguna aplikasi, dan jumlah tayangan halaman menggambarkan perilaku kunjungan dan tingkat keterlibatan konsumen selama berada di platform. Selain itu, variabel waktu pengiriman dan penilaian konsumen mencerminkan pengalaman pembelian serta tingkat kepuasan pelanggan, yang turut berpengaruh terhadap segmentasi konsumen secara keseluruhan.

3.3 Teknik Analisis Data

Berikut merupakan rangkaian beberapa tahapan teknik analisis data untuk Implementasi Metode *Clustering* HDBSCAN Dalam Segmentasi Konsumen *E-commerce* Berdasarkan Perilaku Belanja. Tahapan teknik analisis dapat dilihat pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

Penjelasan langkah-langkah terkait teknik analisis data sebagai berikut:

1. *Preprocessing Data*

   Langkah ini bertujuan untuk mempersiapkan data agar siap digunakan dalam proses *clustering*.

   a. *MinMax Scaler* (Normalisasi data)

      Menormalkan data agar setiap fitur berada pada skala yang sama sehinggan tidak ada variabel yang mendominasi proses pembentukan klaster menggunakan Persamaan (2.1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. *Clustering*

   Sebelum proses *clustering* menggunakan algoritma HDBSCAN dapat dilakukan, dibutuhkan perhitungan jarak *Euclidean* terlebih dahulu antar setiap titik data menggunakan Persamaan (2.2).

$$d(p_i, p_j) = \sqrt{\sum_{l=1}^{z} (p_{il} - p_{jl})^2}$$

3. Proses *Clustering* HDBSCAN

   Metode HDBSCAN digunakan untuk mengelompokkan konsumen berdasarkan kemiripan perilaku pembelian. Langkah-langkah meliputi:

   a. Menghitung *core distance*.

      Tahap awal dilakukan dengan menghitung *core distance* untuk menentukan tingkat kepadatan lokal di sekitar setiap titik data. Nilai ini menunjukkan jarak antara suatu titik dengan tetangga ke-k terdekatnya. Jika melihat Persamaan (2.3), perhitungan ini dijelaskan secara matematis sebagai dasar identifikasi densitas pada setiap area data.

$$core_k(p_i) = d(p_i, tetangga\ terdekat\ ke-k\ dari\ titik\ \ p_i)$$

   b. Menghitung *mutual reachability distance*.

      Setelah *core distance* diperoleh, tahap berikutnya menghitung *mutual reachability distance*, yaitu jarak efektif antar dua titik yang mempertimbangkan kepadatan lokal keduanya. Nilai ini digunakan untuk menstabilkan hubungan antar titik terhadap variasi kepadatan data. Adapun bentuk matematisnya dapat dilihat pada Persamaan (2.6), yang menjelaskan fungsi jarak gabungan antara kepadatan dan jarak *Euclidean.*

$$mreach_k(p_i, p_j) = \max\{core_k(p_i),\ core_k(p_j),\ d(p_i, p_j)\}$$

c. Membangun *minimum spanning tree*.

Selanjutnya dibangun *Minimum Spanning Tree* (MST) dari graf berbobot berdasarkan nilai *mutual reachability distance*. MST menghubungkan seluruh titik dengan total bobot minimum tanpa membentuk siklus. Proses ini membantu membangun struktur hubungan hierarkis antar data, yang secara matematis dijelaskan dalam Persamaan (2.8) menggunakan pendekatan algoritma prim.

$$w(p_i, p_j) = mreach_k(p_i, p_j).$$

Langkah-langkah membentuk MST menggunakan Algoritma prim:

Pertama pilih satu simpul awal $v \in V$, yang kedua Masukkan v ke dalam himpunan simpul MST (T), yang ketiga Selama masih terdapat simpul MST (T): Cari sisi (u, v) dengan bobot minimum, dimana $u \in T$ dan $v \notin T$. Tambahkan sisi (u, v) ke MST, yang keempat ulangi langkah 3 hingga seluruh simpul tergabung dalam T, dan yang terakhir MST terbentuk.

Jika langkah disesuikan denga penelitian ini, pertama pilih titik awal p_1, kedua MST = $\{p_1\}$ masukkan p_1 ke dalam himpunan simpul MST, ketiga cari sisi berbobot minimum yang menghubungkan titik dalam MST dengan titik yang belum tergabung dalam MST berdasarkan nilai Mutual Reachbility Distance, keempat, sisi dengan bobot minimum tersebut beserta titik yang terhubung ditambahkan ke dalam MST. Kelima, langkah ketiga dan keempat diulangi hingga seluruh titik data tergabung ke dalam MST. Hasil akhir berupa MST yang terdiri dari $n -$

1 sisi, dengan n menyatakan jumlah titik data, sehingga seluruh titik data terhubung dengan total bobot minimum tanpa membentuk siklus.

d. Membentuk hierarki klaster.

Setelah MST terbentuk, dilakukan proses pembentukan hierarki klaster dengan memutus sisi-sisi tertentu berdasarkan bobot tertinggi. Langkah ini menghasilkan struktur hierarki yang menggambarkan proses penggabungan dan pemisahan klaster pada berbagai tingkat kepadatan. Penjelasan teoretis serta persamaan yang berkaitan dapat dilihat pada Persamaan (2.11).

$$\lambda = \frac{1}{w}$$

e. Menyederhanakan hierarki.

Tahapan berikutnya adalah penyederhanaan hierarki atau *condensed cluster tree*, yang dilakukan dengan menerapkan parameter *minimum cluster size*. Langkah ini bertujuan untuk menghapus klaster berukuran kecil yang dianggap tidak stabil atau termasuk *noise*. Istilah "tidak stabil" dalam HDBSCAN merujuk pada *cluster* yang memiliki rentang keberadaan yang sangat singkat dalam perubahan tingkat kepadatan (λ). Stabilitas *cluster* diukur berdasarkan lamanya *cluster* tersebut bertahan sejak terbentuk hingga terpecah atau menghilang. Pembahasan mendalam mengenai mekanisme penyederhanaan ini dapat dilihat pada Persamaan (2.15) dan (2.16).

$$C_i \text{ valid} \iff N(C_i) \geq m_{\text{clsize}}$$

Sebaliknya, klaster dieliminasi apabila:

$$C_i \text{ tidak valid} \iff N(C_i) < m_{\text{clsize}}$$

f. Mengekstraksi klaster berdasarkan stabilitas.

Tahap terakhir adalah mengekstraksi klaster hasil akhir berdasarkan nilai stabilitasnya (*cluster stability*). Nilai ini menunjukkan konsistensi klaster terhadap perubahan kepadatan dan digunakan untuk memilih klaster yang paling representatif. Persamaan yang digunakan untuk mengukur stabilitas klaster dijelaskan secara rinci pada Persamaan (2.19).

$$S(C_i) = \sum_{p_j \in C_i} (\lambda_{\max}(p_j, C_i) - \lambda_{\min}(C_i))$$

g. Pemilihan Hasil Akhir *Cluster*

Setelah proses Implementasi *Clustering* HDBSCAN selesai dilakukan, langkah selanjutnya adalah memilih klaster terbaik dari hasil pengelompokan yang terbentuk menggunakan Persamaan (2.23)

$$C_{\text{final}} = \underset{C_i}{\text{argmax}}\ S(C_i)$$

4. Evaluasi klaster

Evaluasi hasil klaster pada penelitian ini dilakukan menggunakan *Silhouette Coefficient* dan *Davies-Bouldin Index* (DBI), dua metrik yang mampu menilai kualitas pengelompokan tanpa memerlukan label eksternal, sehingga sesuai untuk metode *unsupervised learning* seperti HDBSCAN.

a. *Silhouette Coefficient* mengukur kesesuaian objek terhadap klasternya, dengan nilai mendekati 1 menunjukkan pemisahan yang jelas dan kohesi internal yang kuat. *Silhouette Coefficient* dapat dilihat di Persamaan (2.24).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

b. DBI menilai tingkat kesamaan antar klaster, di mana nilai rendah menunjukkan perbedaan signifikan antar klaster dan kepadatan internal yang baik. *Davies Bouldin Index* dapat dilihat di Persamaan (2.25).

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i$$

Kombinasi kedua metrik ini memungkinkan evaluasi seimbang antara pemisahan antar klaster (*inter-cluster separation*) dan kepadatan dalam klaster (*intra-cluster compactness*), sehingga segmentasi pelanggan yang dihasilkan lebih akurat dan relevan untuk analisis perilaku konsumen di *e-commerce.*

Dataset dievaluasi menggunakan dua metrik evaluasi klaster, yaitu *Silhouette Coefficient* dan *Davies Bouldin Index*. Evaluasi dilakukan pada algoritma HDBSCAN dengan memvariasikan dua parameter utama secara bersamaan, yaitu *min_sample* dan *min_cluster_size*, dengan nilai yang sama keduanya. Variasi nilai yang diuji adalah 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, dan 55, total sebelas kombinasi parameter. Setiap kombinasi parameter menghasilkan satu nilai *Silhouette Coefficient* dan satu nilai *Davies Bouldin Index*. Nilai *Silhouette Coefficient* yang lebih tinggi menunjukkan klaster yang lebih padat dan terpisah dengan baik, sementara nilai *Davies Bouldin Index* yang lebih rendah menunjukkan klaster yang lebih kompak dan jauh satu sama lain. Konfigurasi nilai parameter optimal HDBSCAN dibandingkan langsung dengan hasil evaluasi algoritma DBSCAN menggunakan metrik yang sama, yaitu *Silhouette Coefficient*

dan *Davies Bouldin Index*. Perbandingan ini bertujuan untuk menentukan algoritma mana yang mengasilkan struktur klaster yang lebih baik .

BAB IV

HASIL DAN PEMBAHASAN

4.1 *Preprocessing* Data *E-commerce*

Pada tahap ini, dilakukan proses normalisasi terhadap seluruh variabel numerik dalam dataset agar memiliki skala nilai yang seragam sebelum memasuki tahap *clustering*. Normalisasi dilakukan karena setiap variabel dalam penelitian memiliki rentang nilai yang berbeda. Perbedaan rentang tersebut dapat memengaruhi proses perhitungan jarak, karena variabel dengan nilai yang lebih besar berpotensi memberikan pengaruh lebih dominan dibandingkan variabel lainnya. Oleh karena itu, data perlu dinormalisasikan agar seluruh variabel berada pada skala yang sebanding sebelum dilakukan proses *clustering*. Dalam penelitian ini digunakan metode *Min-Max Scaler*, yaitu proses penskalaan nilai data ke dalam rentang 0 hingga 1 berdasarkan nilai minimum dan maksimum dari masing-masing variabel. Normalisasi dilakukan agar variabel dengan nilai besar tidak terlalu memengaruhi hasil *clustering*. Setelah dinormalisasi, setiap variabel memiliki skala yang sama, sehingga proses pengelompokan data menjadi lebih adil dan seimbang.

1. Variabel Umur

   Nilai maksimum = 75

   Nilai minimum = 18

   Nilai data = 27

$$x' = \frac{27 - 18}{75 - 18} = 0{,}1579$$

2. Variabel Harga Perunit

   Nilai maksimum = 715945

Nilai minimum = 131

Nilai data = 5428

$$x' = \frac{5428 - 131}{715945 - 131} = 0{,}0074$$

3. Variabel Jumlah Perunit

Nilai maksimum = 5

Nilai minimum = 1

Nilai data = 1

$$x' = \frac{1 - 1}{5 - 1} = 0$$

4. Variabel Potongan Harga

Nilai maksimum = 152555

Nilai minimum = 0

Nilai data = 0

$$x' = \frac{0 - 0}{152555 - 0} = 0$$

5. Variabel Total Pembayaran

Nilai maksimum = 2147835

Nilai minimum = 127

Nilai data = 5428

$$x' = \frac{5428 - 127}{2147835 - 127} = 0{,}0025$$

6. Variabel Rata-rata durasi sesi

Nilai maksimum = 73

Nilai minimum = 1

Nilai data = 4

$$x' = \frac{4-1}{73-1} = 0{,}0417$$

7. Variabel Jumlah Tayangan Halaman

Nilai maksimum = 24

Nilai minimum = 1

Nilai data = 14

$$x' = \frac{14-1}{24-1} = 0{,}5652$$

8. Variabel Waktu Pengiriman

Nilai maksimum = 25

Nilai minimum = 1

Nilai data = 8

$$x' = \frac{8-1}{25-1} = 0{,}2917$$

9. Variabel Penilaian Konsumen

Nilai maksimum = 5

Nilai minimum = 1

Nilai data = 5

$$x' = \frac{5-1}{5-1} = 1$$

Setelah proses perhitungan manual selesai dilakukan, tahap normalisasi kemudian diterapkan pada seluruh data dalam dataset *e-commerce*. Hasil normalisasi dapat dilihat di tabel 4.1 yang menunjukkan data pertama setelah melalui proses *Min Max Scaler.*

Tabel 4.1 Normalisasi *Min Max Saler*

Titik	p_1	p_2	p_3	p_4	p_5
Umur	0,1579	0,4211	0,4386	0,2456	0,3860
Harga Perunit	0,0074	0,0032	0,0065	0,1121	0,1054
Jumlah Perunit	0,0000	0,0000	1,0000	0,0000	1,0000
Potongan Harga	0,0000	0,0000	0,0000	0,1503	0,0000
Total Pembayaran	0,0025	0,0011	0,0112	0,0267	0,1759
Rata-rata Durasi Sesi	0,0417	0,1389	0,0833	0,0972	0,2778
Jumlah Tayangan Halaman	0,5652	0,0870	0,3043	0,3913	0,3913
Waktu Pengiriman	0,2917	0,0833	0,1667	0,0000	0,2500
Penilaian Konsumen	1,0000	0,5000	0,2500	0,7500	0,7500

Berdasarkan Tabel 4.1 menampilkan hasil normalisasi *Min-Max Scaler* dari beberapa variabel terhadap titik data (p_1 sampai p_5). Normalisasi *Min-Max Scaler* mengubah nilai asli ke rentang 0 sampai 1, dimana $0 =$ nilai minimum dan $1 =$ nilai maksimum dalam dataset. Setiap variabel menunjukkan perbedaan karakteristik pada titik data p_1 sampai p_5. Pada variabel umur, nilai tertinggi terdapat pada p_3 sebesar $0{,}4386$, yang menunjukkan bahwa data tersebut memiliki nilai umur paling besar dibandingkan titik lainnya setelah dinormalisasi. Pada variabel harga perunit, nilai tertinggi terdapat pada p_4 sebesar $0{,}1121$, sedangkan nilai terendah berada pada p_2 sebesar $0{,}0032$. Hal ini menunjukkan bahwa p_4 memiliki harga perunit yang lebih tinggi dibandingkan data lainnya. Pada variabel jumlah perunit, titik p_3 dan p_5 memiliki nilai $1{,}0000$ sehingga menunjukkan jumlah pembelian tertinggi, sedangkan titik lainnya memiliki nilai rendah. Untuk variabel potongan harga,

hanya titik p_4 yang memiliki nilai $0{,}1503$, sementara titik lainnya bernilai $0{,}0000$, yang berarti hanya p_4 yang memperoleh potongan harga. Pada variabel total pembayaran, nilai tertinggi terdapat pada p_5 sebesar $0{,}1759$ yang menunjukkan total pembayaran paling besar dibandingkan titik lainnya. Variabel durasi sesi juga menunjukkan nilai tertinggi pada p_5 sebesar $0{,}2778$, sehingga dapat diartikan bahwa pelanggan pada titik tersebut memiliki durasi aktivitas paling lama. Pada variabel jumlah tayangan halaman, nilai tertinggi terdapat pada p_1 sebesar $0{,}5652$, yang menandahakan aktivitas penelusuran halaman paling banyak. Untuk variabel waktu pengiriman, nilai tertinggi berada pada p_1 sebesar $0{,}2917$, sedangkan p_4 memiliki nilai $0{,}0000$ yang menunjukkan waktu pengiriman paling rendah. Sementara itu, pada variabel penilaian konsumen, titik p_1, memperoleh nilai tertinggi sebesar $1{,}0000$ sehingga menunjukkan tingkat penilaian konsumen paling baik dibandingkan titik data lainnya. Dengan proses normalisasi menggunakan *Min-Max Scaler*, seluruh variabel berhasil disetarakan ke dalam rentang nilai yang sama sehingga data menjadi lebih seimbang untuk digunakan proses *clustering* menggunakan metode HDBSCAN.

4.2 Simulasi *Clustering*

Sebelum proses *clustering* menggunakan algoritma HDBSCAN dilakukan, terlebih dahulu dihitung jarak antar titik data menggunakan jarak *Euclidean*. Perhitungan jarak ini diperlukan karena HDBSCAN membutuhkan informasi kedekatan antar data sebagai dasar dalam mengestimasi kepadatan lokal di sekitar setiap titik data. Jarak *Euclidean* dipilih karena merupakan metode pengukuran

jarak yang umum digunakan dalam ruang berdimensi banyak dan mampu merepresentasikan kedekatan antar data secara geometris.

Perhitungan manual pada penelitian ini menggunakan lima data pertama $(p_1 - p_5)$ sebagai contoh ilustrasi. Sebelum dilakukan perhitungan, data dinormalisasi menggunakan metode *Min-Max* agar seluruh fitur berada pada rentang nilai yang sama, yaitu 0 hingga 1. Hal ini dilakukan karena HDBSCAN menggunakan perhitungan jarak, sehingga perbedaan skala antar fitur dapat memengaruhi hasil *clustering*. Dengan menggunakan lima data pertama, tahapan perhitungan dapat dijelaskan lebih sederhana dan mudah ditelusuri.

Tabel 4.2 Data setelah di normalisasi

Titik	p_1	p_2	p_3	p_4	p_5
Umur	0,1579	0,4211	0,4386	0,2456	0,3860
Harga Perunit	0,0074	0,0032	0,0065	0,1121	0,1054
Jumlah Perunit	0,0000	0,0000	1,0000	0,0000	1,0000
Potongan Harga	0,0000	0,0000	0,0000	0,1503	0,0000
Total Pembayaran	0,0025	0,0011	0,0112	0,0267	0,1759
Rata-rata Durasi Sesi	0,0417	0,1389	0,0833	0,0972	0,2778
Jumlah Tayangan Halaman	0,5652	0,0870	0,3043	0,3913	0,3913
Waktu Pengiriman	0,2917	0,0833	0,1667	0,0000	0,2500
Penilaian Konsumen	1,0000	0,5000	0,2500	0,7500	0,7500

1. Menghitung Jarak *Euclidean*

   Perhitungan jarak *Euclidean* dilakukan terhadap 5 data pertama yang telah melalui proses normalisasi. Penggunaan data yang telah dinormalisasi bertujuan untuk menyamakan skala antar variabel, sehingga tidak terdapat atribut yang mendominasi dalam perhitungan jarak. Hal ini penting karena

setiap variabel memiliki rentang nilai yang berbeda, sehingga normalisasi diperlukan agar perhitungan jarak menjadi lebih proporsional dan akurat.

Jarak *Euclidean* digunakan untuk mengukur tingkat kedekatan antar titik data dalam ruang multidimensi. Perhitungan dilakukan dengan menghitung selisih setiap atribut antar dua titik data, kemudian dikuadratkan, dijumlahkan, dan diakarkan sesuai dengan rumus jarak *Euclidean*. Dalam penelitian ini, perhitungan jarak dilakukan secara manual pada 5 data pertama sebagai ilustrasi untuk mempermudah pemahaman terhadap proses pembentukan matriks jarak yang digunakan dalam tahapan *clustering* HDBSCAN.

