EKSPLORASI DAN OPTIMALISASI ALGORITMA KLASTERING UNTUK SEGMENTASI PELANGGAN PADA *E-COMMERCE*

SKRIPSI

Oleh:

ADITYA PUTRA SUGIARTA

20211000011

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2025

EKSPLORASI DAN OPTIMALISASI ALGORITMA KLASTERING UNTUK SEGMENTASI PELANGGAN PADA *E-COMMERCE*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Buddhi Dharma Jenjang Pendidikan Strata 1

Oleh:

ADITYA PUTRA SUGIARTA

20211000011

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2025

LEMBAR PERSEMBAHAN

"Gagal bukanlah akhir dari segalanya, masih ada hal yang dapat dilakukan untuk menciptakan keberhasilan"

Dengan mengucap puji syukur kepada Tuhan Yang Maha Esa, SKRIPSI ini kupersembahkan untuk:

1. Papa Naga Sugiarta Karlim dan Mama Ariati Adiwijaya tercinta yang telah membesarkan aku dan selalu membimbing, mendukung, memotivasi, memberi apa yang terbaik bagiku serta selalu mendoakan aku untuk meraih kesuksesanku.
2. Kakak Yuliastati Putri Sugiarta Karlim dan kakak Deviastati Putri Sugiarta Karlim yang telah memberikan dukungan semangat dan finansial serta dorongan yang senantiasa diberikan sehingga menambah motivasi dan semangat.
3. Bapak Aditiya Hermawan M.Kom yang telah memberikan bimbingan, ilmu, dan arahan yang begitu berharga selama masa Pendidikan hingga selesainya skripsi ini.
4. Teman-teman anak bimbingan Bapak Aditiya Hermawan yang telah membantu dalam penulisan skripsi.
5. Teman-teman kelompok belajar, Adisty Maharani dan Grediyanto yang selalu berjuang bersama.
6. Teman-teman semasa SMP yang telah memberikan semangat dan dukungan agar perkuliahan dapat diselesaikan bersama-sama.
7. Yang terakhir, saya berterima kasih kepada diri sendiri yang masih mau berjuang dan menyelesaikan skripsi dengan banyaknya pikiran (*overthinking)* akan lulus tidaknya, dan atas kerja keras untuk menyelesaikannya dalam waktu yang ditentukan.

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan di bawah ini.

NIM	: 20211000011
Nama	: Aditya Putra Sugiarta
Jenjang Studi	: Strata 1
Program Studi	: Teknik Informatika
Peminatan	: Data Science

Dengan ini saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik Sarjana atau kelengkapan studi, baik di Universitas Buddhi Dharma maupun di Pereguruan Tinggi lainnya.
2. Skripsi ini saya buat sendiri tanpa bantuan dari pihak lain, kecuali arahan dosen pembimbing.
3. Dalam Skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan daftar pustaka.
4. Dalam Skripsi ini tidak terdapat pemalsuan (kebohongan), seperti buku, artikel, jurnal, data sekunder, pengolahan data, dan pemalsuan tanda tangan dosen atau Ketua Program Studi Universitas Buddhi Dharma yang dibuktikan dengan keasliannya.
5. Lembar pernyataan ini saya buat dengan sesungguhnya, tanpa paksaan dan apabila dikemudian hari atau pada waktu lainnya terdapat penyimpangan dan ketidak benaran dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar akademik yang telah saya peroleh karena Skripsi ini serta sanksi lainnya sesuai dengan peraturan dan norma yang berlaku.

Tangerang, 04-08-2025
Yang membuat pernyataan,

Aditya Putra Sugiarta
20211000011

LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini.

NIM	: 20211000011
Nama	: Aditya Putra Sugiarta
Jenjang Studi	: Strata 1
Program Studi	: Teknik Informatika
Peminatan	: Data Science

Dengan ini menyetujui untuk memberikan ijin kepada pihak Universitas Buddhi Dharma, Hak Bebas Royalti Non – Eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah kami yang berjudul: "Eksplorasi dan Optimalisasi Algoritma Klastering untuk Segmentasi Pelanggan pada *E-Commerce*", beserta alat yang diperlukan (apabila ada). Dengan Hak Bebas Royalti Non – Eksklusif ini pihak Universitas Buddhi Dharma berhak menyimpan, mengalih-media atau format-kan, mengelolanya dalam pangkalan data (*database*), mendistribusikannya, dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis pertama atau pencipta karya ilmiah tersebut. Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Universitas Buddhi Dharma, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.
Demikian pernyataan ini saya buat dengan sebenarnya.