Jarak *Euclidean* antara dua titik dihitung menggunakan Persamaan (2.2)

$$d(p_i, p_j) = \sqrt{\begin{matrix}[(p_{i1} - p_{j1})^2 + (p_{i2} - p_{j2})^2 + (p_{i3} - p_{j3})^2 \\ +(p_{i4} - p_{j4})^2 + (p_{i5} - p_{j5})^2 + (p_{i6} - p_{j6})^2 \\ +(p_{i7} - p_{j7})^2 + (p_{i8} - p_{j8})^2 + (p_{i9} - p_{j9})^2]\end{matrix}}$$

Berikut subsitusi nilai data $p_1 - p_5$ langsung ke dalam rumus untuk semua 10 pasang, maka p_i adalah p_1 dan p_j adalah p_2.

a. Jarak $d\ (p_1, p_2)$

Subsitusi nilai p_1 dan p_2 ke dalam rumus:

$$d(p_1, p_2) = \sqrt{\begin{matrix}[(0{,}1579 - 0{,}4211)^2 + (0{,}0074 - 0{,}0032)^2 + (0{,}0000 - 0{,}0000)^2 \\ +(0{,}0000 - 0{,}0000)^2 + (0{,}0025 - 0{,}0011)^2 + (0{,}0417 - 0{,}1389)^2 \\ +(0{,}5652 - 0{,}0870)^2 + (0{,}2917 - 0{,}0833)^2 + (1{,}0000 - 0{,}5000)^2]\end{matrix}}$$

Hitung setiap selisih $(p_i - p_j)^2$:

$$d(p_1, p_2) = \sqrt{\begin{matrix}[(-0{,}2632)^2 + (0{,}0042)^2 + (0{,}0000)^2 \\ +(0{,}0000)^2 + (0{,}0014)^2 + (-0{,}0972)^2 \\ +(0{,}4782)^2 + (0{,}2084)^2 + (0{,}5000)^2]\end{matrix}}$$

Kuadratkan setiap selisih $(p_i - p_j)^2$:

$$d(p_1, p_2) = \sqrt{\begin{matrix}[(0{,}0692742 + 0{,}0000176 + 0{,}0000000 \\ +0{,}0000000 + 0{,}0000020 + 0{,}0094478 \\ +0{,}2286752 + 0{,}0434306 + 0{,}2500000)]\end{matrix}}$$

Jumlah semua kuadrat

$d(p_1, p_2) = \sqrt{0{,}6008474}$

Lalu kuadratkan:

$$d(p_1, p_2) = 0{,}7751435$$

b. Jarak $d\ (p_1, p_3)$

Substitusi nilai p_1 dan p_3 ke dalam rumus:

$$d(p_1, p_3) = \sqrt{\begin{matrix}[(0{,}1579 - 0{,}4386)^2 + (0{,}0074 - 0{,}0065)^2 + (0{,}0000 - 1{,}0000)^2 \\ +(0{,}0000 - 0{,}0000)^2 + (0{,}0025 - 0{,}0112)^2 + (0{,}0417 - 0{,}0833)^2 \\ +(0{,}5652 - 0{,}3043)^2 + (0{,}2917 - 0{,}1667)^2 + (1{,}0000 - 0{,}2500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_1, p_3) = \sqrt{\begin{matrix}[(-0{,}2807)^2 + (0{,}0009)^2 + (-1{,}0000)^2 \\ +(0{,}0000)^2 + (-0{,}0087)^2 + (-0{,}0416)^2 \\ +(0{,}2609)^2 + (0{,}1250)^2 + (0{,}7500)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_1, p_3) = \sqrt{\begin{matrix}[(0{,}0787925 + 0{,}0000008 + 1{,}0000000 \\ +0{,}0000000 + 0{,}0000756 + 0{,}0017305 \\ +0{,}0680688 + 0{,}0156250 + 0{,}5625000)]\end{matrix}}$$

Jumlah semua kuadrat

$d(p_1, p_3) = \sqrt{1{,}7267932}$

Lalu akar kuadratkan:

$$d(p_1, p_3) = 1{,}3140750$$

c. Jarak $d\ (p_1, p_4)$

Substitusi nilai p_1 dan p_4 :

$$d(p_1, p_4) = \sqrt{\begin{matrix}[(0{,}1579 - 0{,}2456)^2 + (0{,}0074 - 0{,}1121)^2 + (0{,}0000 - 0{,}0000)^2 \\ +(0{,}0000 - 0{,}1503)^2 + (0{,}0025 - 0{,}0267)^2 + (0{,}0417 - 0{,}0972)^2 \\ +(0{,}5652 - 0{,}3913)^2 + (0{,}2917 - 0{,}0000)^2 + (1{,}0000 - 0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_1, p_4) = \sqrt{\begin{matrix}[(-0{,}0877)^2 + (-0{,}1047)^2 + (0{,}0000)^2 \\ +(-0{,}1503)^2 + (-0{,}0242)^2 + (-0{,}0555)^2 \\ +(0{,}1739)^2 + (0{,}2917)^2 + (0{,}2500)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_1, p_4) = \sqrt{\begin{matrix}[(0{,}0076913 + 0{,}0109621 + 0{,}0000000 \\ +0{,}0225901 + 0{,}0005856 + 0{,}0030803 \\ +0{,}0302412 + 0{,}0850889 + 0{,}0625000)]\end{matrix}}$$

Jumlah semua kuadrat

$$d(p_1, p_4) = \sqrt{0{,}2227395}$$

Lalu akar kuadratkan:

$$d(p_1, p_4) = 0{,}4719529$$

d. Jarak $d\ (p_1, p_5)$

Substitusi nilai p_1 dan p_5:

$$d(p_1, p_5) = \sqrt{\begin{matrix}[(0{,}1579 - 0{,}3860)^2 + (0{,}0074 - 0{,}1054)^2 + (0{,}0000 - 1{,}0000)^2 \\ +(0{,}0000 - 0{,}0000)^2 + (0{,}0025 - 0{,}1759)^2 + (0{,}0417 - 0{,}2778)^2 \\ +(0{,}5652 - 0{,}3913)^2 + (0{,}2917 - 0{,}2500)^2 + (1{,}0000 + 0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_1, p_5) = \sqrt{\begin{matrix}[(-0{,}2281)^2 + (-0{,}0980)^2 + (-1{,}0000)^2 \\ +(0{,}0000)^2 + (-0{,}1734)^2 + (-0{,}2361)^2 \\ +(0{,}1739)^2 + (0{,}0417)^2 + (0{,}2500)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_1, p_5) = \sqrt{\begin{matrix}[(0{,}0520296 + 0{,}0096040 + 1{,}0000000 \\ +0{,}0000000 + 0{,}0300676 + 0{,}0557432 \\ +0{,}0302412 + 0{,}0017389 + 0{,}0625000)]\end{matrix}}$$

Jumlah semua kuadrat

$d(p_1, p_5) = \sqrt{1{,}2419245}$

Lalu akar kuadrat

$$d(p_1, p_5) = 1{,}1144167$$

e. Jarak $d\ (p_2, p_3)$

Substitusi nilai p_2 dan p_3:

$$d(p_2, p_3) = \sqrt{\begin{matrix}[(0{,}4211 - 0{,}4386)^2 + (0{,}0032 - 0{,}0065)^2 + (0{,}0000 - 1{,}0000)^2 \\ +(0{,}0000 - 0{,}0000)^2 + (0{,}0011 - 0{,}0112)^2 + (0{,}1389 - 0{,}0833)^2 \\ +(0{,}0870 - 0{,}3043)^2 + (0{,}0833 - 0{,}1667)^2 + (0{,}5000 - 0{,}2500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_2, p_3) = \sqrt{\begin{matrix}[(-0{,}0175)^2 + (-0{,}0033)^2 + (-1{,}0000)^2 \\ +(0{,}0000)^2 + (-0{,}0101)^2 + (0{,}0556)^2 \\ +(-0{,}2173)^2 + (-0{,}0834)^2 + (0{,}2500)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_2, p_3) = \sqrt{\begin{matrix}[(0{,}0003063 + 0{,}0000108 + 1{,}0000000 \\ +0{,}0000000 + 0{,}0001020 + 0{,}0030913 \\ +0{,}0472192 + 0{,}0069555 + 0{,}0625000)]\end{matrix}}$$

Jumlah semua kuadrat

$d(p_2, p_3) = \sqrt{1{,}1201851}$

Lalu akar kuadrat

$$d(p_2, p_3) = 1{,}0583879$$

f. Jarak $d\ (p_2, p_4)$

Substitusi nilai p_2 dan p_4:

$$d(p_2, p_4) = \sqrt{\begin{matrix}[(0{,}4211 - 0{,}2456)^2 + (0{,}0032 - 0{,}1121)^2 + (0{,}0000 - 0{,}0000)^2 \\ +(0{,}0000 - 0{,}1503)^2 + (0{,}0011 - 0{,}0267)^2 + (0{,}1389 - 0{,}0972)^2 \\ +(0{,}0870 - 0{,}3913)^2 + (0{,}0833 - 0{,}0000)^2 + (0{,}5000 - 0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_2, p_4) = \sqrt{\begin{matrix}[(0{,}1755)^2 + (-0{,}1089)^2 + (0{,}0000)^2 \\ +(-0{,}1503)^2 + (-0{,}0256)^2 + (0{,}0417)^2 \\ +(-0{,}3043)^2 + (0{,}0833)^2 + (-0{,}2500)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_2, p_4) = \sqrt{\begin{matrix}[(0{,}0308003 + 0{,}0118592 + 0{,}0000000 \\ +0{,}0225901 + 0{,}0006554 + 0{,}0017389 \\ +0{,}0925985 + 0{,}0069389 + 0{,}062500)]\end{matrix}}$$

Jumlah semua kuadrat

$$d(p_2, p_4) = \sqrt{0{,}2296813}$$

Lalu akar kuadrat

$$d(p_2, p_4) = 0{,}4792508$$

g. Jarak $d\ (p_2, p_5)$

Substitusi nilai p_2 dan p_5:

$$d(p_2, p_5) = \sqrt{\begin{matrix}[(0{,}4211 - 0{,}3860)^2 + (0{,}0032 - 0{,}1054)^2 + (0{,}0000 - 1{,}0000)^2 \\ +(0.0000 - 0{,}0000)^2 + (0{,}0011 - 0{,}1759)^2 + (0{,}1389 - 0{,}2778)^2 \\ +(0{,}0870 - 0{,}3913)^2 + (0{,}0833 - 0{,}2500)^2 + (0{,}5000 - 0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_2, p_5) = \sqrt{\begin{matrix}[(0{,}0351)^2 + (-0{,}1022)^2 + (-1{,}0000)^2 \\ +(0{,}0000)^2 + (-0{,}1748)^2 + (-0{,}1389)^2 \\ +(-0{,}3043)^2 + (-0{,}1667)^2 + (-0{,}2500)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_2, p_5) = \sqrt{\begin{matrix}[(0{,}0012320 + 0{,}0104448 + 1{,}0000000 \\ +0{,}0000000 + 0{,}0305550 + 0{,}0192932 \\ +0{,}0925985 + 0{,}0277889 + 0{,}0625000)]\end{matrix}}$$

Jumlah semua kuadrat

$$d(p_2, p_5) = \sqrt{1{,}2444124}$$

Lalu akar kuadrat

$$d(p_2, p_5) = 1{,}1155323$$

h. Jarak $d\ (p_3, p_4)$

Substitusi nilai p_3 dan p_4:

$$d(p_3,p_4) = \sqrt{\begin{matrix}[(0{,}4386-0{,}2456)^2+(0{,}0065-0{,}1121)^2+(1{,}0000-0{,}0000)^2 \\ +(0{,}0000-0{,}1503)^2+(0{,}0112-0{,}0267)^2+(0{,}0833-0{,}0972)^2 \\ +(0{,}3043-0{,}3913)^2+(0{,}1667-0{,}0000)^2+(0{,}2500-0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_3,p_4) = \sqrt{\begin{matrix}[(0.1930)^2+(-0.1056)^2+(1.0000)^2 \\ +(-0.1503)^2+(-0.0155)^2+(-0.0139)^2 \\ +(-0.0870)^2+(0.1667)^2+(-0.5000)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_3,p_4) = \sqrt{\begin{matrix}[(0{,}0372490+0{,}0111514+1{,}0000000 \\ +0{,}0225901+0{,}0002403+0{,}0001932 \\ +0{,}0075690+0{,}0277889+0{,}2500000)]\end{matrix}}$$

Jumlah semua kuadrat

$$d(p_3,p_4) = \sqrt{1{,}3567819}$$

Lalu akar kuadrat

$$d(p_3,p_4) = 1{,}1648098$$

i. Jarak $d\ (p_3, p_5)$

Substitusi nilai p_3 dan p_5:

$$d(p_3,p_5) = \sqrt{\begin{matrix}[(0{,}4386-0{,}3860)^2+(0{,}0065-0{,}1054)^2+(1{,}0000-1{,}0000)^2 \\ +(0{,}0000-0{,}0000)^2+(0{,}0112-0{,}1759)^2+(0{,}0833-0{,}2778)^2 \\ +(0{,}3043-0{,}3913)^2+(0{,}1667-0{,}2500)^2+(0{,}2500-0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_3,p_5) = \sqrt{\begin{matrix}[(0{,}0526)^2+(-0{,}0989)^2+(0{,}0000)^2 \\ +(0{,}0000)^2+(-0{,}1647)^2+(-0{,}1945)^2 \\ +(-0{,}0870)^2+(-0{,}0833)^2+(-0{,}5000)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_3, p_5) = \sqrt{\begin{matrix}[(0{,}0027668 + 0{,}0097812 + 0{,}0000000 \\ +0{,}0000000 + 0{,}0271261 + 0{,}0378303 \\ +0{,}0075690 + 0{,}0069389 + 0{,}2500000)]\end{matrix}}$$

Jumlah semua kuadrat

$$d(p_3, p_5) = \sqrt{0{,}3420123}$$

Lalu akar kuadrat

$$d(p_3, p_5) = 0{,}5848182$$

j. Jarak $d\ (p_4, p_5)$

Substitusi nilai $p4$ dan $p5$:

$$d(p_4, p_5) = \sqrt{\begin{matrix}[(0{,}2456 - 0{,}3860)^2 + (0{,}1121 - 0{,}1054)^2 + (0{,}0000 - 1{,}0000)^2 \\ +(0{,}1503 - 0{,}0000)^2 + (0{,}0267 - 0{,}1759)^2 + (0{,}0972 - 0{,}2778)^2 \\ +(0{,}3913 - 0{,}3913)^2 + (0{,}0000 - 0{,}2500)^2 + (0{,}7500 - 0{,}7500)^2]\end{matrix}}$$

Hitung setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_4, p_5) = \sqrt{\begin{matrix}[(-0{,}1404)^2 + (0{,}0067)^2 + (-1{,}0000)^2 \\ +(0{,}1503)^2 + (-0{,}1492)^2 + (-0{,}1806)^2 \\ +(0{,}0000)^2 + (-0{,}2500)^2 + (0{,}0000)^2]\end{matrix}}$$

Kuadratkan setiap selisih $\left(p_i - p_j\right)^2$:

$$d(p_4, p_5) = \sqrt{\begin{matrix}[(0{,}0197122 + 0{,}0000449 + 1{,}0000000 \\ +0{,}0225901 + 0{,}0222606 + 0{,}0326164 \\ +0{,}0000000 + 0{,}0625000 + 0{,}0000000)]\end{matrix}}$$

Jumlah semua kuadrat

$$d(p_4, p_5) = \sqrt{1{,}1597242}$$

Lalu akar kuadrat

$$d(p_4, p_5) = 1{,}0769049$$

Tabel 4.3 Rekapitulasi Jarak *Euclidean*

	p_1	p_2	p_3	p_4	p_5
p_1	0	0,7751435	1,3140750	0,4719529	1,1144167
p_2	0,7751435	0	1,0583879	0,4792508	1,1155323
p_3	1,3140750	1,0583879	0	1,1648098	0,5848182
p_4	0,4719529	0,4792508	1,1648098	0	1,0769049
p_5	1,1144167	1,1155323	0,5848182	1,0769049	0

Berdasarkan Tabel 4.3, rekapitulasi jarak *Euclidean* menunjukkan tingkat kedekatan atau kemiripan antar titik data p_1 hingga p_5. Perhitungan jarak *Euclidean* dilakukan setelah proses normalisasi data menggunakan *Min-Max Scaler* sehingga seluruh variabel berada pada skala yang sama. Nilai jarak yang semakin kecil menunjukkan bahwa dua titik data memiliki karakteristik yang semakin mirip, sedangkan nilai jarak yang semakin besar menunjukkan perbedaan karakteristik yang lebih tinggi antar data.

Berdasarkan hasil perhitungan, jarak terkecil terdapat antara titik p_1 dan p_4 sebesar $0{,}4719529$. Hal tersebut menunjukkan bahwa kedua titik data memiliki tingkat kemiripan yang paling tinggi dibandingkan pasangan data lainnya. Selain itu, jarak relatif kecil juga terlihat antara titi p_3 dan p_5 sebesar $0{,}5848182$, yang menunjukkan adanya kedekatan karakteristik antara kedua data tersebut. Sementara itu, jarak terbesar terdapat antara titik p_3 dan p_4 sebesar $1{,}1648098$, yang menunjukkan bahwa kedua titik data memiliki perbedaan karakteristik yang cukup signifikan. Nilai jarak lainnya, seperti antara titik p_1 dan p_3 sebesar $1{,}3140750$ serta antara titik p_2 dan p_5 sebesar $1{,}1155323$, juga menunjukka tingkat perbedaan data yang relatif tinggi.

Secara keseluruhan, matrik jarak *Euclidean* pada tabel 4.3 digunakan untuk menggambarkan hubungan kedekatan antar data sebelum dilakukan proses *clustering* menggunakan metode HDBSCAN. Semakin kecil nilai jarak antar titik data, maka semakin besar kemungkinan data tersebut berada dalam kelompok klaster yang sama. Sebaliknya data dengan nilai jarak yang besar cenderung berada pada klaster yang berbeda karena memiliki karakteristik yang serupa.

4.3 Simulasi Implementasi *Clustering* HDBSCAN

Setelah nilai jarak *Euclidean* antar setiap titik data berhasil dihitung, proses selanjutnya adalah menerapkan algoritma HDBSCAN pada matriks jarak yang telah diperoleh. Algoritma HDBSCAN kemudian memproses matriks jarak tersebut untuk membangun struktur hierarki kepadatan data yang menjadi dasar dalam pembentukan klaster. Pada perhitungan manual, penelitian hanya menggunakan lima data pertama sebagai contoh untuk menjelaskan tahapan algoritma HDBSCAN secara sederhana. Oleh karena itu menggunakan *min_sample* 2 dipilih agar proses pembentukan klaster masih dapat terlihat. Jika pada lima data tersebut digunakan *min_sample* 5, maka setiap titik akan mempertimbangkan hampir seluruh data sebagai tetangga, sehingga hasilnya membentuk satu kelompok saja dan kurang menunjukkan proses pemisahan klaster.

1. Menghitung *Core Distance* (*min_sample* 2)

   Perhitungan *core distance* pada penelitian ini dilakukan menggunakan 5 data pertama sebagai sampel perhitungan manual, dengan tujuan untuk mempermudah proses penjelasan langkah-langkah algoritma HDBSCAN secara rinci dan sistematis.

   Pemilihan nilai *min_samples* 2 dilakukan karena jumlah data yang digunakan dalam perhitungan manual relatif kecil, sehingga nilai tersebut dianggap paling sesuai untuk menggambarkan konsep kepadatan tanpa menghilangkan terlalu banyak titik sebagai *noise*. Karena *min_sample* 2, makan *core distance* setiap titik adalah jarak ke tetangga terdekat ke-1 (*nearest neighbor*). Ini karena *min_sample* dihitung termasuk titik itu sendiri, sehingga tetangga yang diperhitungkan adalah tetangga ke (*min_sample* 1) =1. *Core distane* titik

p adalah jarak ke k terdekat dari 5 data pertama menggunakan Persamaan (2.3):

$$core_k(p_i) = d(p_i, tetangga\ terdekat\ ke-k\ dari\ titik\ p_i)$$

$$core_2(p_i) = d(p_i, tetangga\ terdekat\ ke-k\ dari\ titik\ p_i)$$

Substitusi rumus:

Core Distance p_1, $p_i = p_1$

Urutan tetangga terdekat dari $core_2(\ p_1)$:

$$\{d(p_1, p_4), d(p_1, p_2), d(p_1, p_5), d(p_1, p_3)\}$$

$$= \{0{,}4719529, 0{,}7751435, 1{,}1144167, 1{,}3140750\}$$

Jadi, $core_2(p_1) = d(p_1, p_2) = 0{,}7751435$

Core Distance p_2, $p_i = p_2$

Urutan tetangga terdekat dari $core_2(\ p_2)$:

$$\{d(p_2, p_4), d(p_2, p_1), d(p_2, p_3), d(p_2, p_5)\}$$

$$= \{0{,}4792508, 0{,}7751435, 1{,}0583879, 1{,}1155323\}$$

Jadi, $core_2(p_2) = d(p_2, p_1) = 0{,}7751435$

Core Distance p_3, $p_i = p_3$

Urutan tetangga terdekat dari $core_2(\ p_3)$:

$$\{d(p_3, p_5), d(p_3, p_2), d(p_3, p_4), d(p_3, p_1)\}$$

$$= \{0{,}5848182, 1{,}0583879, 1{,}1648098, 1{,}3140750\}$$

Jadi, $core_2(p_3) = d(p_3, p_2) = 1{,}0583879$

Core Distance p_4, $p_i = p_4$

Urutan tetangga terdekat dari $core_2(\ p_4)$:

$$\{d(p_4, p_1), d(p_4, p_2), d(p_4, p_5), d(p_4, p_3)\}$$

$$= \{0{,}4719529, 0{,}4792508, 1{,}0769049, 1{,}1648098\}$$

Jadi, $core_2(p_4) = d(p_4, p_2) = 0{,}4719529$

Core Distance p_5, $p_i = p_5$

Urutan tetangga terdekat dari $core_2(\ p_5)$:

$$\{d(p_5, p_3), d(p_5, p_4), d(p_5, p_1), d(p_5, p_2)\}$$

$$= \{0{,}5848182, 1{,}0769094, 1{,}1144167, 1{,}1155323\}$$

Jadi, $core_2(p_5) = d(p_5, p_4) = 1{,}0769094$

Tabel 4.4 Rekapitulasi *Core Distance*

Titik	*Nearest Neighbor*	*Core Distance*
p_1	p_2	0,7751435
p_2	p_1	0,7751435
p_3	p_2	1,0583879
p_4	p_2	0,4719529
p_5	p_4	1,0769094

Berdasarkan Tabel 4.4, setiap titik data memiliki *nearest neighboor* yang digunakan untuk menentukan nilai *core distance*. Nilai *core distance* menunjukkan jarak antara suatu titik dengan tetangga terdekatnya. Semakin kecil nilai yang diperoleh.