Tangerang, 04-08-2025
Yang membuat pernyataan,

Aditya Putra Sugiarta
20211000011

LEMBAR PENGESAHAN PEMBIMBING

EKSPLORASI DAN OPTIMALISASI ALGORITMA KLASTERING UNTUK SEGMENTASI PELANGGAN PADA *E-COMMERCE*

Dibuat Oleh:

NIM : 20211000011

Nama : Aditya Putra Sugiarta

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Data Science

Tahun Akademik 2024/2025

Disahkan oleh,

Tangerang, 23-06-2025

Pembimbing,

(Aditiya Hermawan, S.Kom., M.Kom., M.M)

NUPTK. 9538766667130223

LEMBAR PENGESAHAN SKRIPSI
EKSPLORASI DAN OPTIMALISASI ALGORITMA KLASTERING UNTUK SEGMENTASI PELANGGAN PADA E-COMMERCE

Dibuat Oleh:

NIM : 20211000011

Nama : Aditya Putra Sugiarta

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Data Science

Tahun Akademik 2024/2025

Disahkan oleh,

Tangerang, 04-08-2025

Dekan,	**Ketua Program Studi,**
Dr. Yakub, M.Kom., M.M.	Hartana Wijaya, M.Kom.
NUPTK. 1836747648130172	NUPTK. 3844759660131192

LEMBAR PENGESAHAN TIM PENGUJI

Nama	:	Aditya Putra Sugiarta
NIM	:	20211000011
Fakultas	:	Sains dan Teknologi
Judul Skripsi	:	Eksplorasi dan Optimalisasi Algoritma Klastering untuk Segmentasi Pelanggan pada *E-Commerce*

Dinyatakan LULUS setelah mempertahankan di depan Tim Penguji pada hari Senin, 04-08-2025.

		Nama Penguji:	Tanda Tangan:
Ketua Sidang	:	Benny Daniawan, M.Kom NUPTK. 8756768669130412	.....
Penguji I	:	Edy, ST., M.Kom NUPTK. 7560760661130203	.....
Penguji II	:	Aditiya Hermawan, S.Kom., M.Kom., M.M NUPTK. 9538766667130223	.....

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Dr. Yakub, M.Kom., M.M

NUPTK. 1836747648130172

KATA PENGANTAR

Dengan mengucapkan Puji Syukur kepada Tuhan Yang Maha Esa, yang telah memberikan Rahmat dan karunia-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan SKRIPSI ini dengan judul **Eksplorasi dan Optimalisasi Algoritma Klastering untuk Segmentasi Pelanggan pada *E-Commerce***. Tujuan utama dari pembuatan Skripsi ini adalah sebagi salah satu syarat kelengkapan dalam menyelesaikan program pendidikan Strata 1 Program Studi Teknik Informatika di Universitas Buddhi Dharma. Dalam penyusunan Skripsi ini penulis banyak menerima bantuan dan dorongan baik moril maupun materiil dari berbagai pihak, maka pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ibu Dr. Limajatini, S.E., M.M., B.K.P, sebagai Rektor Universitas Buddhi Dharma
2. Bapak Dr. Yakub, S.Kom., M.Kom., M.M, Dekan Fakultas Sains dan Teknologi
3. Bapak Hartana Wijaya, S.Kom., M.Kom, sebagai Ketua Program Studi Teknik Informatika
4. Bapak Aditiya Hermawan, S.Kom., M.Kom, sebagai pembimbing yang telah membantu dan memberikan dukungan serta harapan untuk menyelesaikan penulisan SKRIPSI ini.
5. Orang tua dan keluarga yang selalu memberikan dukungan baik moril maupun materiil.
6. Teman-teman yang selalu membantu dan memberikan semangat

Serta semua pihak yang terlalu banyak untuk disebutkan satu-persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan SKRIPSI ini masih belum sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga SKRIPSI ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Tangerang, 04-08-2025

Penulis

Eksplorasi dan Optimalisasi Algoritma Klastering untuk Segmentasi Pelanggan pada *E-Commerce*
90 + xv / 28 tabel / 20 gambar / 1 lampiran