Titik p_4 memiliki nilai *core distance* yaitu, 0,4719529 dengan *nearest neighboor* p_2, yang menunjukkan bahwa p_4 memiliki kedekatan paling tinggi dibandingkan titik lainnya. Sementara itu, nilai *core distance* terbesar terdapat pada p_5 yaitu, 1,0769094 dengan *nearest neighboor* p_4, yang menunjukkan bahwa jarak p_5 terhadap tetangga terdekatnya relatif jauh. Selanjutnya, nilai c*ore distance* ini digunakan dalam perhitungan *mutual reachbility distanc*e sebagai dasar pembentukan *Minimum Spanning Tree* (MST) pada metode HDBSCAN.

2. Menghitung *Mutual Reachbility Distance*

Mutual reachability distance dihitung menggunakan Persamaan (2.9), yaitu dengan mengambil nilai maksimum dari *core distance* masing-masing titik dan jarak langsung antar titik. Secara matematis, *mutual reachability distance* antara dua titik tidak hanya mempertimbangkan jarak *Euclidean*, tetapi juga memperhitungkan tingkat kepadatan lokal di sekitar kedua titik tersebut yang direpresentasikan melalui nilai *core distance*.

Pendekatan ini bertujuan untuk mengurangi pengaruh perbedaan kepadatan dalam data, sehingga titik-titik yang berada di area dengan kepadatan rendah tidak secara langsung dianggap dekat hanya berdasarkan jarak *Euclidean*. Dengan demikian, *mutual reachability distance* memberikan representasi jarak yang lebih stabil dan sesuai dengan struktur kepadatan data, yang kemudian digunakan sebagai dasar dalam pembentukan struktur *Minimum Spanning Tree* (MST) pada algoritma HDBSCAN.

Mutual Reachbility Distance dihitung dengan persamaan (2.7):

$$mreach_k(p_i, p_j) = \max\{core_k(p_i),\ core_k(p_j),\ d(p_i, p_j)\}$$

Artinya jarak antar dua titik tidak hanya melihat jarak langsung, tetapi juga mempertimbangkan kepadatan lokal (*core distance*) masing-masing titik.

a. $mreach_2(p_1, p_2)$

Substitusi $core_2(p_1)$, $core_2(p_2)$ dan $d(p_1, p_2)$:

$$mreach_2(p_1, p_2) = \max\{core_2(p_1), core_2(p_2), d(p_1, p_2)\}$$
$$= \max\{0{,}4719529, 0{,}4792508, 0{,}7751435\}$$
$$= mreach_2(p_1, p_2) = 0{,}7751435$$

b. $mreach_2\ (p_1, p_3)$

Substitusi $core_2(p_1), core_2(p_3)$, dan $d(p_1, p_3)$:

$$mreach_2(p_1, p_3) = \max\{core_2(p_1), core_2(p_3), d(p_1, p_3)\}$$
$$= \max\{0{,}4719529, 0{,}5848182, 1{,}3140750\}$$
$$= mreach_2(p_1, p_3) = 1{,}3140750$$

c. $mreach_2\ (p_1, p_4)$

Substitusi $core_2(p_1), core_2(p_4)$, dan $d(p_1, p_4)$:

$$mreach_2(p_1, p_4) = \max\{core_2(p_1), core_2(p_4), d(p_1, p_4)\}$$
$$= \max\{0{,}4719529, 0{,}4719529, 0{,}4719529\ )$$
$$= mreach_2(p_1, p_4) = 0{,}4719529$$

d. $mreach_2\ (p_1, p_5)$

Substitusi $core_k(p_1), core_k(p_5)$, dan $d(p_1, p_5)$:

$$mreach_2(p_1, p_5) = \max\{core_2(p_1), core_2(p_5), d(p_1, p_5)\}$$
$$= \max\{0{,}4719529, 0{,}5848182, 1{,}1144167\}$$
$$= mreach_2(p_{!}, p_5) = 1{,}1144167$$

e. $mreach_2\ (p_2, p_3)$

Substitusi $core_k(p_2), core_k(p_3)$,dan $d(p_2, p_3)$:

$$mreach_2(p_2, p_3) = \max\{core_2(p_2), core_2(p_3), d(p_2, p_3)\}$$
$$= \max\{0{,}4792508, 0{,}5848182, 1{,}0583879\}$$
$$= mreach_2(p_2, p_3) = 1{,}0583879$$

f. $mreach_2\ (p_2, p_4)$

Substitusi $core_k(p_2), core_k(p_4)$,dan $d(p_2, p_4)$:

$$mreach_2(p_2, p_4) = \max\{core_2(p_2), core_2(p_4), d(p_2, p_4)\}$$
$$= \max\{0{,}4792508, 0{,}4719529, 0{,}4792508$$
$$= mreach_2(p_2, p_4) = 0{,}4792508$$

g. $mreach_2\ (p_2, p_5)$

Substitusi $core_k(p_2), core_k(p_5)$,dan $d(p_2, p_5)$:

$$mreach_2(p_2, p_5) = \max\{core_2(p_2), core_2(p_5), d(p_2, p_5)\}$$
$$= \max\{0{,}4792508, 0{,}5848182, 1{,}1155323$$
$$= mreach_2(p_2, p_5) = 1{,}1155323$$

h. $mreach_2\ (p_3, p_4)$

Substitusi $core_k(p_3), core_k(p_4)$, dan $d(p_3, p_4)$:

$$mreach_2(p3, p4) = \max\{core_2(p_3), core_2(p_4), d(p_3, p_4)\}$$
$$= \max\{0{,}5848182, 0{,}4719529, 1{,}1648098\}$$
$$= mreach_2(p_3, p_4) = 1{,}1648098$$

i. $mreach_2(p_3, p_5)$

Substitusi $core_k(p_3), core_k(p_5)$, dan $d(p_3, p_5)$:

$$mreach_2(p_3, p_5) = \max\{core_2(p_3), core_2(p_5), d(p_3, p_5)\}$$
$$= \max\{0{,}5848182, 0{,}5848182, 0{,}5848182\}$$
$$= mreach_2(p_3, p_5) = 0{,}5848182$$

j. $mreach_2\ (p_4, p_5)$

Substitusi $core_k(p_4), core_k(p_5)$, dan $d(p_4, p_5)$:

$$mreach_2(p_4, p_5) = \max\{core_2(p_4), core_2(p_5), d(p_4, p_5)\}$$
$$= \max\{0{,}4719529, 0{,}5848182, 1{,}0769049\}$$
$$= mreach_2(p_4, p_5) = 1{,}0769049$$

Tabel 4.5 Rekapitulasi *Mutual Reachbility Distance*

	p_1	p_2	p_3	p_4	p_5
p_1	0	0,7751435	1,3140750	0,4719529	1,1144167
p_2	0,7751435	0	1,0583879	0,4792508	1,1155323
p_3	1,3140750	1,0583879	0	1,1648098	0,5848182
p_4	0,4719529	0,4792508	1,1648098	0	1,0769049
p_5	1,1144167	1,1155323	0,5848182	1,0769049	0

Berdasarkan Tabel 4.5 rekapitulasi *Mutual Reachability Distance* menunjukkan nilai jarak antar titik data yang telah mempertimbangkan *core distance* masing-masing titik. Dalam metode HDBSCAN, *mutual reachability distance* digunakan untuk mengukur tingkat keterhubungan antar data dengan mempertimbangkan kepadatan lokal di sekitar titik data tersebut. Nilai ini diperoleh dari nilai maksimum antara jarak *Euclidean* dan *core distance* dari dua titik yang dibandingkan. Semakin kecil nilai *mutual reachability distance*, maka semakin tinggi tingkat kedekatan dan kepadatan antar titik data, sehingga peluang berada dalam *cluster* yang sama menjadi lebih besar.

Berdasarkan tabel 4.5, nilai *mutual reachbility distance* terkecil terdapat antara titik p_1 dan p_4 sebesar $0{,}4719529$. Nilai tersebut menunjukkan bahwa kedua titik memiliki tingkat kedekatan dan kepadatan yang paling tinggi dibandingkan pasangan lainnya. Selain itu, pasangan p_3 dan p_5 juga memiliki nilai yang relatif kecil yaitu sebesar $0{,}5848182$, yang menunjukkan karakteristik antar kedua titik data tersebut. Sementara itu, nilai terbesat terdapat pada pasangan titik p_3 dan p_4 sebesar $1{,}1648098$. Hal ini menunjukkan bahwa kedua titik memiliki tingkat perbedaan karakteristik yang cukup tinggi sehingga kemungkinan berada pada klaster yang berbeda.

Secara keseluruhan, matriks *mutual reachability distance* digunakan sebagai dasar dalam pembentukan struktur klaster pada metode HDBSCAN. Nilai jarak yang kecil menunjukkan hubungan antar data yang lebih rapat dan padat, sedangkan nilai yang besar menunjukkan hubungan yang lebih renggang. Oleh karena itu, hasil pada tabel 4.5 tersebut membantu dalam

menentukan pola kedekatan antar titik data sehingga proses pembentukan klaster dapat dilakukan dengan lebih optimal dan mempertimbangkan tingkat kepadatan data.

3. Membangun MST (*Minimum Spanning Tree*)

   Minimum Spanning Tree (MST) dibangun menggunakan algoritma Prim dengan memanfaatkan nilai *mutual reachability distance* sebagai bobot antar titik. Proses pembentukan MST dimulai dari salah satu titik, yaitu p_1, sebagai titik awal, karena titik tersebut memiliki nilai *mutual reachbility distance* yang paling kecil dengan titik p_4, yaitu sebesar $0{,}4719529$. Nilai yang kecil menunjukkan tingkat kedekatan yang tinggi sehingga hubungan antara kedua titik tersebut tanpa memengaruhi hasil akhir MST. Selanjutnya, pada setiap iterasi, dipilih sisi dengan bobot *mutual reachability distance* terkecil yang menghubungkan titik yang sudah termasuk dalam MST dengan titik lain yang belum tergabung dalam MST.

   Proses ini dilakukan secara berulang hingga seluruh titik data terhubung dalam satu struktur pohon tanpa membentuk siklus. Oleh karena itu, MST yang terbentuk mampu menggambarkan kedekatan antar titik data berdasarkan ukuran jarak yang telah disesuaikan dengan tingkar kepadatannya, sehingga menjadi dasar dalam pembentukan hierarki *cluster* pada algoritma HDBSCAN. MST dipilih dengan memilih sisi berbobot minimum dari titik dalam MST ke titik di luar MST terdapat di Persamaan (2.10).

$$w(p_i, p_j) = mreach_k(p_i, p_j).$$

   Iterasi 1 – MST = $\{p_1\}$ → cari sisi minimum ke $\{p_2, p_3, p_4, p_5\}$

$mraech_2\ (p_1, p_2) = 0{,}7751435$

$mreach_2\ (p_1, p_3) = 1{,}3140750$

$mreach_2\ (p_1, p_4) = 0{,}4719529 \leftarrow minimum$

$mreach_2\ (p_1, p_5) = 1{,}1144167$

Pilih sisi $p_1 - p_4$ dengan bobot $= 0{,}4719529 \rightarrow p_4$ bergabung ke MST.

Iterasi 2 – MST = $\{p_1, p_4\}$ → cari sisi minimum ke $\{p_2, p_3, p_5\}$

$Dari\ p_1 : mreach_2(p_1, p_2) = 0{,}7751435$

$mreach_2(p_1, p_3) = 1{,}3140750$

$mreach_2\ (p_1, p_5) = 1{,}1144167$

$Dari\ p_4 : mreach_2\ (p_4, p_2) = 0{,}4792508 \leftarrow minimum$

$mreach_2(p_4, p_3) = 1{,}1648098$

$mreach_2\ (p_4, p_5) = 1{,}0769049$

Pilih sisi $p_4 - p_2$ dengan bobot $= 0{,}4792508 \rightarrow p_2$ bergabung ke MST.

Iterasi 3 – MST $= \{p_1, p_4, p_2\}$ → cari sisi minimum ke $\{p_3, p_5\}$

$Dari\ p_1 : mreach_2(p_1, p_3) = 1{,}3140750$

$mreach_2(p_1, p_5) = 1{,}1144167$

$Dari\ p_2 : mreach_2(p_2, p_3) = 1{,}0583879$

$mreach_2(p_2, p_5) = 1{,}1155323$

$Dari\ p_4 : mrach_2\ (p_4, p_3) = 1{,}1648098$

$mreach_2\ (p_4, p_5) =\ 1{,}0769049$

Pilih sisi $p_2 - p_3$ dengan bobot $= 1{,}0583879 \rightarrow p_3$ bergabung ke MST.

Iterasi 4 – MST $= \{p_1, p_4, p_2, \ p_3\} \rightarrow$ cari sisi minimum ke $\{p_5\}$

$Dari\ p_1 : mreach_2\ (p_1, p_5) = 1{,}1144167$

$Dari\ p_2 : mreach_2\ (p_2, p_5) = 1{,}1155323$

$Dari\ p_3 : mreach_2\ (p_3, p_5) = 0{,}5848182$

$Dari\ p_4 : mreach_2 (p_4, p_5) = 1{,}0769049$

Pilih sisi $p_3 - p_5$ dengan bobot $= 0{,}5848182 \rightarrow p_5$ bergabung ke MST.

MST terdiri dari 4 sisi berikut (diurutkan dari bobot terbesar ke terkecil)

Tabel 4.6 Hasil bobot MST

Urutan	Sisi	w (bobot)
1	$p_2 - p_3$	1,0583879
2	$p_3 - p_5$	0,5848182
3	$p_4 - p_2$	0,4792508
4	$p_1 - p_4$	0,4719529

Berdasarkan tabel 4.6 hasil bobot MST, terdapat empat sisi yang menghubungkan seluruh titik data, yaitu $p_2 - p_3, p_3 - p_5, p_4 - p_2$, dan $p_1 - p_4$. Nilai w menunjukkan bobot hubungan antar titik yang diperoleh dari nilai *mutual reachability distance*. Semakin kecil nilai bobot, maka hubungan antar titik semakin dekat, sedangkan semakin besar nilai bobot menunjukkan hubungan yang lebih jauh atau lebih lemah. Pada tabel tersebut, sisi $p_1 - p_4$ memiliki bobot paling kecil sebesar 0,4719529, sehingga kedua titik ini memiliki hubungan paling dekat. Sebaliknya, sisi $p_2 - p_3$ memiliki bobot paling besar sebesar 1,0583879, sehingga menjadi hubungan paling lemah

dalam struktur MST. Oleh karena itu, pada proses pembentukan hierarki HDBSCAN, sisi dengan bobot terbesar akan diputus lebih dahulu untuk melihat pemisahan klaster yang terbentuk.

$MST\ SELESAI - Semua\ titik\ terhubung\ dengan\ 4\ sisi$:

$Sisi\ MST: p_2 - p_3\ (1{,}058387) + p_3 - p_5\ (0{,}5848182) + p_4$

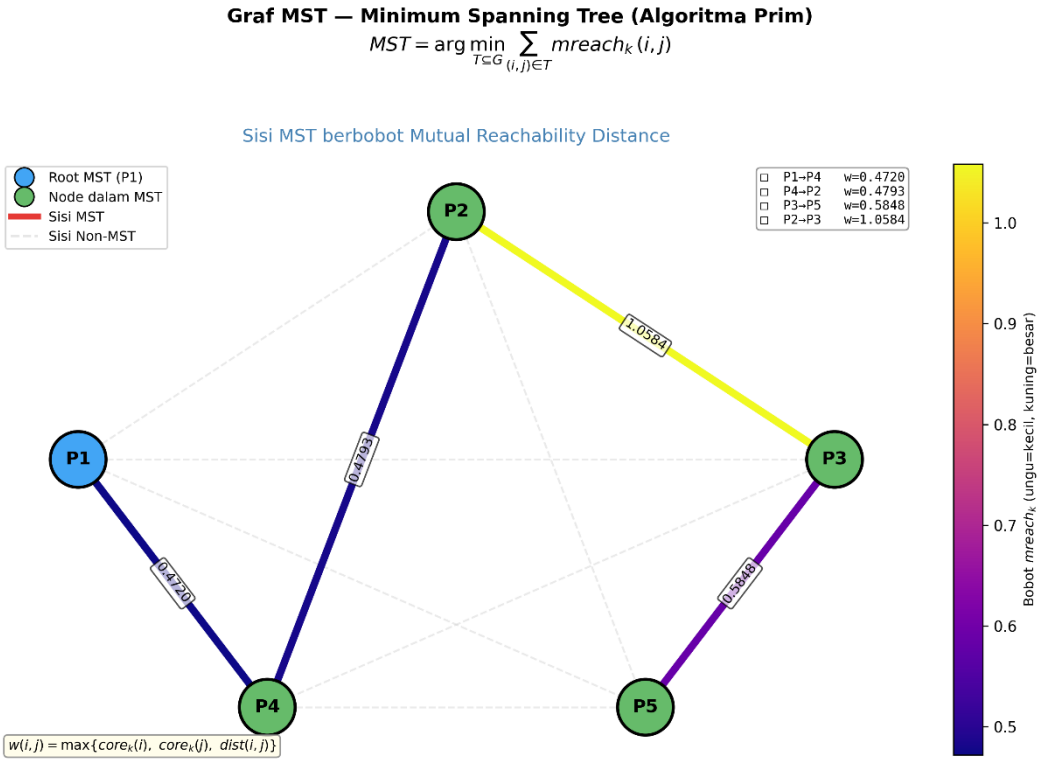
$- p_2(0{,}4792508) + p_1 - p_4\ (0{,}4719529)$

$Total\ bobot\ MST = 2{,}5944098$

$Struktur: p_2 - p_3 - p_5 - p_4 - p_1.$

Berdasarkan Gambar 4.1 menunjukkan hasil pembentukan *Minimum Spanning Tree* (MST) menggunakan algoritma Prim. Setiap titik $p1$ sampai $p5$ dihubungkan berdasarkan bobot *mutual reachability distance* terkecil, sehingga seluruh titik dapat terhubung tanpa membentuk siklus.

Pada penelitian ini, *Minimum Spanning Tree* (MST) dibentuk menggunakan algoritma Prim karena algoritma tersebut mampu menghubungkan seluruh titik data berdasarkan bobot *mutual reachability distance* terkecil tanpa membentuk siklus. Proses pemilihan sisi dilakukan secara bertahap sehingga hubungan antar titik yang memiliki tingkat kedekatan dan kepadatan tinggi dapat direpresentasikan dengan baik dalam struktur MST. Dari hasil MST, diperoleh empat sisi yang menghubungkan seluruh data, yaitu (p_1, p_4), (p_4, p_2), (p_2, p_3), dan (p_3, p_5). Sisi $p_1 - p_4$ memiliki bobot paling kecil, yaitu $0{,}4720$, sehingga menunjukkan hubungan paling dekat. Sebaliknya, sisi $p_2 - p_3$ memiliki bobot paling besar, yaitu $1{,}0584$, sehingga menjadi hubungan paling lemah dalam struktur MST.