ABSTRAK

Pertumbuhan *e-commerce* yang pesat menyebabkan meningkatnya *volume* data pelanggan, sehingga diperlukan strategi segmentasi yang tepat untuk memahami perilaku konsumen secara lebih mendalam. Sehingga perusahaan dapat merancang strategi pemasaran yang tepat untuk mempertahankan serta mengembangkan keuntungan bisnis. Namun, metode klastering tradisional seperti *K-Means* masih memiliki keterbatasan dalam menghasilkan klaster optimal karena sifatnya yang sensitif terhadap titik awal pusat klaster sehingga mudah terjebak pada solusi lokal. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi dan membandingkan efektivitas lima metode klastering berbasis optimasi, yaitu *K-Means* dengan *Adaptive Learning Particle Swarm Optimization* (KM-ALPSO), *Improved Harris Hawk Optimization* dengan *K-Means* (IHHO-KM), dan *Sine Cosine Algorithm* dengan *K-Means* (SCA-KM), serta pendekatan *Density Peak–Genetic Algorithm–Possibilistic Fuzzy C-Means* (DP-GA-PFCM), dan metode *Hierarchical Clustering* sebagai pembanding. Proses klastering akan dilakukan dengan menggunakan masing-masing algoritma pada *dataset* yang sama, kemudian dievaluasi menggunakan *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*. Ketiga evaluasi akan dirata-ratakan untuk dibandingkan. Hasil penelitian menunjukkan bahwa algoritma SCA-KM memberikan hasil terbaik dengan nilai evaluasi rata-rata tertinggi sebesar 0,7393272, sedikit lebih baik dari KM-ALPSO dan menunjukkan nilai yang lebih baik dari algoritma *K-Means*. Dapat disimpulkan bahwa pendekatan *hybrid* antara metode klastering dan algoritma optimasi dapat meningkatkan kualitas segmentasi pelanggan secara signifikan, serta dapat menjadi solusi yang efektif dalam pengambilan keputusan berbasis data pada bisnis *e-commerce*.

Kata Kunci*: E-Commerce*, Klastering, *Machine Learning*, Optimisasi, Segmentasi Pelanggan

Exploration and Optimization of Clustering Algorithms for Customer Segmentation in E-Commerce
90 + xv / 28 tables / 20 figures / 1 appendix

ABSTRACT

The rapid growth of e-commerce has led to an increase in customer data volume, necessitating an appropriate segmentation strategy to gain a deeper understanding of consumer behavior. This enables companies to design effective marketing strategies to maintain and grow business profits. However, traditional clustering methods such as K-Means still have limitations in producing optimal clusters due to their sensitivity to the initial cluster center points, making them prone to getting stuck in local solutions. Therefore, this study aims to evaluate and compare the effectiveness of five optimization-based clustering methods: K-Means with Adaptive Learning Particle Swarm Optimization (KM-ALPSO), Improved Harris Hawk Optimization (IHHO-KM), and Sine Cosine Algorithm (SCA-KM), as well as the Density Peak–Genetic Algorithm –Possibilistic Fuzzy C-Means (DP-GA-PFCM) approach, and the Hierarchical Clustering method as a comparison. The clustering process will be performed using each algorithm on the same dataset, then evaluated using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The three evaluations will be averaged for comparison. The research results show that the SCA-KM algorithm provides the best results with the highest average evaluation score of 0.7393272, slightly better than KM-ALPSO and showing better results than the K-Means algorithm. It can be concluded that the hybrid approach combining clustering methods and optimization algorithms can significantly improve customer segmentation quality and serve as an effective solution for data-driven decision-making in e-commerce businesses.

Key word: E-Commerce; Clustering; Customer Segmentation; Machine Learning; Optimization

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan teknologi yang pesat, semakin banyak perusahaan yang mulai memasuki dunia *e-commerce* guna mencapai keuntungan yang lebih tinggi. Dalam upaya tersebut, segmentasi pelanggan menjadi penting untuk mengkategorikan pelanggan berdasarkan karakteristik dan perilaku mereka (Rajput & Singh, 2023). Segmentasi pelanggan memungkinkan perusahaan untuk membagi basis pelanggan menjadi kelompok yang lebih homogen berdasarkan karakteristik, preferensi, dan perilaku pembelian, sehingga strategi pemasaran dapat dirancang lebih efektif (Kuo et al., 2023). Melalui segmentasi ini, perusahaan dapat meningkatkan kualitas pelayanan produk atau jasa, meningkatkan tingkat konversi, serta memperkuat hubungan dengan pelanggan. Dengan memahami berbagai segmen konsumen, perusahaan dapat mengalokasikan sumber daya secara efektif dan efisien, serta mencapai tujuan bisnis yang lebih optimal (Li et al., 2021).