Gambar 4.1 Pembentukan MST

4. Membentuk Herarki Klaster

Hierarki klaster dalam algoritma HDBSCAN dibentuk berdasarkan struktur *Minimum Spanning Tree* (MST) dengan memanfaatkan nilai *mutual reachability distance* sebagai bobot antar titik. Proses pembentukan hierarki dilakukan dengan memotong sisi-sisi pada MST secara bertahap, dimulai dari sisi dengan bobot terbesar hingga terkecil. Setiap pemotongan sisi akan memisahkan struktur pohon menjadi beberapa komponen yang kemudian merepresentasikan klaster-klaster baru pada tingkat kepadatan tertentu.

Dalam proses ini, digunakan parameter λ (lambda) yang didefinisikan sebagai kebalikan dari nilai w $\left(\lambda = \frac{1}{w}\right)$, yang menunjukkan tingkat kepadatan data. Semakin besar nilai λ, maka tingkat kepadatan yang dipertimbangkan semakin tinggi, sehingga klaster yang terbentuk cenderung lebih kecil namun

lebih padat. Dengan demikian, hierarki klaster yang dihasilkan menggambarkan perubahan struktur klaster pada berbagai tingkat kepadatan, yang selanjutnya digunakan dalam proses penentuan klaster yang paling stabil.

Sisi MST dihapus dari bobot terbesar ke terkecil. Tingkat kepadatan $\lambda = \frac{1}{w}$ terdapat persamaan (2.13):

$$\lambda = \frac{1}{w}$$

Hapus sisi ke-1: $p_2 - p_3$ (bobot terbesar $= 1{,}0583879$)

$$w = 1{,}0583879$$

$$\lambda = \frac{1}{1{,}0583879} = 0{,}9448332$$

Asal komponen $\{p_1, p_2, p_3, p_4, p_5\}$ Klaster terbagi menjadi: $\{p_1, p_2, p_4\}$ dan $\{p_3, p_5\}$.

Hapus sisi ke-2: $p_5 - p_3$ (bobot $= 0{,}5848182$)

$$w = 0{,}5848182$$

$$\lambda = \frac{1}{0{,}5848182} = 1{,}7099333$$

Klater terpecah menjadi : $\{p_1, p_2, p_4\}$ dan $\{p_3\}$ dan $\{p_5\}$

Hapus sisi ke-3: $p_4 - p_2$ (bobot $= 0{,}4792508$)

$$w = 0{,}4792508$$

$$\lambda = \frac{1}{0{,}4792508} = 2{,}0865902$$

Klaster terpercah menjadi: $\{p_1, p_4\}$ dan $\{p_2\}$ dan $\{p_3\}$ dan $\{p_5\}$

Hapus sisi ke-4: $p_1 - p_4$ (bobot terkecil $= 0{,}4719529$)

$$w = 0{,}4719529$$

$$\lambda = \frac{1}{0{,}4719529} = 2{,}1188555$$

Klaster terpecah menjadi: $\{p_1\}\ \{p_2\}\ \{p_3\}\ \{p_4\}\ \{p_5\}$ (semua terpisah).

Pemecahan klaster terjadi karena pada setiap tahap dilakukan penghapusan sisi pada *Minimum Spanning Tree* (MST) berdasarkan urutan bobot *mutual reachability distance* dari yang terbesar ke terkecil, di mana sisi dengan bobot terbesar merepresentasikan hubungan paling lemah antar titik (jarak efektif paling jauh setelah mempertimbangkan kepadatan). Ketika sisi tersebut dihapus, konektivitas antar titik yang sebelumnya terhubung menjadi terputus, sehingga struktur pohon terpecah menjadi beberapa komponen yang membentuk klaster baru. Proses ini dikaitkan dengan peningkatan nilai λ $\left(\lambda = \frac{1}{w}\right)$, yang menunjukkan bahwa tingkat kepadatan yang disyaratkan semakin tinggi, sehingga hanya titik-titik yang memiliki kedekatan dan kepadatan yang cukup tinggi yang tetap tergabung dalam satu klaster, sementara titik yang kurang padat atau memiliki jarak lebih besar akan terpisah menjadi klaster tersendiri atau bahkan menjadi *noise*.

5. Menyederhanakan Hierarki Berdasarkan *Minimum Cluster Size*

   Setelah hierarki klaster terbentuk dari penghapusan sisi MST, setiap sub-klaster yang muncul dicek apakah memenuhi syarat minimum ukuran. Klaster dipertahankan jika dan hanya jika memenuhi persamaan (2.17) dan (2.18):

   $C_i\ valid \iff N(C_i) \geq m_{clustersize} = 2$

   $C_i\ tidak\ valid \iff N(C_i) < m_{clustersize} = 2$

   Proses pengecekan dilakukan pada setiap pemecahan klaster yang terjadi saat sisi MST dihapus satu persatu:

a. Penghapusan sisi $p_2 - p_3$ $(w = 1{,}0583879, \lambda = 0{,}9448332)$

Sebelum penghapusan sisi ini seluruh 5 titik data berada dalam satu komponen terhubung $\{p_1, p_2, p_3, p_4, p_5\}$. Setelah sisi $p_2 - p_3$ dihapus, klaster *root* terbagi menjadi dua komponen:

Komponen A: $\{p_1, p_2, p_4\}$

Cek validitas: $N\ (\{p_1, p_2, p_4\}) = 3$

Terapkan Persamaan (2.13) dan (2.14): $N = 3 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_1, p_2, p_4\}$ valid tapi terbagi lagi

Komponen B: $\{p_3, p_5\}$

Cek validitas: $N(\{p_3, p_5\}) = 2$

Terapkan Persamaan (2.13) dan (2.14): $N = 2 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_3, p_5\}$ valid klaster final 0.

b. Penghapusan sisi $p_4 - p_2$ $(w = 0{,}4792508, \lambda = 2{,}0865902)$

Sub-klaster $\{p_1, p_2, p_4\}$ terbagi menjadi $\{p_1, p_4\}$ dan $\{p_2\}$:

Komponen C: $\{p_1, p_4\}$

Cek validitas: $N(\{p_1, p_4\}) = 2$

Terapkan Persamaan (2.13) dan (2.14): $N = 2 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_1, p_4\}$ valid klaster final 1.

Komponen D: $\{p_2\}$

Cek validitas: $N(\{p_2\}) = 1$

Terapkan Persamaan (2.13) dan (2.14):$N = 1 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_2\}$ *noise*.

c. Penghapusan sisi $p_3 - p_5$ $(w = 0{,}5848182, \lambda = 1{,}70993331)$

Sub-klaster $\{p_3, p_5\}$ terbagi menjadi $\{p_3\}$ dan $\{p_5\}$ secara terpisah:

Komponen E: $\{p_3\}$ (dari Komponen B)

Cek validitas:$N(\{p_3\}) = 1$

Terapkan Persamaan (2.13) dan (2.14):$N(\{p_3\}) = 1 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_3\}$ Keluar dari Klaster B

Komponen F: $\{p_5\}$ (dari Komponen B)

Cek Validitas:$N(\{p_5\}) = 1$

Terapkan Persamaan (2.13) dan (2.14):$N(\{p_5\}) = 1 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_5\}$ Keluar dari Klaster B

d. Penghapusan sisi $p_1 - p_4$ ($w = 0{,}4719529, \lambda = 2{,}1188555$

Sub-klaster $\{p_1, p_4\}$ terbagi menjadi $\{p_1\}$ dan $\{p_4\}$ secara terpisah:

Komponen G: $\{p_1\}$ (dari Komponen C)

Cek Validitas: $N(\{p_1\}) = 1$

Terapkan Persamaan (2.13) dan (2.14):$N(\{p_1\}) = 1 \geq m_{clustersize} = 2$

Kesimpulan: $\{p_1\}$ Keluar dari Klaster C

Komponen H: $\{p_4\}$ (dari Komponen C)

Cek Validitas:$N(\{p_4\}) = 1$

Terapkan Persamaan (2.13) dan (2.14):$N(\{p_4\}) = 1 \geq m_{clustersize} = 2$

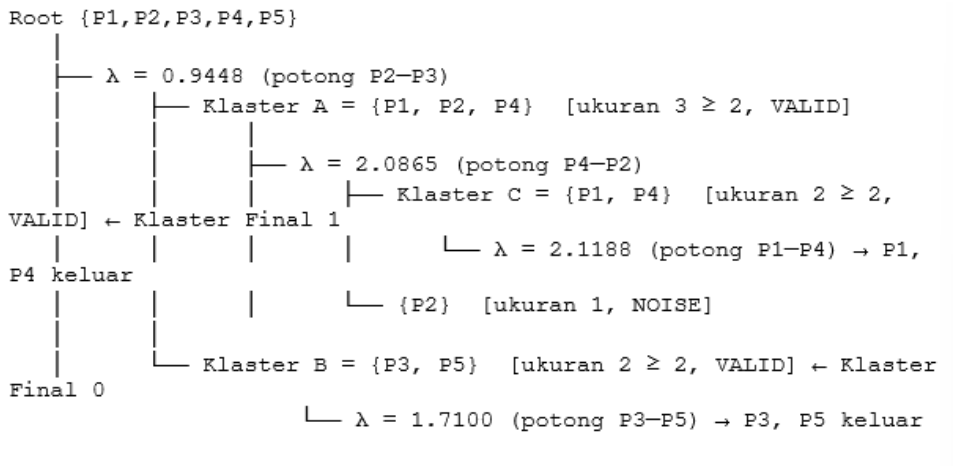
Kesimpulan: $\{p_4\}$ Keluar dari Klaster C

Tabel 4.7 Rekapitulasi Hasil Penyederhanaan HIerarki

Komponen	$N(C)$	Syarat	Hasil Cek	Status
Root$\{p_1, p_2, p_3, p_4, p_5\}$	5	≥ 2	$5 \geq 2$	Valid sebagai komponen awal.
Klaster A$\{p_1, p_2, p_4\}$	3	≥ 2	$3 \geq 2$	Valid sebagai klaster sementara.
Klaster B$\{p_3, p_5\}$	2	≥ 2	$2 \geq 2$	Valid (Klaster 0)
Klaster C$\{p_1, p_4\}$	2	≥ 2	$2 \geq 2$	Valid (Klaster 1)
$\{p_2\}$	1	≥ 2	$1 < 2$	Noise

$\{p_3\}$ (dari Komponen B)	1	≥ 2	$1 < 2$	Keluar dari Klaster B
$\{p_5\}$ (dari Komponen B)	1	≥ 2	$1 < 2$	Keluar dari Klaster B
$\{p_1\}$ (dari Komponen C)	1	≥ 2	$1 < 2$	Keluar dari Klaster C
$\{p_4\}$ (dari Komponen C)	1	≥ 2	$1 < 2$	Keluar dari Klaster C

Berdasarkan Tabel 4.7, proses penyederhanaan hierarki dilakukan dengan melihat jumlah anggota pada setiap komponen. Suatu komponen dinyatakan valid apabila jumlah anggotanya memenuhi syarat minumum, yaitu $N(C_i) \geq m_{clustersize} = 2$. Komponen awal $\{p_1, p_2, p_3, p_4, p_5\}$ dinyatakan valid karena memiliki 5 anggota, tetapi masih berperan sebagai komponen awal dalam stuktur hierarki. Selanjutnya, Klaster A $\{p_1, p_2, p_4\}$ juga valid karena memiliki 3 anggota, namun masih dianggap sebagai klaster sementara karena pada tahap berikutnya masih membentuk komponen yang lebih kecil. Klaster B $\{p_3, p_5\}$ dan Klaster C $\{p_1, p_4\}$ dinyatakan sebagai klaster akhir karena masing-masing memiliki 2 anggota dan memenuhi syarat $N(C_i) \geq m_{clustersize} = 2$. Sementara itu, komponen yang hanya terdiri dari satu titik, seperti $\{p_2\}, \{p_3\}, \{p_5\}, \{p_1\}$, dan $\{p_4\}$, tidak dipertahankan sebagai klaster tersendiri karena jumlah anggotanya kurang dari batas minimum. p_2 dikategorikan sebagai *noise*, sedangkan p_3, p_5, p_1 dan p_4 tetap menjadi bagian dari klaster akhir yang sudah terbentuk sebelumnya.



Gambar 4.2 Struktur *Condesed Tree*

Gambar 4.2 menunjukkan proses pembentukan hierarki klaster pada HDBSCAN setelah sisi-sisi pada MST dipotong berdasarkan nilai λ. Pada awalnya, seluruh titik berada dalam satu komponen utama, yaitu $\{p_1, p_2, p_3, p_4, p_5\}$. Ketika sisi $p_2 - p_3$ dipotong pada $\lambda = 0{,}9448$, komponen awal membentuk dua bagian, yaitu Klaster A $\{p_1, p_2, p_4\}$ dan Klaster B $\{p_3, p_5\}$. Klaster B dinyatakan valid karena memiliki 2 anggota dan memenuhi syarat *min_cluster_size* 2, sehingga dipertahankan sebagai klaster final. Selanjutnya, Klaster A membentuk bagian baru ketika sisi $p_4 - p_2$ dipotong pada $\lambda = 2{,}0865$, sehingga terbentuk Klaster C $\{p_1, p_4\}$ dan titik $p2$ yang berdiri sendiri. Klaster C dinyatakan valid karena memiliki 2 anggota, sedangkan p_2 dikategorikan sebagai *noise* karena hanya terdiri dari satu titik dan tidak memenuhi batas minimum ukuran klaster. Dengan demikian, hasil akhir dari proses hierarki ini menghasilkan dua klaster final, yaitu Klaster Final 0 $\{p_3, p_5\}$, Klaster Final 1 $\{p_1, p_4\}$, dan satu data *noise* yaitu p_2.

6. Perhitungan Stabilitas Klaster

Stabilitas klister C dihitung menggunakan persamaan (2.21):

$$S(C_i) = \sum_{p_j \in C_i} (\lambda_{\max}(p_j, C_i) - \lambda_{\min}(C_i))$$

Keterangan:

$\lambda_{\min(c_i)} = \lambda\ saat\ klaster\ \ lahir\ (pertama\ kali\ terbentuk)$

$\lambda_{\max(pj,C_i)} = \lambda\ saat\ titik\ xj\ keluar\ dari\ klaster$

$\Delta\lambda(pj) = \lambda_{\max(pj,c)} - \lambda_{\min(c)}$

a. Stabilitas Klaster B $= \{p_3, p_5\}$

$p_j = p_3$

$\lambda_{min}(B) = 0{,}9448332$ (Klaster B didapatkan dari pemotongan $p2 - p3, w = 1{,}0583879$)

$\lambda_{\max}(B) = 1{,}7099333$ (Hasil didapatkan dari pemotongan (p_3, p_5)

$p_j = p_3$

$\lambda_{min}(B) = 0{,}9448332$

$\lambda_{\max}(B) = 1{,}7099333$

$\Delta\lambda(p_3) = \lambda_{\max} - \lambda_{\min}$

$= 1{,}7099333 - 0{,}9448332$

$= 0{,}7651001$

$\Delta\lambda(p_3) = 0{,}7651001$

$p_j = p_5$

$\lambda_{min}(B) = 0{,}9448332$

$\lambda_{\max}(B) = 1{,}7099333$

$\Delta\lambda(p_5) = \lambda_{\max} - \lambda_{\min}$

$= 1{,}7099333 - 0{,}9448332$

$= 0{,}7651001$

$\Delta\lambda(p_5) = 0{,}7651001$

$$S(B) = 0{,}7651001 + 0{,}7651001 = 1{,}5302002$$

b. Stabilitas Klaster C = $\{p_1, p_4\}$

$\lambda_{min}(C) = 2{,}0865902$ (Klaster C didapatkan dari pemotongan $p_4 - p_2$,

$w = 0{,}4792508$

$\lambda_{max}(C) = 2{,}1188555$ (Hasil dari penggabungan bobot klaster p_1, p_4

atau berdasarkan Persamaan (2.11)

$p_j = p_1$

$\lambda_{min}(C) = 2{,}0865902$

$\lambda_{\max}(C) = 2{,}1188555$

$\Delta\lambda(p_1) = \lambda_{\max} - \lambda_{\min}$

$= 2{,}1188555 - 2{,}0865902$

$= 0{,}0322653$

$\Delta\lambda(p_1) = 0{,}0322653$

$p_j = p_4$

$\lambda_{min}(C) = 2{,}0865902$

$\lambda_{\max}(C) = 2{,}1188555$

$\Delta\lambda(p_4) = \lambda_{\max} - \lambda_{\min}$

$= 2{,}1188555 - 2{,}0865902$

$= 0{,}0322653$

$\Delta\lambda(p_4) = 0{,}0322653$

$$S(C) = 0{,}0322653 + 0{,}0322653 = 0{,}0645306$$

c. Perbandingan Stabilitas Klaster A dengan Sub-klaster C

Karena Klaster A = $\{p_1, p_2, p_4\}$ memiliki sub-klaster C = $\{p_1, p_4\}$, maka

harus membandingkan apakah lebih baik memiliki A secara keseluruhan atau memilih C sebagai sub-klaster.

S(A) dihitung seolah tidak ada sub klaster (semua titik bertahan sampai masing masing keluar):

$p_j = p_1$

$\lambda_{max}(p_1) = 2{,}1188555$

$\lambda_{min}(A) = 0{,}9448332$

$\Delta\lambda(p_1) = \lambda_{\max} - \lambda_{\min}$

$= 2{,}1188555 - 0{,}9448332$

$= 1{,}1740223$

$p = p_2$

$\lambda_{max}(p_2) = 2{,}0865902$

$\lambda_{min}(A) = 0{,}9448332$

$\Delta\lambda(p_2) = \lambda_{\max} - \lambda_{\min}$

$= 2{,}0865902 - 0{,}9448332$

$= 1{,}1417570$

$p_j = p_4$

$\lambda_{max}(p_4) = 2{,}1188555$

$\lambda_{min}(A) = 0{,}9448332$

$\Delta\lambda(p_2) = \lambda_{\max} - \lambda_{\min}$

$= 2{,}1188555 - 0{,}9448332$

$= 1{,}1740223$

$S(A) = 1{,}1740223 + 1{,}1417570 + 1{,}1740223 = 3{,}4898016$

$S(A)$ jika tidak dipecah memperoleh nilai stabilitas $3{,}4898016$

$S(C)$ jika dipilih sub-klaster C memperoleh nilai stabilitas $0{,}0645306$

Jadi, nilai stabilitas (A) memperoleh $3{,}4898016$ lebih besar dari pada stabilitas (C) memperoleh $0{,}0645306$, pilih $S(A)$ sebagai klaster akhir (bukan sub klasternya).

Artinya, klaster A dianggap lebih stabil dalam bentuk utuh, namun karena A kemudia terpecah dan hanya meninggalkan $C = \{p_1, p_4\}$ yang valid, maka C adalah representasi akhir dari A.

7. Hasil Akhir *Clustering*

Klaster akhir adalah klaster dengan stabilitas tertinggi dengan Persamaan (2.25):

$$C_{\text{final}} = \underset{C_i}{argmax}\, S(C_i)$$

Ringkasan semua klaster valid beserta stabilitasnya:

Setelah *condensed tree* terbentuk, setiap klaster yang valid (ukuran ≥ *min_cluster_size*) dievaluasi berdasarkan nilai stabilitasnya. Klaster dengan stabilitas tertinggi dipilih sebagai klaster akhir. Jika sebuah klaster memiliki sub-klaster, maka dibandingkan apakah lebih baik mempertahankan klaster induk atau memecahnya menjadi sub-klaster.

Substitusi Nilai Stabilitas ke Persamaan 2.19:

Dari Langkah ke 7, diperoleh nilai stabilitas untuk setiap klaster valid:

$$S(Klaster\ B) = S(\{p_3, p_5\}) = \Delta\lambda(p_3) + \Delta\lambda(p_5)$$

$$= \left(\lambda_{max}(p_3, B) - \lambda_{min}(B)\right) + \left(\lambda_{max}(p_5, B) - \lambda_{min}(B)\right)$$

$$= (1{,}7099333 - 0{,}9448332) + (1{,}7099333 - \ 0{,}9448332)$$

$$= 0{,}7651001 + 0{,}7651001$$

$$= 1{,}5302002$$

$$S(Klaster\ C) = S(\{p_1, p_4\}) = \Delta\lambda(p_1) + \Delta\lambda(p_4)$$

$$= \left(\lambda_{max}(p_1, B) - \lambda_{min}(B)\right) + \left(\lambda_{max}(p_4, B) - \lambda_{min}(B)\right)$$

$$= (2{,}1188555 - 2{,}0865902) + (2{,}1188555 - 2{,}0865902)$$

$$= 0{,}0322653 + 0{,}0322653$$

$$= 0{,}0645306$$

Penerapan Persamaan 2.19 (pilih stabilitas tertinggi)

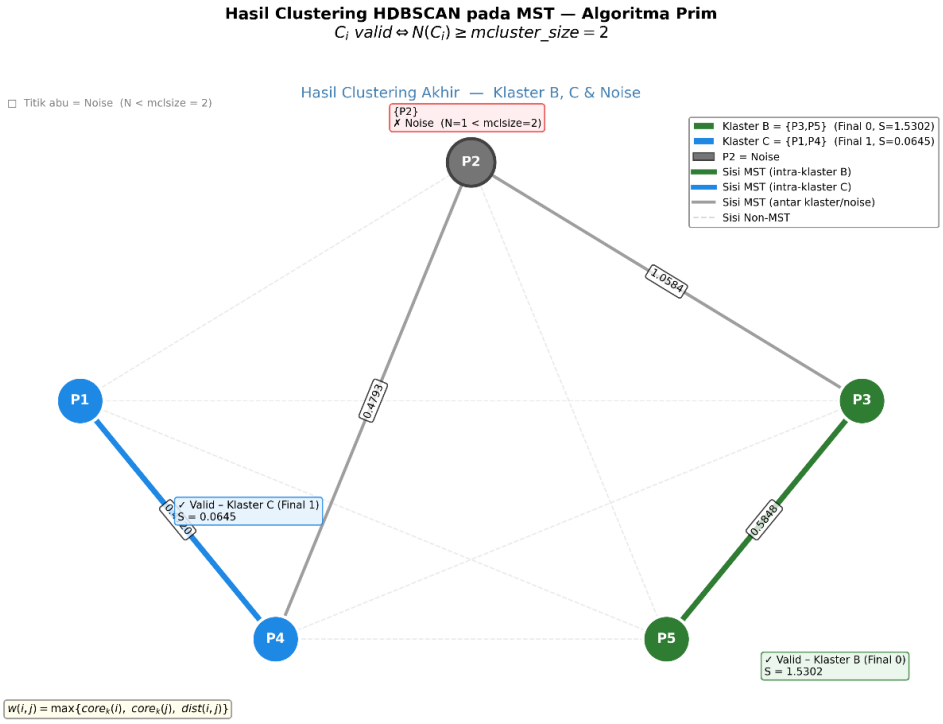
$$C_{final} = argmax_{\{C_i\}} S(C_i)$$

$$= argmax\,\{S(Klaster\ B), S(Klaster\ C)\}$$

$$= argmax\,\{1{,}5302002, 0{,}0645306\}$$

Klaster B mendapatkan nilai stabilitas lebih besar 1,5302002 dibandingkan dengan Klaster C mendapatkan nilai 0,0645306.

Jadi, argumen *maximal cluster* yaitu Klaster B dengan nilai stabilitas 1,5302002.



Gambar 4.3 Hasil akhir *clustering*

Berdasarkan Gambar 4.3, Klaster B = $\{p_3, p_5\}$ dipilih sebagai klaster dengan stabilitas tertinggi. Hal ini menunjukkan bahwa klaster tersebut mampu bertahan lebih lama dalam struktur hierarki HDBSCAN sehingga memiliki tingkat kestabilan yang lebih baik. Sementara itu, Klaster C = $\{p_1, p_4\}$ memiliki nilai stabilitas yang lebih rendah, namun tetap dipertahankan sebagai klaster akhir karena memenuhi syarat minimum ukuran klaster. Adapun p_2 dikategorikan sebagai noise karena tidak tergabung dalam klaster yang memenuhi ukuran minimum dan hanya membentuk komponen tunggal.

Tabel 4.8 Rekapitulasi Hasil Stabilitas

Label	Anggota	Stabilitas	Status
Klaster B	$\{p_3, p_5\}$	1,5302002	Klaster 0
Klaster C	$\{p_1, p_4\}$	0,0645306	Klaster 1
p_2	$\{p_2\}$	–	Noise

Berdasarkan Tabel 4.8 hasil proses *clustering* menggunakan metode HDBSCAN, diperoleh dua klaster utama dan satu data yang dikategorikan sebagai *noise*. Klaster 0 terdiri dari p_3 dan p_5, sedangkan Klaster 1 terdiri dari p_1 dan p_4. Kedua pasangan tersebut tergabung dalam klaster karena memiliki kedekatan jarak dan kepadatan yang relatif serupa. Sementara itu, p_2 dikategorikan sebagai *noise* karena terpisah sebagai komponen tunggal dan tidak memenuhi syarat batas minimum ukuran klaster. Oleh karena itu, p_2 tidak dimasukkan ke dalam klaster akhir dan tidak digunakan dalam perhitungan evaluasi. Hasil ini hanya digunakan sebagai ilustrasi tahapan perhitungan, sehingga kesimpulan utama penilitian tetap mengacu pada hasil clustering seluruh dataset.

Tabel 4.9 Hasil Akhir Label per Klaster

Titik	Klaster	$\lambda_{\min}(C)$	$\lambda_{\max}(C)$	$\Delta\lambda$	Label Final
p_1	Klaster C	2,0865902	2,1188555	0,0322653	Klaster 1
p_2	-	-	-	-	-
p_3	Klaster B	0,9448332	1,7099333	0,7651001	Klaster 0
p_4	Klaster C	2,0865902	2,1188555	0,0322653	Klaster 1
p_5	Klaster B	0,9448332	1,7099333	0,7651001	Klaster 0

Berdasarkan Tabel 4.9 menunjukkan hasil penentuan label klaster untuk setiap titik data setelah proses HDBSCAN. Setiap titik diberikan nilai λ_{min} dan λ_{max} yang digunakan utuk menghitung durasi keberadaan klaster ($\Delta\lambda$). Titik p_1 dan p_4 tergabung dalam klaster C, dan karena keduanya memenuhi syarat minimum ukuran klaster, akhirnya diberi label Klaster 1, Titik p_3 dan p_5 tergabung dalam klaster B, dengan durasi $\Delta\lambda$ yang lebih panjang, sehingga diberi label klaster 0. Sementara itu, titik p_2 tidak tergabung dalam klaster manapun karena jumlah anggotanya kurang dari batas minimum, sehingga dikategorikan sebagai data *noise* dan tidak mendapatkan label.

4.4 Evaluasi Hasil *Clustering* HBSCAN

Setelah proses pengelompokan selesai dilakukan, tahap selanjutnya adalah mengevaluasi seberapa baik hasil pengelompokan tersebut menggunakan dua evaluasi, yaitu *Silhouette Coefficient* dan *Davies Bouldin Index*.

1. *Silhouette Coefficient*

   Silhouette Coeffiecient mengukur kualitas setiap titik dalam klasternya. Nilainya berkisar dari -1 hingga +1, semakin tinggi semakin baik. Rumus di Persamaan (2.26)

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Keterangan:

$a(i)$: Rata-rata jarak titik i ke semua anggota *cluster* yang sama (kohesi *intra-cluster*).

$b(i)$: Rata-rata jarak terkecil titik i ke anggota *cluster* lain terdekat (seprasi *inter-cluster*).

Dalam proses evaluasi menggunakan *Silhouette Coefficient*, titik p_2 yang dikategorikan sebagai *noise* tidak disertakan dalam perhitungan. Hal ini disebabkan karena data *noise* tidak termasuk dalam klaster manapun, sehingga tidak memiliki nilai rata-rata jarak *intra-cluster* (a) maupun jarak ke klaster terdekat (b) yang diperlukan dalam perhitungan *Silhouette*. Oleh karena itu, untuk menjaga validitas dan akurasi hasil evaluasi, hanya titik-titik yang tergabung dalam klaster yang digunakan dalam perhitungan nilai *Silhouette*.

a. Perhitungan $S(p_1)$ – Klaster 1 = $(\{p_1, p_4\})$, $i = p_1$

Anggota klaster sama = p_4

Klaster lain = p_3, p_5

Hitung $a(p_1)$ pada Klaster 1

$d(p_1, p_4) = 0{,}4719529$

$a(p_1) = \frac{0{,}4719529}{1} = 0{,}4719529$

Hitung $b(i)$

$b(p_1) = \frac{[d(p_1,p_3)+d(p_1,p_5)]}{2} = \frac{[1{,}3140750+1{,}1144167]}{2} = \frac{2{,}4788848}{2} = 1{,}2394424$

Hitung $\max\{a, b\}$

$\max\{0{,}4719529\ ,1{,}2142458\} = 1{,}2142458$

$S(p_1) = \frac{1{,}2142458-0{,}4719529}{1{,}2142458} = \frac{0{,}7422929}{1{,}2142458} = 0{,}6113201$

b. Perhitungan $S(p_3)$ – Klaster 0 = $\{p_3, p_5\}$, $i = p_3$

Anggota klaster sama = p_5

Klaster lain = p_1, p_4

Hitung $a(p_3)$ pada Klaster 0

$d(p_3, p_5) = 0{,}5848182$

$a(p_3) = \frac{0{,}5848182}{1} = 0{,}5848182$

Hitung $b(i)$

$b(p_3) = \frac{[d(p_3,p_1)+d(p_3,p_4)]}{2} = \frac{[1{,}3140750+1{,}1648098]}{2} = \frac{2{,}4788848}{2} =$

$1{,}2394424$

Hitung max $\{a, b\}$

$\max\{0{,}5848182, 1{,}2394424\} = 1{,}2394424$

$S(p_3) = \frac{1{,}2394424-0{,}5848182}{1{,}2394424} = \frac{0{,}6546242}{1{,}2394424} = 0{,}5281602$

c. Perhitungan $S(p_4)$ –Klaster 1 = $\{p_1, p_4\}$,$i = p_4$

Anggota klaster sama = p_1

Klaster lain = p_3, p_5

Hitung $a(p_4)$ pada Klaster 1

$d(p_4, p_1) = 0{,}4719529$

$a(p_4) = \frac{0{,}4719529}{1} = 0{,}4719529$

Hitung $b(i)$

$b(p_4) = \frac{[d(p_4,p_3)+d(p_4,p_5)]}{2} = \frac{[1{,}1648098+1{,}0769049]}{2} = \frac{2{,}2417147}{2} =$

$1{,}1208572$

Hitung max$\{a, b\}$

$\max\{0{,}4719529, 1{,}1208572\} = 1{,}1208572$

$$S(p_4) = \frac{1{,}1208572 - 0{,}4719529}{1{,}1208572} = \frac{0{,}6489043}{1{,}1208572} = 0{,}5789357$$

d. Perhitungan $S(p_5)$ −Klaster 0 = $\{p_3 - p_5\}$, $i = p_5$

Anggota klaster sama = p_3

Klaster lain = p_1, p_4

Hitung $a(p_5)$ pada Klaster 0

$d(p_5, p_3) = 0{,}5848182$

$$a(p_5) = \frac{0{,}5848182}{1} = 0{,}5848182$$

Hitung $b(i)$

$$b(p_5) = \frac{[d(p_5,p_1)+d(p_5,p_4)}{2} = \frac{[1{,}1144167+1{,}0769049]}{2} = \frac{2{,}1913216}{2} = 1{,}0956608$$

Hitung max $\{a, b\}$

$\max\{0{,}5848182, 1{,}0956608\} = 1{,}0956608$

$$S(p_5) = \frac{1{,}0956608 - 0{,}5848182}{1{,}0956608} = \frac{0{,}5108426}{1{,}0956608} = 0{,}4662416$$

Tabel 4.10 Hasil Evaluasi *Silhouette Coefficient*

Titik	Label	$s(i)$
p_1	Klaster 1	0,6113201
p_2	Noise	$Dikecualikan$
p_3	Klaster 0	0,5281602
p_4	Klaster 1	0,5789357
p_5	Klaster 0	0,4662416

Rata-rata *Silhouette Score* $= \frac{(0{,}6113201+0{,}5281602+0{,}5789357+0{,}4662416)}{4}$

$$= \frac{2{,}1846576}{4} = 0{,}5461644$$

Berdasarkan Tabel 4.10 hasil perhitungan, diperoleh nilai *Silhouette Coeffiecient* sebesar 0,5461644 nilai tersebut menunjukkan bahwa hasil *clustering* telah mampu membentuk struktur klaster yang cukup jelas, di mana objek dalam klaster cenderung memiliki kemiripan yang lebih tinggi

dibandingkan dengan objek klaster lain. Oleh karena itu, hasil ini tetap mencerminkan bahwa struktur klaster yang dihasilkan telah mampu merepresentasikan pola data dengan cukup baik.

2. *Davies Bouldin Index*

Davies Bouldin Index mengukur rata-rata kemiripan antar *cluster*. Semakin kecil nilainya semakin baik pemisahan *cluster*. Dilihat Persamaan (2.27):

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i$$

dengan,

$$R_i = maxR_{ij}$$

dan,

$$R_{ij} = \frac{var(C_i) + var(C_j)}{|c_i - c_j|}$$

Keterangan:

$var(c_i)$: Scatter = rata-rata jarak tiap titik dalam c_i ke *centroid* c_i.

$|c_i - c_j|$: Jarak Euclidean antara *centroid cluster* i dan j.

k : Jumlah *cluster* 2 (klaster 0 dan klaster 1).

Dalam proses evaluasi menggunakan *Davies-Bouldin Index* (DBI), titik p_2 yang dikategorikan sebagai *noise* tidak disertakan dalam perhitungan. Hal ini disebabkan karena data *noise* tidak termasuk ke dalam klaster manapun, sehingga tidak dapat digunakan dalam perhitungan kohesi (*intra-cluster*) maupun separasi (*inter-cluster*) yang menjadi komponen utama dalam DBI. Oleh karena itu, perhitungan DBI hanya dilakukan pada titik-titik yang tergabung dalam klaster yang valid, yaitu klaster 0 dan klaster 1, guna

memperoleh hasil evaluasi yang lebih akurat dan representatif terhadap kualitas *clustering*.

a. Menghitung *centroid* setiap klaster

centroid = rata-rata nilai tiap fitur dari semua titik dalam klaster.

Klaster 0 = $\{p_3, p5\}$

Variabel Umur

$$= \frac{0{,}4385965 + 0{,}3859649}{2} = \frac{0{,}8245614}{2} = 0{,}4122807$$

Variabel Harga Perunit

$$= \frac{0{,}0065436 + 0{,}1053765}{2} = \frac{0{,}1119201}{2} = 0{,}0559601$$

Variabel Jumlah Perunit

$$= \frac{1{,}0000000 + 1{,}0000000}{2} = \frac{2{,}0000000}{2} = 1{,}0000000$$

Variabel Potongan Harga

$$= \frac{0{,}0000000 + 0{,}000000}{2} = \frac{0{,}0000000}{2} = 0{,}0000000$$

Variabel Total Pembayaran

$$= \frac{0{,}0111505 + 0{,}1758517}{2} = \frac{0{,}1870021}{2} = 0{,}0935011$$

Variabel Rata-rata Durasi Sesi

$$= \frac{0{,}0833333 + 0{,}2777778}{2} = \frac{0{,}3611111}{2} = 0{,}1805556$$

Variabel Jumlah Tayangan Halaman

$$= \frac{0{,}3043478 + 0{,}3913043}{2} = \frac{0{,}6956522}{2} = 0{,}3478261$$

Variabel Waktu Pengiriman

$$= \frac{0{,}1666667 + 0{,}2500000}{2} = \frac{0{,}4166667}{2} = 0{,}2083333$$

Variabel Penilaian Konsumen

$$= \frac{0{,}2500000 + 0{,}7500000}{2} = \frac{1{,}0000000}{2} = 0{,}5000000$$

Tabel 4.11 Hasil centroid c_0

Fitur	*centroid* c_0
Umur	0,4122807
Harga Perunit	0,0559601
Jumlah Perunit	1,0000000
Potongan Harga	0,0000000
Total Pembayaran	0,0935011
Rata-rata Durasi Sesi	0,1805556
Jumlah Tayangan Halaman	0,3478261
Waktu Pengiriman	0,2083333
Penilaian Konsumen	0,5000000

Klaster 1 = $\{p_1, p_4\}$

Variabel Umur

$$= \frac{0{,}1578947 + 0{,}2456140}{2} = \frac{0{,}4035088}{2} = 0{,}2017544$$

Variabel Harga Perunit

$$= \frac{0{,}0074000 + 0{,}1121451}{2} = \frac{0{,}1195450}{2} = 0{,}0597725$$

Variabel Jumlah Perunit

$$= \frac{0{,}0000000 + 0{,}0000000}{2} = \frac{0{,}0000000}{2} = 0{,}0000000$$

Variabel Potongan Harga

$$= \frac{0{,}0000000 + 0{,}1502933}{2} = \frac{0{,}1502933}{2} = 0{,}0751467$$

Variabel Total Pembayaran

$$= \frac{0{,}0024682 + 0{,}0267034}{2} = \frac{0{,}0291716}{2} = 0{,}0145858$$

Variabel Rata-rata durasi sesi

$$= \frac{0{,}0416667 + 0{,}0972222}{2} = \frac{0{,}1388889}{2} = 0{,}0694444$$

Variabel Jumlah Tayangan Halaman

$$= \frac{0{,}5652174 + 0{,}3913043}{2} = \frac{0{,}9565217}{2} = 0{,}4782609$$

Variabel Waktu Pengiriman

$$= \frac{0{,}2916667 + 0{,}0000000}{2} = \frac{0{,}2916667}{2} = 0{,}1458333$$

Variabel Penilaian Konsumen

$$= \frac{1{,}0000000 + 0{,}7500000}{2} = \frac{1{,}7500000}{2} = 0{,}8750000$$

Tabel 4.12 Hasil *centroid* c_1

Fitur	*centroid* c_1
Umur	0,2017544
Harga Perunit	0,0597725
Jumlah Perunit	0,0000000
Potongan Harga	0,0751467
Total Pembayaran	0,0145858
Rata-rata Durasi Sesi	0,0694444
Jumlah Tayangan Halaman	0,4782609
Waktu Pengiriman	0,1458333
Penilaian Konsumen	0,8750000

b. Menghitung $var(c_i)$ = Rata-rata Jarak Titik ke *centroid*

$var(c)$ dihitung sebagai rata-rata jarak setiap titik anggota klaster ke *centroid* klasternya.