Untuk menghasilkan segmentasi yang efektif, perusahaan perlu mengumpulkan dan menganalisis data secara menyeluruh, termasuk mengidentifikasi variabel-variabel relevan seperti waktu yang dihabiskan pelanggan dan total pengeluaran dalam setahun (Kuo et al., 2023). Penerapan teknik analisis yang tepat sangat penting untuk mengelompokkan pelanggan dengan akurat dan dapat menggambarkan karakteristik mereka secara mendalam (Rajput & Singh, 2023). Metode yang digunakan harus mampu menangkap keragaman dan kompleksitas data pelanggan, sehingga hasil segmentasi menjadi andal dan relevan. Dengan pendekatan yang sistematis dan berbasis data, perusahaan dapat menciptakan segmentasi pelanggan yang bermakna dan bermanfaat.

Pendekatan yang umum digunakan dalam segmentasi pelanggan meliputi metode *k-means* (Cullu et al., 2023; Gumasing et al., 2023; Li et al., 2021; Prastyabudi et al., 2024; Rajput & Singh, 2023; Singh et al., 2020; Tabianan et al., 2022; Uddin et al., 2024; Wu, 2024; Zhang & Wu, 2024), *fuzzy c-means* (Kuo et al., 2020, 2021, 2023; Setiawan et al., 2022; Uddin et al., 2024), dan *hierarchical clustering* (Afzal et al., 2024; Uddin et al., 2024). Algoritma *K-Means* populer karena banyak digunakan untuk mengelompokkan data ke dalam beberapa klaster berdasarkan kesamaan data (Li et al., 2021), *Fuzzy c-means* dapat menangkap ketidakpastian atau ambiguitas dalam data, yang sering kali lebih mencerminkan kondisi nyata dibandingkan klasterisasi konvensional yang bersifat kaku (Kuo et al., 2020), sedangkan *hierarchical clustering* menyediakan representasi visual dari hubungan antara setiap klaster (Li et al., 2021).

Setiap algoritma segmentasi memiliki tingkat akurasi dan *erro*r yang berbeda. Oleh karena itu, untuk meningkatkan kualitas segmentasi, digunakan berbagai algoritma optimisasi seperti *Improved Harris Hawks Optimization* (IHHO) yang merupakan pengembangan dari algoritma *Harris Hawk Optimization* (HHO) yang terinspirasi dari perilaku berburu burung *Harris Hawk* (Wu, 2024), *Adaptive Learning Particle Swarm Optimization* (ALPSO) (Li et al., 2021) merupakan salah satu variasi dari *Particle Swarm Optimization* (PSO) (Kuo et al., 2020, 2023; Li et al., 2021) yang telah terbukti lebih baik daripada algoritma optimisasi PSO dasar, *Sine Cosine Algorithm* (SCA) adalah metode optimasi yang menggunakan fungsi *sine* dan *cosine* untuk mendekati solusi optimal dengan pergerakan acak dalam ruang pencarian (Kuo et al., 2020), dan *Genetic Algorithm* (GA) merupakan metode optimasi berbasis evolusi yang meniru prinsip seleksi alam untuk mencari solusi optimal pada masalah kompleks. Melalui proses seleksi, *crossover*, dan mutasi pada populasi kromosom (Kuo et al., 2020, 2021). Algoritma-algoritma optimisasi ini dapat digunakan untuk meningkatkan akurasi, mengurangi tingkat *error*, dan

menghasilkan segmentasi yang lebih akurat serta memaksimalkan efektivitas strategi pemasaran yang diterapkan.

Optimasi-optimasi ini digunakan dengan menggabungkan algoritma-algoritma klastering sehingga mendapatkan hasil yang lebih baik dari algoritma klastering dasar. Seperti KM-ALPSO, DP-GA-PFCM, KM-IHHO, dan SCA-PFCM yang digunakan untuk mencari tahu mana algoritma yang memberikan hasil terbaik, yang dapat membantu meningkatkan akurasi, mengurangi tingkat *error*, dan menghasilkan segmentasi yang lebih akurat untuk memaksimalkan efektivitas strategi pemasaran.