Jarak *euclidean* antara *centroid cluster* i dan j dihitung dengan rumus yang sama seperti persamaan (2.7):

$var(c_0)$: Klaster 0 = $\{p_3, p_5\}$, *centroid* c_0

c_0 = rata- rata tiap fitur dari $\{p_3, p_5\}$

Tabel 4.13 Hasil Klaster 0 dan *centroid* c_0

Fitur	p_3	p_5	*centroid* c_0
Umur	0,4385965	0,3859649	0,4122807
Harga Perunit	0,0065436	0,1053765	0,0559601
Jumlah Perunit	1,0000000	1,0000000	1,0000000
Potongan Harga	0,0000000	0,0000000	0,0000000
Total Pembayaran	0,0111505	0,1758517	0,0935011
Rata-rata Durasi Sesi	0,0833333	0,2777778	0,1805556
Jumlah Tayangan Halaman	0,3043478	0,3913043	0,3478261
Waktu Pengiriman	0,1666667	0,2500000	0,2083333
Penilaian Konsumen	0,2500000	0,7500000	0,5000000

Hitung $dist(p_3, c_0)$ jarak p_3 ke *centroid* c_0

Hitung selisih tiap fitur $(p_3 - c_0)$:

Umur $= 0{,}4385965 - 0{,}4122807 = 0{,}0263158$

Harga Perunit $= 0{,}0065436 - 0{,}0559601 = -0{,}0494165$

Jumlah Perunit $= 1{,}0000000 - 1{,}0000000 = 0{,}0000000$

Potongan Harga $= 0{,}0000000 - 0{,}0000000 = 0{,}0000000$

Total Pembayaran $= 0{,}0111505 - 0{,}0935011 = -0{,}0823506$

Rata-rata Durasi Sesi $= 0{,}0833333 - 0{,}1805556 = -0{,}0972223$

Jumlah Tayangan Halaman $= 0{,}3043478 - 0{,}3478261 = -0{,}0434783$

Waktu Pengiriman $= 0{,}1666667 - 0{,}2083333 = -0{,}0416667$

Penilaian Konsumen $= 0{,}2500000 - 0{,}5000000 = -0{,}2500000$

Kuadratkan setiap selisih, jumlahkan, lalu akar:

$(0{,}0263158)^2 = 0{,}0006925$

$(-0{,}0494165)^2 = 0{,}0024420$

$(0{,}0000000)^2 = 0{,}0000000$

$(0{,}0000000)^2 = 0{,}0000000$

$(-0{,}0823506)^2 = 0{,}0067816$

$(-0{,}0972223)^2 = 0{,}0094522$

$(-0{,}0434783)^2 = 0{,}0018904$

$(-0{,}0416667)^2 = 0{,}0017361$

$(-0{,}2500000)^2 = 0{,}0625000$

Jumlah kuadrat $= 0{,}0006925 + 0{,}0024420 + 0{,}0000000 + 0{,}0000000$

$+0{,}0067816 + 0{,}0094522 + 0{,}0018904 + 0{,}0017361 + 0{,}0625000$

$= 0{,}0854948$

$$dist(p_3, c_0) = \sqrt{0{,}0854948} = 0{,}2923949$$

Hitung $dist(p_5, c_0)$ jarak p_5 ke *centroid* c_0

Hitung setiap tiap fitur $(p_5 - c_0)$

Umur $= 0{,}3859649 - 0{,}4122807 = 0{,}0263158$

Harga Perunit $= 0{,}1053765 - 0{,}0559601 = -0{,}0494165$

Jumlah Perunit $= 1{,}0000000 - 1{,}0000000 = 0{,}0000000$

Potongan Harga $= 0{,}0000000 - 0{,}0000000 = 0{,}0000000$

Total Pembayaran $= 0{,}1758517 - 0{,}0935011 = -0{,}0823506$

Rata-rata Durasi Sesi $= 0{,}2777778 - 0{,}1805556 = -0{,}0972223$

Jumlah Tayangan Halaman $= 0{,}3913043 - 0{,}3478261 = -0{,}0434783$

Waktu Pengiriman $= 0{,}2500000 - 0{,}2083333 = -0{,}0416667$

Penilaian Konsumen $= 0{,}7500000 - 0{,}5000000 = -0{,}2500000$

Kuadratkan, jumlahkan, lalu akar:

$(0{,}0263158)^2 = 0{,}0006925$

$(-0{,}0494165)^2 = 0{,}0024420$

$(0{,}0000000)^2 = 0{,}0000000$

$(0{,}0000000)^2 = 0{,}0000000$

$(-0{,}0823506)^2 = 0{,}0067816$

$(-0{,}0972223)^2 = 0{,}0094522$

$(-0{,}0434783)^2 = 0{,}0018904$

$(-0{,}0416667)^2 = 0{,}0017361$

$(-0{,}2500000)^2 = 0{,}0625000$

Jumlah kuadrat $= 0{,}0006925 + 0{,}0024420 + 0{,}0000000 +$

$0{,}0000000 + 0{,}0067816 + 0{,}0094522 + 0{,}0018904 + 0{,}0017361 +$

$0{,}0625000 = 0{,}0854948$

$dist(p_5, c_0) = \sqrt{0{,}0854948 =}\ 2923949$

$$var(c_0) = \frac{\big(dist, (p_3, c_0) + dist(p_5, c_0)\big)}{2} = \frac{(2923949 + 2923949)}{2}$$

$$= \frac{0{,}5847898}{2} = 0{,}2923949$$

$var(c_1)$: Klaster 1 = $\{p_1, p_4\}$, *centroid* c_1

c_1 = rata-rata tiap fitur dari $\{p_1, p_4\}$:

Tabel 4.14 Hasil Klaster 1 dan *centroid* c_1

Fitur	p_1	p_4	*centroid* C_1
Umur	0,1578947	0,2456140	0,2017544
Harga Perunit	0,0074000	0,1121451	0,0597725
Jumlah Perunit	0,0000000	0,0000000	0,0000000
Potongan Harga	0,0000000	0,1502933	0,0751467
Total Pembayaran	0,0024682	0,0267034	0,0145858
Rata- rata Durasi Sesi	0,0416667	0,0972222	0,0694444
Jumlah Tayangan Halaman	0,5652174	0,3913043	0,4782609
Waktu Pengiriman	0,2916667	0,0000000	0,1458333
Penilaian Konsumen	1,0000000	0,7500000	0,8750000

Hitung $dist(p_1, c_1)$ jarak p_1 ke *centroid* c_1:

Hitung setiap selisih fitur (p_1, c_1):

Umur $= 0{,}1578947 - 0{,}2017544 = -0{,}0438597$

Harga Perunit $= 0{,}1121451 - 0{,}0597725 = -0{,}0523726$

Jumlah Perunit $= 0{,}0000000 - 0{,}0000000 = 0{,}0000000$

Potongan Harga $= 0{,}0000000 - 0{,}0751467 = -0{,}0751467$

Total Pembayaran $= 0{,}0024682 - 0{,}0145858 = -0{,}0121176$

Rata-rata Durasi Sesi $= 0{,}0416667 - 0{,}0694444 = -0{,}0277777$

Jumlah Tayangan Halaman $= 0{,}5652174 - 0{,}4782609 = 0{,}0869565$

Waktu Pengiriman $= 0{,}2916667 - 0{,}1458333 = 0{,}1458334$

Penilaian Konsumen $= 1{,}0000000 - 0{,}8750000 = 0{,}1250000$

Kuadratkan, jumlahkan, lalu akar:

$(-0{,}0438597)^2 = 0{,}0019237$

$(-0{,}0523726)^2 = 0{,}0027429$

$(0{,}0000000)^2 = 0{,}0000000$

$(-0{,}0751467)^2 = 0{,}0056470$

$(-0{,}0121176)^2 = 0{,}0001468$

$(-0{,}0277777)^2 = 0{,}0007716$

$(0{,}0869565)^2 = 0{,}0075614$

$(0{,}1458334)^2 = 0{,}0212674$

$(0{,}1250000)^2 = 0{,}0156250$

Jumlah kuadrat $= 0{,}0019237 + 0{,}0027429 + 0{,}0000000 +$

$0{,}0056470 + 0{,}0001468 + 0{,}0007716 + 0{,}0075614 + 0{,}0212674 +$

$0{,}0156250 = 0{,}0556858$

$dist(p1, c_1) = \sqrt{0{,}0556858} = 0{,}2359785$

Hitung $dist(p_4, c_1)$ Jarak p_4 ke centroid c_1:

Hitung setiap selisih tiap fitur (p_4, c_1):

Umur $= 0{,}2456140 - 0{,}2017544 = 0{,}0438596$

Harga Perunit $= 0{,}1121451 - 0{,}0597725 = 0{,}0523726$

Jumlah Perunit $= 0{,}0000000 - 0{,}0000000 = 0{,}0000000$

Potongan Harga $= 0{,}1502933 - 0{,}0751467 = 0{,}0751466$

Total Pembayaran $= 0{,}0267034 - 0{,}0145858 = 0{,}0121176$

Rata-rata Durasi Sesi $= 0{,}0972222 - 0{,}0694444 = 0{,}0277778$

Jumlah Tayangan Halaman $= 0{,}3913043 - 0{,}4782609 = -0{,}0869566$

Waktu Pengiriman $= 0{,}0000000 - 0{,}1458333 = -0{,}1458333$

Penilaian Konsumen $= 0{,}7500000 - 0{,}8750000 = -0{,}1250000$

Kuadratkan, jumlahkan , lalu akar:

$(0{,}0438596)^2 = 0{,}0019237$

$(0{,}0523726)^2 = 0{,}0027429$

$(0{,}0000000)^2 = 0{,}0000000$

$(0{,}0751466)^2 = 0{,}0056470$

$(0{,}0121176)^2 = 0{,}0001468$

$(0{,}0277778)^2 = 0{,}0007716$

$(-0{,}0869566)^2 = 0{,}0075615$

$(-0{,}1458333)^2 = 0{,}0212674$

$(-0{,}1250000)^2 = 0{,}0156250$

Jumlah Kuadrat $= 0{,}0019237 + 0{,}0027429 + 0{,}0000000 + 0{,}0056470 + 0{,}0001468 + 0{,}0007716 + 0{,}0075615 + 0{,}0212674 + 0{,}0156250 = 0{,}0556858$

$dist(p_4, c_1) = \sqrt{0{,}0556858} = 0{,}2359785$

$$var(c_1) = \frac{\big(dist(p_1, c_1) + dist(p_4, c_1)\big)}{2} = \frac{(0{,}2359785 + 0{,}2359785)}{2}$$

$$= \frac{0{,}4719569}{2} = 0{,}2359784$$

c. Hitung Jarak Antar *centroid* $|c_0 - c_1|$

Hitung setiap selisih fitur $|c_0 - c_1|$:

Umur $= 0{,}4122807 - 0{,}2017544 = 0{,}2105263$

Harga Perunit $= 0{,}0559601 - 0{,}0597725 = -0{,}0038124$

Jumlah Perunit $= 1{,}0000000 - 0{,}0000000 = 1{,}0000000$

Potongan Harga $= 0{,}0000000 - 0{,}0751467 = -0{,}0751467$

Total Pembayaran $= 0{,}0935011 - 0{,}0145858 = 0{,}0789153$

Rata-rata Durasi Sesi $= 0{,}1805556 - 0{,}0694444 = 0{,}1111111$

Jumlah Tayangan Halaman $= 0{,}3478261 - 0{,}4782609 = -0{,}1304348$

Waktu Pengiriman $= 0{,}2083333 - 0{,}1458333 = 0{,}0625000$

Penilaian Konsumen $= 0{,}5000000 - 0{,}8750000 = -0{,}3750000$

Kuadratkan, lalu jumlahkan:

$(0{,}2105263)^2 = 0{,}0443213$

$(-0{,}0038124)^2 = 0{,}0000145$

$(1{,}0000000)^2 = 1{,}0000000$

$(-0{,}0751467)^2 = 0{,}0056470$

$(0{,}0789153)^2 = 0{,}0062276$

$(0{,}1111111)^2 = 0{,}0123457$

$(-0{,}1304348)^2 = 0{,}0170132$

$(0{,}0625000)^2 = 0{,}0039063$

$(-0{,}3750000)^2 = 0{,}1406250$

Jumlahkan:$0{,}0443213 + 0{,}0000145 + 1{,}0000000 + 0{,}0056470 + 0{,}0062276 + 0{,}0123457 + 0{,}0170132 + 0{,}0039063 + 0{,}1406250 = 1{,}2301007$

$|c_0 - c_1| = \sqrt{1{,}2301007} = 1{,}1090990$

Hitung R_{01} dan DBI

$var(c_0) = 0{,}2923949$

$var(c_1) = 0{,}2359784$

$var(c_0) + var(c_1) = 0{,}2923949 + 0{,}2359784 = 0{,}5283733$

$|c_0 - c_1) = 1{,}1090990$

$$R_{01} = R_{10} = \frac{0{,}5283733}{1{,}1090990} = 0{,}4763987$$

$$DBI = \frac{1}{2}(maxR_{01} + R_{10}) = \frac{1}{2}(0{,}4763987 + 0{,}4763987)$$

$$= \frac{1}{2}(0{,}9527974) = 0{,}4763987$$

$DBI = 0{,}4763987$ Nilai mendekati 0 menunjukkan klaster yang terpisah dengan baik.

Tabel 4.15 Hasil Akhir Evaluasi *Davies Bouldin Index*

Komponen	Hasil
min_sample	2
min_cluster_size	2
Jumlah klaster	2 klaster
Klaster 0	$\{p_3, p_5\}$ stabilitas = 1,5302002
Klaster 1	$\{p_1, p_4\}$ stabilitas = 0,0645306
Noise	$\{p_2\}$
Silhouette Score	0,5461644 (nilai mendekati 1 = klaster cukup terpisah baik)
Davies Bouldin Index	0,4763987 (nilai mendekati 0 = klaster terpisah baik)

Berdasarkan Tabel 4.15 hasil simulasi perhitungan manual HDBSCAN pada lima data pertama, diperoleh dua klaster dan satu data noise dengan parameter *min_sample* 2 dan *min_cluster_size* 2. Klaster 0 terdiri dari p_3 dan p_5 dengan nilai stabilitas sebesar $1{,}5302002$, sedangkan Klaster 1 terdiri dari p_1, dan p_4 dengan nilai stabilitas sebesar $0{,}0645306$. Nilai yang lebih tinggi pada Klaster 0 menunjukkan bahwa klaster tersebut lebih kuat dalam struktur hierarki dibandingkan Klaster 1. Sementara, p_2 dikategorikan sebagai *noise* karena tidak memenuhi syarat minimum ukuran klaster. Nilai *Silhouette Coeffiecient* sebesar $0{,}5461644$ menunjukkan bahwa hasil pengelompokan cukup baik, sedangkan nilai *Davies Bouldin Index* sebesar $0{,}4763987$ menunjukkan bahwa klaster yang terbentuk relatif terpisah dengan baik.

4.5 Analisis Hasil *Clustering*

Analisis hasil clustering dilakukan untuk menjelaskan proses pembentukan klaster menggunakan metode HDBSCAN serta mengevaluasi kualitas klaster yang dihasilkan. Pada penelitian ini, perhitungan manual menggunakan lima data pertama sebagai ilustrasi tahapan algoritma, sedangkan hasil clustering utama diperoleh dari keseluruhan dataset.

Tabel 4. 16 Ringkasan Perbanding *min_sample* dan DBSCAN

Perbandingan	Banyak Klaster	*Silhouette Coeffiecient*	*Davies Bouldin Index*
min_sample 5	25	0,1667	1,9822
min_sample 10	22	0,1749	1,8704
min_sample 15	18	0,1817	1,8668
min_sample 20	17	0,1868	1,7889
min_sample 25	15	0,1896	1,7502
min_sample 30	13	0,1935	1,7420
min_sample 35	13	0,1977	1,6714

Lanjutan Tabel 4.16

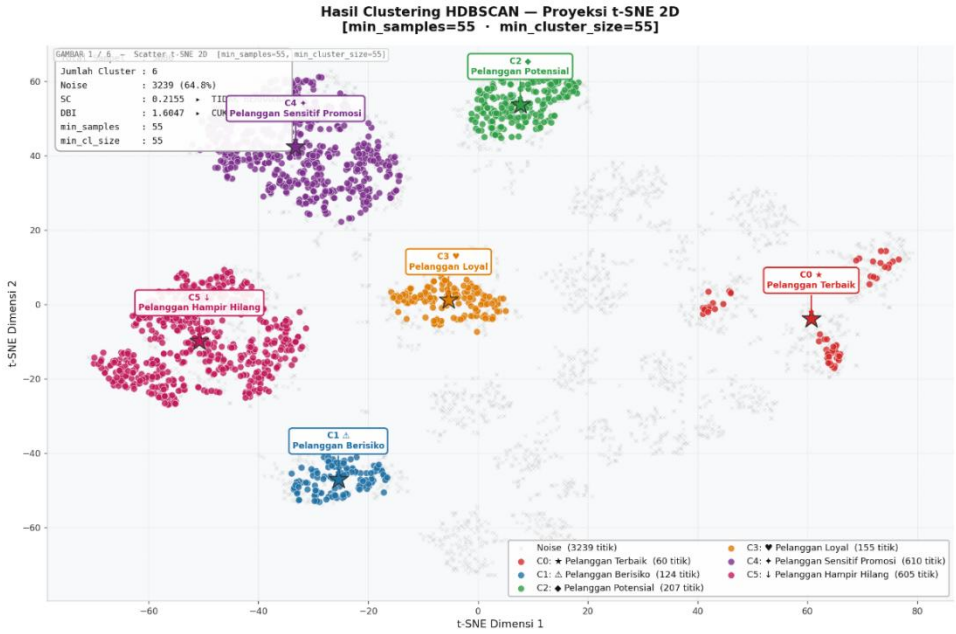
min_sample 40	12	0,1980	1,6277
min_sample 45	9	0,1952	1,7339
min_sample 50	7	0,2082	1,6611
min_sample 55	6	0,2155	1,6047
HDBSCAN	2	0,1216	2,7666

Berdasarkan Tabel 4.16 hasil penelitian parameter *min_sample*, diperoleh bahwa perubahan nilai *min_sample* berpengaruh terhadap jumlah klaster yang terbentuk dan nilai evaluasi *clustering*. Pada nilai *min_sample* yang kecil, banyak klaster yang dihasilkan cenderung lebih banyak, sehingga segmentasi terlalu menjadi telalu terpecah. Sebaliknya, Ketika *min_sample* diperbesar, jumlah klaster menjadi lebih sedikit dan struktur kelompok terlihat lebih ringkas.

Hasil terbaik diperoleh pada *min_sample* 55 dengan banyak klaster sebanyak 6 klaster. Penentuan label setiap klaster dilakukan dengan melihat nilai rata-rata variabel pada masing-masing kelompok klaster. Klaster 0 diberi label Pelanggan terbaik karena memiliki jumlah produk dan total pembayaran paling tinggi dibandingkan klaster lainnya, serta didukung oleh penilaian konsumen yang baik. Klaster 1 diberi label Pelanggan berisiko karena memiliki jumlah produk, total pembayaran, dan penilaian konsumen yang relatif rendah, sehingga menunjukkan aktivitas belanja yang lemah. Klaster 2 diberi label Pelanggan potensial karena memiliki penilaian konsumen sangat baik dan masih menunjukkan aktivitas transaksi, meskipun total pembayarannya belum menjadi yang tertinggi. Selanjutnya, Klaster 3 diberi label Pelanggan loyal karena memiliki pola belanja yang cukup stabil, dengan total pembayaran relatif baik dan penilaian konsumen yang juga baik. Klaster 4 diberi label Pelanggan Sensitif Promosi karena memiliki

nilai potongan harga paling menonjol dan jumlah anggota klaster paling besar sehingga kelomok ini cenderung berkaitan dengan konsumen yang tertarik pada diskon atau promosi. Adapun Klaster 5 diberi label Pelanggan hampir hilang karena memiliki jumlah produk yang rendah, total pembayaran yang tidak terlalu tinggi, serta waktu pengiriman yang relatif paling lama dibandingkan klaster lainnya.

Parameter *min_sample* 55 menghasilkan nilai *Silhouette Coefficient* sebesar 0,2155 dan *Davies Bouldin Index* sebesar 1,6047. Nilai *Silhouette Coefficient* tersebut merupakan nilai tertinggi dibandingkan percobaan parameter lainnya, sedangkan nilai *Davies Bouldin Index* merupakan nilai terendah. Hal ini menunjukkan bahwa hasil *clustering* pada *min_sample* 55 memiliki pemisahan klaster yang lebih baik dibandingkan parameter lainnya.



Gambar 4.4 Hasil Plot klaster *min_sample* 55

Berdasarkan Gambar 4.4 menujukkan hasil *clustering* HDBSCAN dengan *min_sample* 55 dan *min_cluster_size* 55, menghasilkan 6 klaster serta jumlah titik yanh dikategorikan sebagai *noise*. Titik-titik yang berwarna abu-abu adalah data

noise, yaitu titik yang tidak bergabung dalam klaster karena tidak memiliki kedekatan pola perilaku belanja yang cukup dengan titik lain.

Keenam klaster memiliki sebaran yang berbeda dan menggambarkan pola perilaku belanja yang spesifik. Klaster 0 (Pelanggan Terbaik) berada pada area dengan total pembayaran dan aktivitas pembelian yang tinggi. Klaster 1 (Pelanggan Berisiko) menempati area dengan aktivitas belanja rendah dan penilaian konsumen rendah. Klaster 2 (Pelanggan Potensial) menunjukkan konsumen dengan penilaian tinggi namun nilai transaksinya belum maksimal. Klaster 3 (Pelanggan Loyal) menggambarkan konsumen dengan pola belanja stabil. Klaster 4 (Pelanggan Sensisitif Promosi) cenderung berada pada area dengan potongan harga tinggi, sedangkan Klaster 5 (Pelanggan Hampir Hilang) berada pada area dengan aktivitas belanja rendah dan waktu pengiriman relatif tinggi.

Oleh karena itu, parameter *min_sample* 55 dipilih sebagai parameter utama dalam proses *clustering* menggunakan HDBSCAN. Pemilihan ini didasarkan pada hasil evaluasi yang menunjukkan bahwa klaster yang terbentuk lebih baik dari segi kedekatan anggota dalam klaster maupun pemisahan antar klaster.

4.6 Pembahasan

Berdasarkan hasil analisis *clustering*, parameter terbaik pada metode HDBSCAN diperoleh pada *min_sample* 55. Parameter tersebut menghasilkan 6 klaster dengan nilai *Silhouette Coefficient* sebesar 0,2155 dan *Davies Bouldin Index* sebesar 1,6047. Nilai ini menunjukkan bahwa hasil *clustering* pada parameter tersebut lebih baik dibandingkan percobaan nilai *min_sample* lainnya, karena memiliki nilai *Silhouette Coefficient* paling tinggi dan *Davies Bouldin Index* paling rendah.

Hasil optimal HDBSCAN kemudian dibandingkan dengan metode DBSCAN. Pada metode DBSCAN, klaster yang terbentuk hanya 2 klaster dengan nilai *Silhouette Coefficient* sebesar 0,1216 dan *Davies Bouldin Index* sebesar 2,7666. Jika dibandingkan, nilai *Silhouette Coefficien*t HDBSCAN lebih tinggi daripada DBSCAN, sedangkan nilai *Davies Bouldin Index* lebih rendah. Hal ini menunjukkan bahwa HDBSCAN mampu membentuk kelompok konsumen yang lebih baik dibandingkan DBSCAN, karena hasil klasternya lebih terpisah dan tidak terlalu saling tumpang tindih.

Perbedaan hasil antara HDBSCAN dan DBSCAN dapat terjadi karena data perilaku belanja konsumen tidak selalu sama. Ada konsumen yang sangat aktif berbelanja, ada yang cukup aktif, dan ada juga yang jarang berbelanja. Perbedaan pola tersebut membuat data memiliki tingkat kepadatan yang berbeda-beda. DBSCAN kurang sesuai untuk data dengan tingkat kepadatan yang berbeda-beda karena menggunakan satu batas jarak yang sama untuk seluruh data, sedangkan HDBSCAN dapat menyesuaikan pembentukan klaster berdasarkan perbedaan kepadatan tersebut. Oleh karen itu, HDBSCAN lebih sesuai digunakan karena mampu menghasilkan klaster yang lebih menggambarkan variasi perilaku belanja konsumen.

4.7 Kajian Integrasi Al-Qur'an Terhadap Segmentasi E-commerce Berdasarkan Perilaku Belanja

Dalam segmentasi konsumen *e-commerce* berdasarkan perilaku belanja, konsumen loyal merupakan kelompok yang memiliki kecenderungan melakukan pembelian secara berulang pada suatu platform atau penjual tertentu. Perilaku ini terbentuk melalui pengalaman transaksi yang konsisten dan memuaskan, seperti

kesesuaian antara deskripsi dan kualitas produk, ketepatan pengiriman, serta pelayanan yang responsif. Loyalitas tidak hanya dipengaruhi oleh faktor harga atau promosi, tetapi lebih pada tingkat kepercayaan konsumen terhadap penjual. Dalam konteks ini, kepercayaan menjadi elemen utama yang membedakan konsumen loyal dengan segmen lainnya. Dalam perspektif Islam, hubungan antara penjual dan pembeli tidak hanya bersifat ekonomi, tetapi juga mengandung nilai moral berupa amanah, yaitu kepercayaan yang harus dijaga dan ditunaikan secara bertanggung jawab dalam setiap transaksi.

Dalam perspektif Islam, hal ini berkaitan dengan prinsip amanah yang harus dijaga dalam setiap aktivitas muamalah. Al-Qur'an menegaskan hal tersebut dalam firman Allah SWT pada Q.S. An-Nisa ayat 58:

اِنَّ اللّٰهَ يَأْمُرُكُمْ اَنْ تُؤَدُّوا الْاَمٰنٰتِ اِلٰٓى اَهْلِهَاۙ وَاِذَا حَكَمْتُمْ بَيْنَ النَّاسِ اَنْ تَحْكُمُوْا بِالْعَدْلِۗ اِنَّ اللّٰهَ نِعِمَّا
يَعِظُكُمْ بِهٖۗ اِنَّ اللّٰهَ كَانَ سَمِيْعًاۢ بَصِيْرًا ﴿٥٨﴾

Terjemahan:
"Sesungguhnya Allah menyuruh kamu menyampaikan amanah kepada pemiliknya. Apabila kamu menetapkan hukum di antara manusia, hendaklah kamu tetapkan secara adil. Sesungguhnya Allah memberi pengajaran yang paling baik kepadamu. Sesungguhnya Allah Maha Mendengar lagi Maha Melihat." (QS. An-Nisa' [4]:58)
(Kementrian Agama, 2022)

Ayat tersebut, dapat dipahami bahwa prinsip amanah memiliki peran mendasar dalam menjaga kualitas interaksi dalam aktivitas ekonomi, termasuk dalam sistem *e-commerce*. Amanah mencerminkan kewajiban untuk bersikap jujur, adil, dan bertanggung jawab dalam setiap transaksi. Dalam konteks perilaku belanja, penerapan amanah oleh penjual akan menghasilkan pengalaman transaksi yang positif, yang kemudian membentuk persepsi dan kepercayaan konsumen.

Kepercayaan ini menjadi dasar terbentuknya pola perilaku tertentu, seperti kecenderungan melakukan pembelian ulang atau mempertahankan loyalitas terhadap suatu penjual.

Dalam penafsiran para ulama seperti dalam Tafsir Ibnu Katsir, Q.S. An-Nisa ayat 58 dijelaskan sebagai perintah Allah SWT kepada manusia untuk menunaikan amanah kepada yang berhak secara adil dan tidak menyalahgunakannya. Amanah dalam ayat ini mencakup segala bentuk tanggung jawab, baik yang berkaitan dengan hak individu maupun dalam hubungan sosial, termasuk dalam aktivitas muamalah seperti transaksi ekonomi. Ibnu Katsir menekankan bahwa setiap bentuk kepercayaan yang diberikan harus dijaga dengan penuh tanggung jawab, karena hal tersebut menjadi dasar terciptanya keadilan dan keteraturan dalam kehidupan manusia.

Dalam konteks implementasi metode *clustering* HDBSCAN dalam segmentasi konsumen *e-commerce* berdasarkan perilaku belanja, konsep amanah tersebut dapat dipahami sebagai faktor yang memengaruhi terbentuknya pola perilaku konsumen dalam data transaksi. Ketika penjual mampu menjaga amanah melalui kejujuran informasi, kualitas produk, dan pelayanan yang konsisten, maka akan terbentuk kepercayaan konsumen. Kepercayaan ini kemudian tercermin dalam perilaku belanja yang stabil, seperti kecenderungan melakukan pembelian berulang dan mempertahankan hubungan dengan penjual tertentu. Dalam proses *clustering* menggunakan HDBSCAN, pola perilaku yang konsisten ini akan membentuk kelompok data dengan tingkat kepadatan tinggi, sehingga teridentifikasi sebagai suatu klaster yang jelas, yang dapat diinterpretasikan sebagai segmen konsumen loyal.

Sebaliknya, jika amanah tidak dijaga dalam transaksi, kepercayaan konsumen dapat menurun dan memengaruhi perilaku belanja menjadi tidak konsisten. Konsumen dapat berpindah-pindah, jarang melakukan pembelian, atau berhenti bertransaksi.

Dalam perspektif Islam, amanah merupakan prinsip utama yang menjadi landasan dalam setiap aktivitas ekonomi, termasuk dalam perilaku belanja pada *e-commerce*. Nilai amanah tidak hanya membentuk hubungan yang adil dan terpercaya antara penjual dan konsumen, tetapi juga memengaruhi pola perilaku konsumen yang tercermin dalam data transaksi. Oleh karena itu, hasil segmentasi konsumen melalui metode *clustering* HDBSCAN tidak hanya menunjukkan pengelompokan berdasarkan kemiripan perilaku belanja, tetapi juga dapat dipahami sebagai representasi dari tingkat kepercayaan, tanggung jawab, dan kualitas interaksi yang sesuai dengan nilai-nilai etika dalam Islam.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa metode HDBSCAN mampu digunakan untuk melakukan segmentasi konsumen *e-commerce* berdasarkan perilaku belanja. Proses implementasi dilakukan melalui normalisasi *Min-Max Scaler*, perhitungan jarak *Euclidean*, penerapan algoritma HDBSCAN, serta evaluasi menggunakan *Silhouette Coefficient* dan *Davies Bouldin Index*. Parameter terbaik diperoleh pada *minimum sample* 55 dan *minimum cluster size* 55, dengan hasil 6 klaster, nilai *Silhouette Coefficient* sebesar 0,2155, dan *Davies Bouldin Index* sebesar 1,6047. Hasil *clustering* menujukkan bahwa HDBSCAN mampu membentuk enam kelompok konsumen dengan karakteristik perilaku belanja yang berbeda.

a. Klaster 0 merupakan Pelanggan terbaik, yaitu konsumen dengan aktivitas yang paling kuat.
b. Klaster 1 Pelanggan Berisiko, yaitu konsumen dengan aktivitas belanja dan penilaian yang relatif rendah.
c. Klaster 2 merupakan Pelanggan potensial, yaitu konsumen yang memiliki penilaian terbaik dan masih berpeluang untuk dikembangkan menjadi pelanggan bernilai tinggi.
d. Klaster 3 merupakan Pelanggan loyal, yaitu konsumen yang menunjukkan pola belanja cukup stabil.
e. Klaster 4 merupakan Pelanggan sensitif promosi, yaitu konsumen yang cenderung berkaitan dengan potongan harga atau promo.

f. Klaster 5 merupakan Pelanggan hampir hilang, yaitu konsumen dengan aktivitas pembelian yang rendah dan perlu mendapat perhatian lebih.

Oleh karena itu, HDBSCAN mampu melakukan segmentasi *konsumen e-commerce* karena dapat mengelompokkan konsumen yang memiliki kemiripan perilaku belanja, seperti total pembayaran, jumlah perunit, potongan harga, durasi sesi, waktu pengiriman, dan penilaian konsumen. Konsumen dengan pola yang mirip akan tergabung dalam klaster yang sama, sedangkan konsumen yang polanya terlalu beda dapat dikategorikan sebagai *noise*.

5.2 Saran untuk Penelitian Lanjutan

Penelitian selanjutnya disarankan menggunakan dataset *e-commerce* yang berasal dari transaksi nyata perusahaan atau platform tertentu, sehingga hasil segmentasi dapat menggambarkan perilaku konsumen secara lebih aktual. Selain itu, variabel yang digunakan dapat diperluas dengan menambahkan informasi seperti frekuensi pembelian, jenis produk yang sering dibeli, metode pembayaran, riwayat kunjungan, dan waktu pembelian. Penambahan variabel tersebut diharapkan dapat menghasilkan segmentasi konsumen yang lebih detail dan sesuai dengan kondisi perilaku belanja yang sebenarnya.

Penelitian selanjutnya juga dapat melakukan pengujian parameter HDBSCAN secara lebih luas dan membandingkannya dengan metode *clustering* lain, seperti K-Means, atau Agglomerative Clustering. Hal ini bertujuan untuk melihat metode mana yang paling sesuai dalam membentuk klaster pada data perilaku belanja konsumen. Selain itu, hasil klaster yang diperoleh sebaiknya dikembangkan ke dalam strategi pemasaran, seperti pemberian promosi, rekomendasi produk, atau

program loyalitas yang disesuaikan dengan karakteristik masing-masing segmen konsumen.

DAFTAR PUSTAKA

Alemán, González- R., Platero-Rochart, D., Rodríguez-Serradet, A., Hernández-Rodríguez, E. W., Caballero, J., Leclerc, F., & Montero-Cabrera, L. (2022a). MDSCAN: RMSD-based HDBSCAN clustering of long molecular dynamics. *Bioinformatics*, *38*(23), 5191–5198.

Alwi, B., & Muliono, R. (2025). Analisis klustering menggunakan algoritma DBSCAN untuk deteksi anomali dalam data transaksi keuangan. INCODING: Journal of Informatics and Computer Science Engineering, 5(1), 121–133.

Amory, J. D. S., Mudo, M., & J, R. (2025). Transformasi Ekonomi Digital dan Evolusi Pola Konsumsi: Tinjauan Literatur tentang Perubahan Perilaku Belanja di Era Internet. *Jurnal Minfo Polgan*, *14*(1), 28–37.

Apriyanto, B., & Sitio, S. L. M. (2025). Penerapan K-Means dalam Menganalisis Pola Pembelian Pelanggan Pada Data Transaksi E-Commerce. *Bit-Tech*, *7*(3), 790–797.

Bukhari, & Muslim. (2017). *Hadist Shahih BUKHARI MUSLIM*. http://pustaka-indo.blogspot.com

Bushra, A. A., Kim, D., Kan, Y., & Yi, G. (2024). AutoSCAN: automatic detection of DBSCAN parameters and efficient clustering of data in overlapping density regions. *PeerJ Computer Science*, *10*

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2013). A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery*, *27*(3), 344–371.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, *10*(1).

Chandrasekaran, K. S., Jeevanantham, T., Jemima Blessy, R., & Foumiya, Z. J. (2023). Customer segmentation using UMAP and HDBSCAN. International Journal of Multidisciplinary Research Transactions, 5(7), 117–136.

Fadilah Aswar, N. (2025). Perilaku konsumen. Klaten, Jawa Tengah: Tahta Media.

Febrianty, E., Awalina, L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail Optimizing Marketing Strategies with Customer Segmentation Using K-Means Clustering on Online Retail Transactions. *Jurnal Teknologi Dan Informasi (JATI)*, *13*.

Fu, King. Sun, Lu, S. Y., Davies, D. L., & Bouldin, D. W. (1977). The string-to-string correction problem. Journal of the Association for Computing Machinery, 24(2).

Ghiffary Ghaly, G., Alifviansyah, K., Fitrianto, A., Risman, L. M., & Jumansyah, D. (2024). Perbandingan Algoritma Hdbscan Dan Agglomerative Hierarchical Clustering Dalam Klasterisasi Pada Data Yang Mengandung Pencilan. *Jurnal Riset Dan Aplikasi Matematika*, *08*(02), 122–135.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Waltham, MA: Morgan Kaufmann Publishers.

Hapsari, R. K., Sumarni, T., Hidayatulloh, T., Farida, Herianto, Elfaladonna, F., Aini, N., Nurfitria, Kurniawan, H., Kraugusteeliana, Yusuf, L., & Widiatama, Y. (2024). Data mining. Indie Press.

Hunt, E. L., & Reffert, S. (2021). Improving the open cluster census: I. Comparison of clustering algorithms applied to Gaia DR2 data. *Astronomy and Astrophysics*, *646*

Ismail, H. A., Trimiati, E., & Prihati, Y. (2020). Membangun model konseptual faktor sinergitas perilaku konsumen dalam konteks pembelian impulsive secara online. Al-Tijarah, 6(3), 10–20.

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, *2*(4), 809–823.

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kementrian Agama. (2022). *Al-Qur'an dan Terjemahan. Lajnah Pentashihan Mushaf Al-Qur'an Badan Litbag Kementrian Agama RIhttps://quran.kemenag.go.id/*.

Larasati, M., Nasrudin, & Tojiri, Y. (2024). E-commerce dan transformasi pemasaran: Strategi menghadapi era digital. Padang, Sumatera Barat: Takaza Innovatix Labs.

Lestari, M. D., & Iswari, L. (2023). Identifikasi hotspot area dan waktu penjemputan taksi menggunakan spasial clustering berbasis densitas. Jurnal Teknik Informatika (JUTIF), 4(5), 1135–1142.

Levitin, A. (2012). Introduction to the Design and Analysis of Algorithms (3rd ed.). Boston, MA: Pearson Education, Inc.

Malzer, C., & Baum, M. (2020). A hybrid approach to hierarchical density-based cluster selection. 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 223–228.

McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. The Journal of Open Source Software, 2(11), 205.

Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-based clustering validation. *SIAM International Conference on Data Mining 2014, SDM 2014*, *2*, 839–847.

Mukhlis Juliyanto, R., Patria, F., Danggo Pamungkas, H., & Surya Nugraha, F. (2024). Penerapan algoritma clustering untuk segmentasi pelanggan berdasarkan perilaku pembelian. Prosiding Seminar Nasional Amikom Surakarta (SEMNASA) 2024, 117.

Mutiah, S., Hasnataeni, Y., Fitrianto, A., Erfiani, E., & Jumansyah, L. M. R. D. (2024). Perbandingan Metode Klastering K-Means dan DBSCAN dalam Identifikasi Kelompok Rumah Tangga Berdasarkan Fasilitas Sosial Ekonomi di Jawa Barat. *Teorema: Teori Dan Riset Matematika*, *9*(2), 247.

Natania, A. T., & Dwijayanti, R. (2024). Pemanfaatan platform digital sebagai sarana pemasaran bagi UMKM. Jurnal Pendidikan Tata Niaga (JPTN), 12(1), 1–8.

Novy, N. P., Dewi, C., Aditia, D., & Nasution, D. (2023). Jurnal Pijar Studi Manajemen dan Bisnis Pentingnya Penerapan E-Commerce Bagi Umkm Sebagai Salah Satu Bentuk Pemasaran Digital Dalam Menghadapi Revolusi Industri 4.0. 1(3), 566–577.

Nurjannah & Ratnah S. (2024). Analisis Penggunaan E-Commerce Dalam Meningkatkan Usaha Mikro Kecil Menengah Di Kota Makassar (Studi Kasus Omorfoo Shop). *Jurnal Rumpun Manajemen Dan Ekonomi*, *1*(4), 100–112.

Perdana, N., & Santoso, H. (2025). Klasterisasi judul berita online isu Pemilu Prabowo Subianto dengan kombinasi LLMS embedding dengan HDBSCAN. Journal Locus Penelitian dan Pengabdian, 4(10), 9284–9298.

Perdana Satria Ardi 1), dan Suryandari A. S., A. S., & Amalia, R. N. (2022). Analisis Segmentasi Pelanggan Menggunakan K-Means Clustering Studi Kasus Aplikasi Alfagift. *Sebatik*, *26*(2), 420–427.

Purnomo, N., & Gata, W. (2024). Pengelompokan analisis sentimen komentar YouTube terhadap pengambilalihan jalan rusak di Lampung menggunakan algoritma clustering. Progresif: Jurnal Ilmiah Komputer, 20(2).

Ramkumar, G., Bhuvaneswari, J., Venugopal, S., Kumar, S., Ramasamy, C. K., & Karthick, R. (2025). Enhancing customer segmentation: RFM analysis and K-Means clustering implementation. In S. P. J. Christydass, N. Nurhayati, & S. Kannadhasan (Eds.), Hybrid and advanced technologies (pp. 70–76). Taylor & Francis. https://doi.org/10.1201/9781003559139-9

Rizki, A., Pratama, Y., Mirza, A. H., Nasir, M., & Udariansyah, D. (2023). Penerapan Metode Clustering Dalam Analisis Data Produk Laptop di Tokopedia untuk Membuat Rekomendasi Produk: Studi Perbandingan K-Means, Hierarchical, dan DBSCAN. *Media Online)*, *99*(99), 999–999.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics* (Vol. 20).

Rusvinasari, D., & Annisa, L. H. (2025). Klasterisasi Pola Penjualan Menu Makanan pada Rumah Makan menggunakan Metode K-Means Clustering. *Jurnal Informatika: Jurnal Pengembangan IT*, *10*(2), 398–409.

Shahih, A.-B. (2107), & Muslim (1531)], 2008, h. 388. (2020). *Khiyār Al-majlis Dan Aplikasinya Dalam Jual Beli Modern (Studi Komparatif Antara Jumhur Ulama Dan Imam Malik)*.

Soetiyani, A., Lukiyana, L., Ariandi, A., Kamaruddin, M. J., & Jundi, M. (2024). Sosialisasi: Manfaat E-commerce Untuk Meningkatkan Penjualan Pada UMKM Obat. *Yumary: Jurnal Pengabdian Kepada Masyarakat*, *5*(1), 69–77.

Syah, F., Wiyono, P., Kaffi, L., Maulana, M. H., Damaliana, A. T., Sella, S., & Wara, M. (2025). Segmentasi wilayah provinsi di Indonesia berdasarkan indeks penanganan stunting menggunakan PCA dan partition clustering. Prosiding Seminar Nasional Sains Data, 5(1), 406–417.

Tuygur, U. (2023–2024). E-commerce customer behavior and sales analysis – TR (Turkish synthetic e-commerce data) [Dataset]. Kaggle. https://www.kaggle.com/datasets/umuttuygurr/e-commerce-customer-behavior-and-sales-analysis-tr/data

Wardhana, A. (2024). Perilaku konsumen di era digital. Purbalingga, Jawa Tengah: Eureka Media Aksara.