Penelitian yang berjudul " Eksplorasi dan Optimalisasi Algoritma Klastering untuk Segmentasi Pelanggan pada *E-Commerce*" melakukan perbandingan algoritma-algoritma klastering yang optimal yang dilakukan pada penelitian sebelumnya untuk menentukan metode klastering yang paling efektif dalam segmentasi pelanggan. Penelitian ini menekankan pentingnya segmentasi pelanggan sebagai salah satu langkah strategis yang dapat meningkatkan efektivitas pemasaran dan efisiensi alokasi sumber daya perusahaan. Dengan melakukan pendekatan berbasis data dan penggunaan algoritma yang tepat serta algoritma optimisasi yang tepat, perusahaan dapat memperoleh wawasan mendalam mengenai karakteristik dan preferensi pelanggan. Hal ini tidak hanya akan membantu dalam membangun hubungan yang lebih kuat dengan pelanggan, tetapi juga berkontribusi dalam mencapai tujuan bisnis yang berkelanjutan. Sehingga diharapkan jika kombinasi-kombinasi teknologi analitik dengan metode optimisasi yang inovatif diharapkan akan terus mendorong perkembangan strategi segmentasi pelanggan yang semakin akurat dan relevan untuk mendukung keberhasilan perusahaan dalam menghadapi tantangan di era digital.

1.2 Identifikasi Masalah

Pada penelitian ini dilakukan segmentasi pelanggan yang menggunakan beberapa algoritma klastering dengan algoritma optimasi, sehingga dapat dirumuskan masalah sebagai berikut:

a. Dibutuhkannya teknik segmentasi yang mampu menangkap karakteristik dan perilaku pelanggan dengan akurat untuk mengelompokkan pelanggan secara efektif.
b. Algoritma klastering dasar memiliki keterbatasan dalam memberikan hasil segmentasi yang optimal sehingga diperlukannya algoritma optimasi tambahan.
c. Tingkat akurasi dan *error* pada segmentasi pelanggan masih memerlukan peningkatan untuk mendukung pengambilan keputusan yang lebih baik.

1.3 Ruang Lingkup

Ruang Lingkup penelitian ini adalah sebagai berikut:

a. Data sekunder berjudul *"Ecommerce Customer"* yang digunakan pada sebuah jurnal Penelitian *"Customer Segmentation of E-commerce data using K-means Clustering Algorithm"* (Rajput & Singh, 2023)
b. Dataset memiliki 8 atribut, yang berupa *E-mail, Address, Avatar, Avg. Session Length, Time on App, Time on Website, Length of Membership,* dan *Yearly Amount Spent.*
c. Pengembangan aplikasi berbasis *web* untuk menerapkan hasil segmentasi

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

a. Mengembangkan teknik segmentasi yang mampu menangkap karakteristik dan perilaku pelanggan secara akurat untuk mengelompokkan pelanggan dengan lebih efektif.
b. Meningkatkan kualitas hasil segmentasi melalui penerapan algoritma klastering yang dikombinasikan dengan algoritma optimasi tambahan.
c. Mengurangi tingkat *error* dan meningkatkan akurasi segmentasi pelanggan untuk mendukung pengambilan keputusan yang lebih tepat.

1.4.2 Manfaat

Manfaat yang dapat dicapai dalam penelitian ini adalah:

a. Membantu perusahaan memahami pelanggan dengan lebih baik melalui segmentasi yang lebih akurat dan relevan.
b. Memperbaiki hasil klasterisasi pelanggan sehingga strategi pemasaran dapat dirancang dengan lebih efektif dan terarah.
c. Meningkatkan keandalan analisis pelanggan guna mendukung pengambilan keputusan bisnis yang lebih baik.

1.5 Sistematika Penulisan

Sistematika penulisan laporan ini terdiri dari lima bab yang disusun secara terstruktur.

BAB I PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang penelitian, identifikasi masalah, ruang lingkup penelitian, tujuan dan manfaat penelitian, serta sistematika penulisan untuk memberikan gambaran isi laporan secara keseluruhan.