Wororomi, K J., Felix Reba, M., Aleksander Mandowen, S., Alvian Sroyer, M. M., Halomoan Edy Manurung, M., Mayko Edison Koibur, M., Erna Hudianti Pujiarini, M., Marwan Ramdhany Edy, M., Hajiar Yuliana, M., Mukarramah Yusuf, M., & Marta Dinata, R. (2024). Data mining: Memahami pola di balik angka. Purbalingga, Jawa Tengah: Eureka Media Aksara.

LAMPIRAN

Lampiran 1 *Dataset E-commerce*

https://docs.google.com/spreadsheets/d/1jcVMC6w_qn2zhvjeQjy7HMLWhla3zbu/edit?usp=sharing&ouid=103490621727501536402&rtpof=true&sd=true	

Lampiran 2 *Dataset* setelah *Preprocessing*

https://docs.google.com/spreadsheets/d/1mW0IU73zSDMWfcofUrJQV0SNlRbaMnvU/edit?usp=sharing&ouid=103490621727501536402&rtpof=true&sd=true	

Lampiran 3 Hasil *Clustering* HDBSCAN dengan *min_sample* 55

https://docs.google.com/spreadsheets/d/1DGTq75qdQO9TZRxoou49L62QGfq42Jt/edit?usp=sharing&ouid=103490621727501536402&rtpof=true&sd=true	

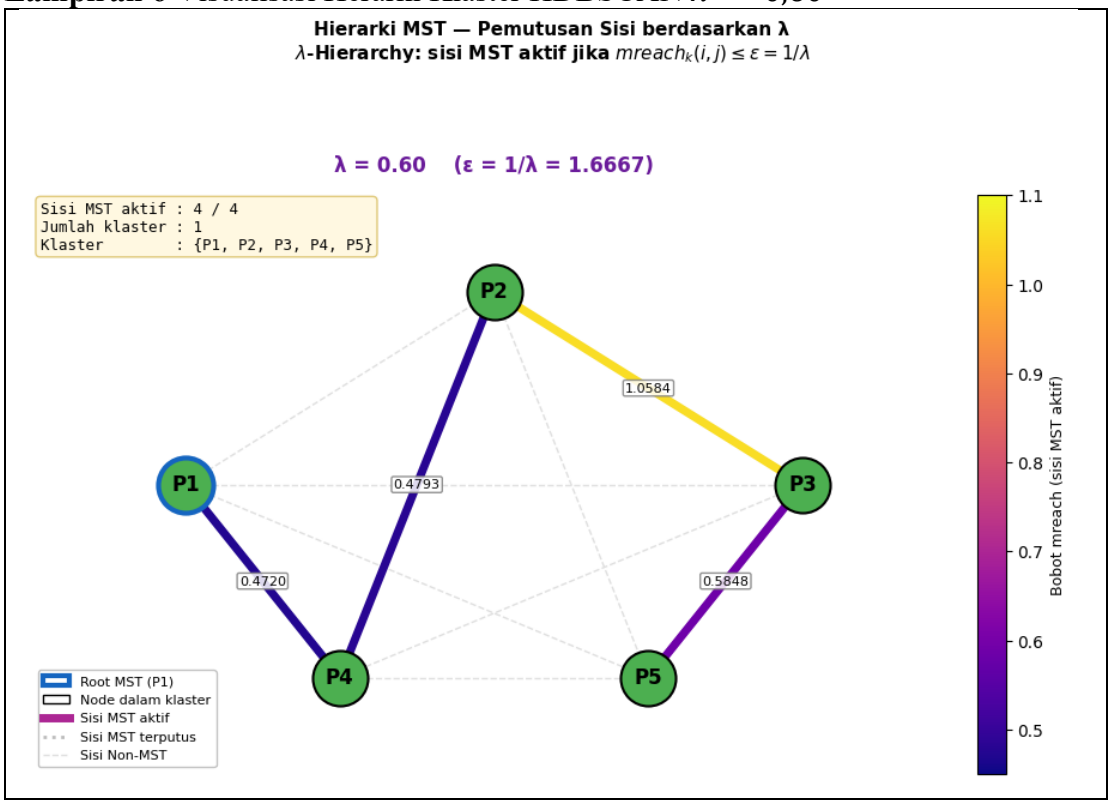
Lampiran 4 *Source Code* metode HDBSCAN dengan *min_sample* 55

https://colab.research.google.com/drive/1ZFAiPbyOjUDutNTP821uzNb_kyx9vnI0?usp=sharing	

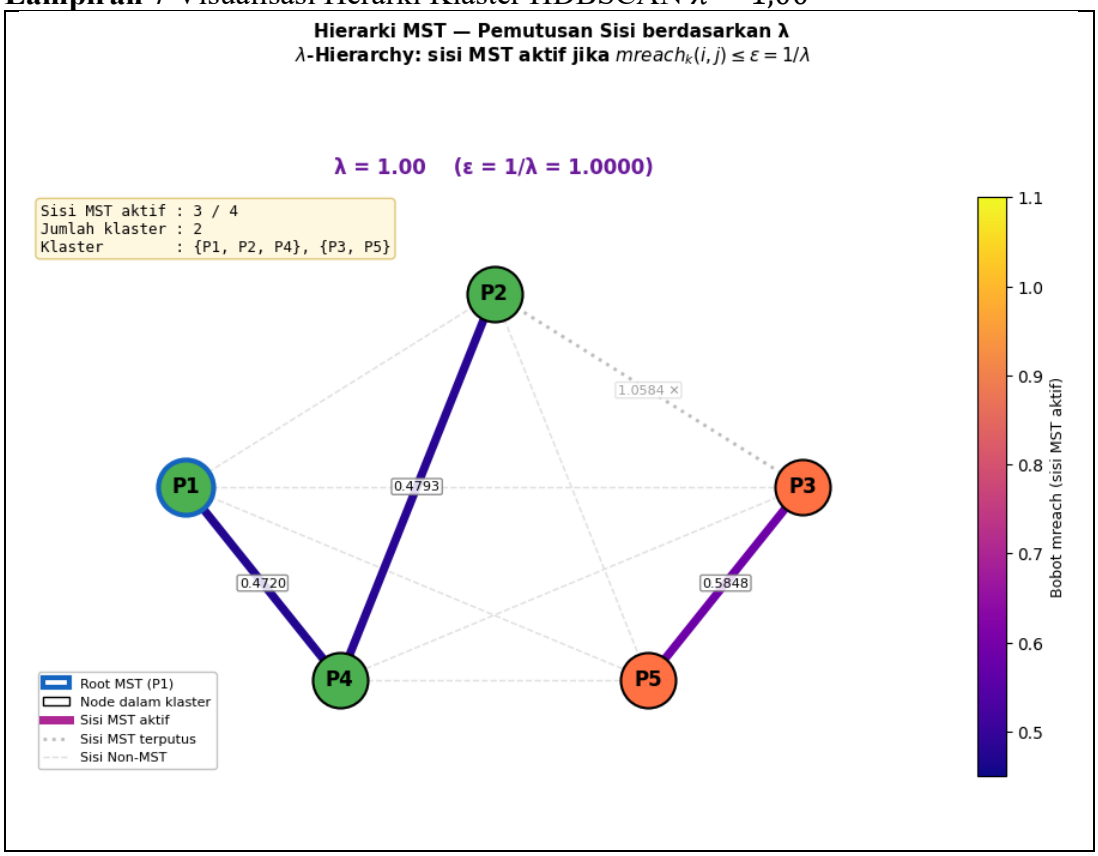
Lampiran 5 *Source Code* Visualisasi Hasil *Clustering* dengan *min_sample* 55

https://colab.research.google.com/drive/1ouiCbXIbQM9uAJvnAAIrKmVDbL6cOxIq?usp=sharing	

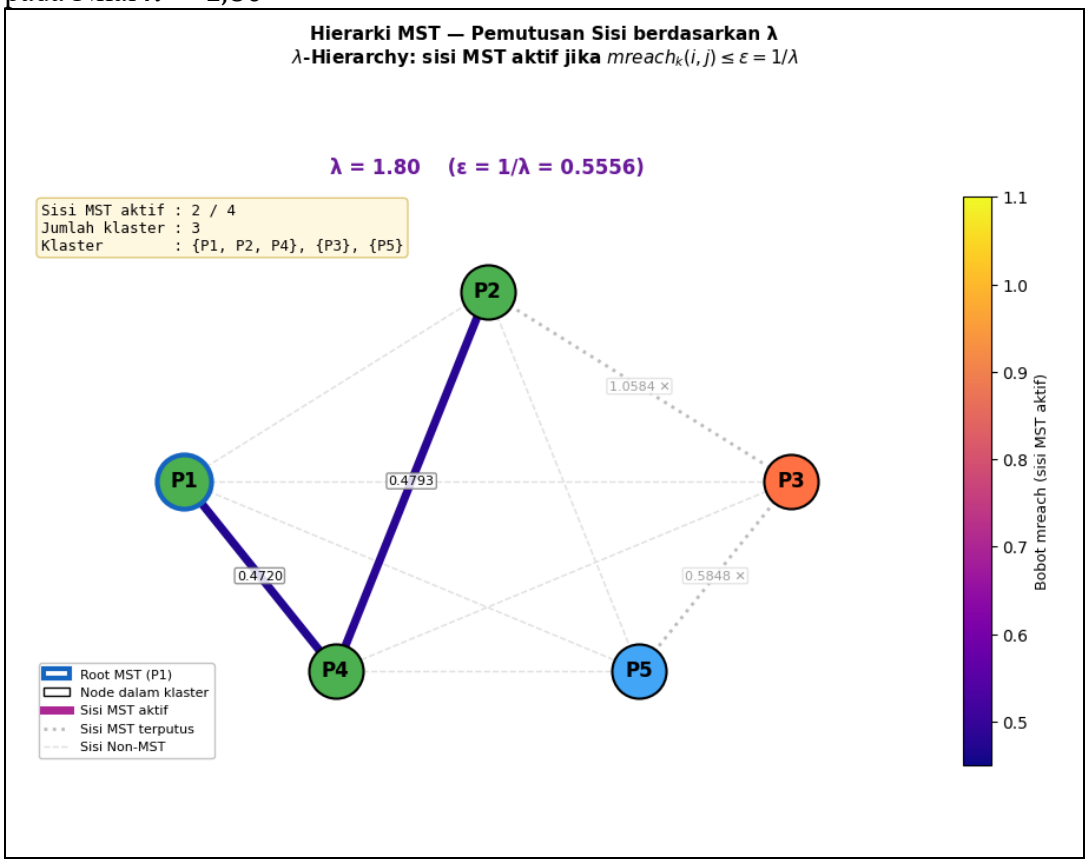
Lampiran 6 Visualisasi Herarki Klaster HDBSCAN $\lambda = 0{,}60$



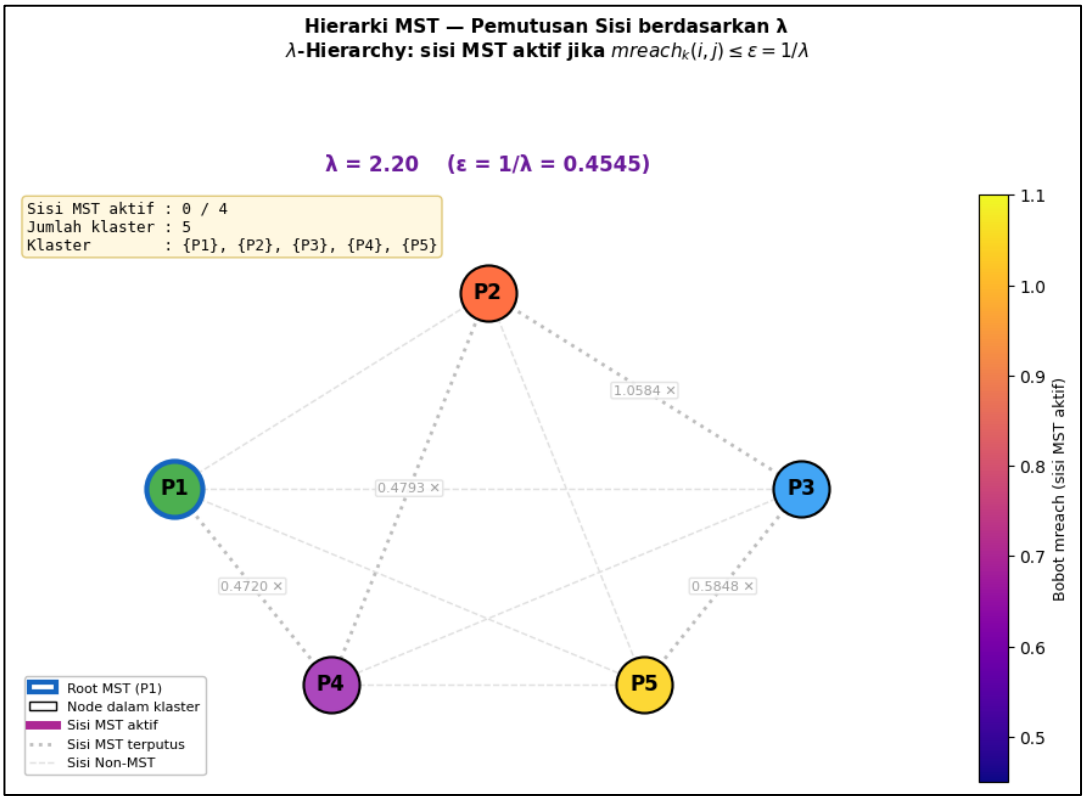
Lampiran 7 Visualisasi Herarki Klaster HDBSCAN $\lambda = 1{,}00$



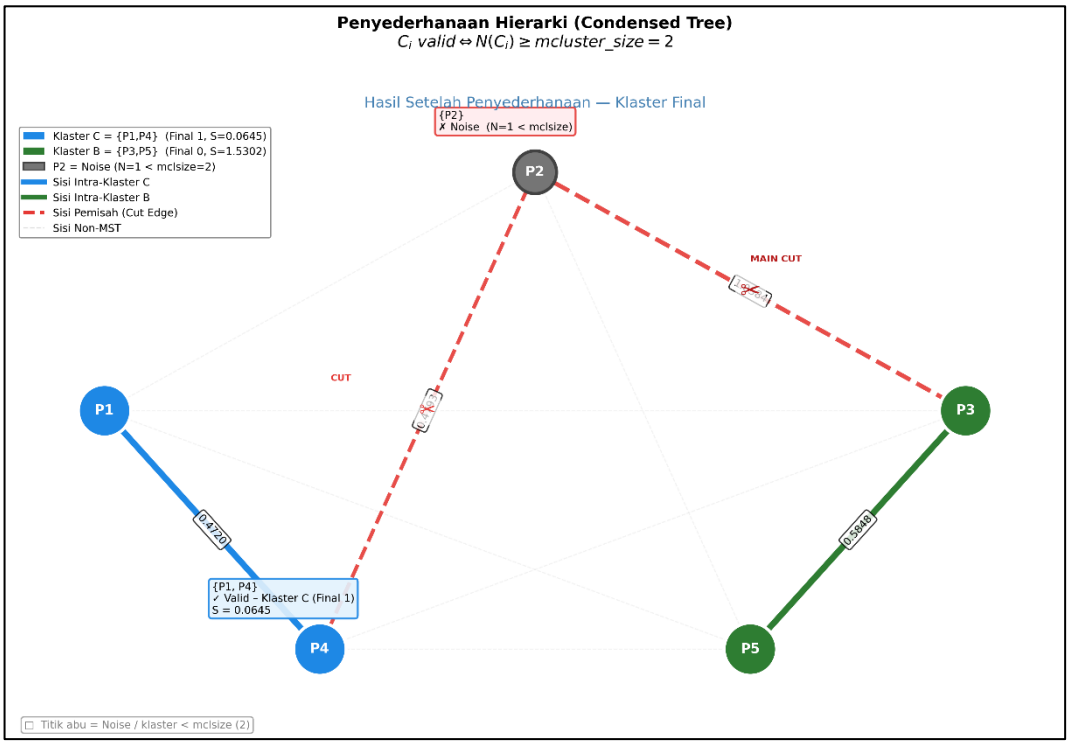
Lampiran 8 Visualisasi Pemisahan Lanjutan Hierarki Klaster HDBSCAN pada Nilai $\lambda = 1{,}80$



Lampiran 9 Visualisasi Pemisahan Akhir Hierarki Klaster HDBSCAN pada Nilai $\lambda = 2{,}20$



Lampiran 10 Visualisasi Klaster Akhir Setelah Penyederhanaan Hierarki HDBSCAN



RIWAYAT HIDUP

Rossima Eva Yuliana, lahir di Bekasi pada 12 Juli 2004. Penulis merupakan anak pertama dari dua bersaudara. Penulis menempuh pendidikan dimulai dari TK Negeri Pembina dan lulus pada tahun 2010. Selanjutnya, penulis melanjutkan pendidikan sekolah dasar di SDN Karang Asih 12 dan lulus pada tahun 2016. Pendidikan kemudian dilanjutkan ke SMPN 1 Cikarang Utara dan diselesaikan pada tahun 2019. Setelah itu, penulis melanjutkan ke jenjang sekolah menengah atas di MA Al-Hikmah Puwoasri, Kediri dan lulus pada tahun 2022. Pada tahun yang sama, penulis melanjutkan pendidikan di Universitas Islam Negeri Maulana Malik Ibrahim Malang pada Program Studi Matematika, Fakultas Sains dan Teknologi. Selama menempuh pendidikan di perguruan tinggi, penulis mengikuti UKM Unit Olahraga (UNIOR) sebagai bentuk pengalaman menjadi anggota dan pengurus untuk mengembangkan diri menjadi mahasiswa.Selama menempuh perguruan tinggi penulis telah melakukan Kegiatan Kerja Mahasiswa (KKM) di Wagir Kabupaten Malang. Selain itu, penulis juga telah melaksanan Praktik Kerja Lapangan (PKL) di Badan Narkotika Nasional (BNN) Kabupaten Malang.

KEMENTERIAN AGAMA RI
UNIVERSITAS ISLAM NEGERI
MAULANA MALIK IBRAHIM MALANG
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Gajayana No. 50 Malang 65144 Telp. / Fax. (0341) 558933

BUKTI KONSULTASI SKRIPSI

Nama : Rossima Eva Yuliana
NIM : 220601110068
Fakultas / Jurusan : Sains dan Teknologi / Matematika
Judul Skripsi : Implementasi Metode *Clustering* HDBSCAN dalam Segmentasi Konsumen E-*commerce* Berdasarkan Perilaku Belanja
Dosen Pembimbing I : Hisyam Fahmi, M.Kom
Dosen Pembimbing II : Mohammad Nafie Jauhari, M.Si

No	Tanggal	Hal	Tanda Tangan
1.	27 Agustus 2025	Konsultasi Topik dan Data	1.
2.	12 September 2025	Konsultasi Topik dan Data	2.
3.	23 September 2025	Konsultasi Bab I, II, dan III	3.
4.	29 September 2025	Konsultasi Bab I, II, dan III	4.
5.	07 Oktober 2025	Konsultasi Bab I, II, dan III	5.
6.	09 Oktober 2025	Konsultasi Kajian Agama Bab I dan II	6.
7.	13 Oktober 2025	Konsultasi Bab I, II, dan III	7.
8.	14 Oktober 2025	ACC Bab I, II, dan III	8.
9.	16 Oktober 2025	Konsultasi Kajian Agama Bab I dan II	9.
10.	16 Oktober 2025	ACC Kajian Agama Bab I dan II	10.
11.	04 November 2025	ACC Seminar Proposal	11.
12.	20 November 2025	Konsultasi Revisi Seminar Proposal	12.
13.	24 Desember 2025	Konsultasi Bab IV dan V	13.
14.	17 April 2026	ACC Bab IV dan V	14.
15.	20 April 2026	Konsultasi Kajian Agama Bab IV	15.
16.	20 April 2026	Konsultasi Kajian Agama Bab IV	16.

KEMENTERIAN AGAMA RI
UNIVERSITAS ISLAM NEGERI
MAULANA MALIK IBRAHIM MALANG
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Gajayana No. 50 Malang 65144 Telp. / Fax. (0341) 558933

17.	20 April 2026	ACC Kajian Agama Bab IV	17.
18.	29 April 2026	ACC Seminar Hasil	18.
19.	6 Mei 2026	Konsultasi Revisi Seminar Hasil	19.
20.	12 Mei2026	Konsultasi Revisi Seminar Hasil	20.
21.	21 Mei 2026	ACC Sidang Skripsi	21.
22.	4 Juni 2026	ACC Keseluruhan	22.

Malang, 4 Juni 2026
Mengetahui,
Ketua Program Studi Matematika

Dr. Fachrur Rozi, M.Si
NIP. 19800527 200801 1 012