BAB II LANDASAN TEORI

Bab ini menguraikan landasan teori yang menjadi dasar penelitian, seperti konsep dan prinsip umum yang relevan, serta teori khusus terkait algoritma klastering dan optimisasi yang dijelaskan secara mendalam.

BAB III METODELOGI PENELITIAN

Bab ini mencakup kerangka pemikiran penelitian, proses pengolahan data, serta metode dan algoritma yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan spesifikasi *hardware* dan *software* yang digunakan, disertai dengan tampilan program hasil segmentasi. Selain itu, ditampilkan hasil penerapan metode dan algoritma serta pengujian menggunakan *blackbox testing*.

BAB V SIMPULAN DAN SARAN

Bab ini menyajikan kesimpulan penelitian yang merangkum hasil utama serta memberikan saran untuk pengembangan lebih lanjut.

BAB II

LANDASAN TEORI

2.1 Data

Data adalah kumpulan informasi atau fakta yang digunakan untuk berbagai tujuan, seperti analisis, pengambilan keputusan, dan penelitian yang mencangkup nilai-nilai berupa rata-rata ataupun nilai tengah dari sebuah data (Umar Hussain, 2023). Data sendiri terbagi menjadi 3 kategori data, Data Terstruktur, Data Semi-Terstruktur, dan Data Tidak-Terstruktur. Data terstruktur adalah data yang sudah memiliki format tetap seperti tabel sehingga mudah diakses dan dianalisis, Data semi-terstruktur adalah data yang elemen untuk pengorganisasian seperti tag atau format markup tetapi tidak terikat pada format tetap seperti tabel sehingga lebih fleksibel, data tidak terstruktur adalah data yang tidak memiliki format tertentu, sehingga sulit untuk diolah secara langsung serta mencakup data seperti dokumen, gambar, *video*, atau *audio* yang memerlukan teknologi tertentu untuk dianalisis lebih mendalam (Breitinger & Jotterand, 2023).

2.2 *Data Mining*

Data Mining adalah proses untuk mengekstrak pengetahuan, pola, dan wawasan dari kumpulan data yang melibatkan perpaduan teknik dari statistik, *Machine-Learning*, *Database System*, dan *Artificial Intelligence* untuk mendapatkan relasi dan tren yang tidak terlihat dari pengamatan yang sederhana (Lasisi & Kolade, 2025). Metode dalam *Data Mining* terbagi menjadi 5 metode, yaitu : klastering, klasifikasi, regresi, prediksi, dan asosiasi. Dimana setiap metode memiliki algoritma-algoritma tersendiri yang banyak digunakan dalam bidang edukasi. Dalam bidang edukasi, data mining berfokus pada mengembangkan, meneliti, dan mengaplikasikan metode untuk mendeteksi pola dalam data

edukasional yang besar yang dimana sulit untuk di analisa berdasarkan jumlah data yang besar (Papadogiannis et al., 2024).

2.3 *Website*

Website adalah kumpulan dokumen berupa halaman *web* berisi teks dengan format *Hyper Text Markup Language* (HTML) yang dapat diakses pada sebuah browser menggunakan *Hyper Text Transfer Protocol* (HTTP) Atau melalui HTTP *Secure* (HTTPS) melalui alamat *Internet* berupa *Uniform Resource Locator* (URL) (Widia & Asriningtias, 2021). Website terbagi menjadi 2 jenis, *website* statis dan *website* dinamis. *Website* Statis adalah *website* yang bersifat tetap, dimana isi informasinya berupa searah, hanya bisa diubah oleh pemilik *website* dan hanya dapat diubah melalui *Source Code*, sedangkan *website* dinamis adalah *website* yang bersifat interaktif dua arah, dimana isi informasinya dapat diubah oleh pemilik *website* dan pengguna *website* (Noviantoro et al., 2022).

2.4 Segmentasi Pelanggan

Segmentasi pelanggan adalah area penelitian yang terus berkembang dan akan semakin meningkat dengan ketersediaan dan aksesbilitas terhadap data *online*, untuk memahami pelanggan serta memprediksi perilaku dan pola pelanggan menjadi hal yang dapat digunakan untuk mempertahankan bisnis (Rajput & Singh, 2023). Segmentasi pelanggan sudah banyak digunakan untuk mendapatkan keuntungan lebih dan mempertahankan bisnis yang dijalankan, segmentasi pelanggan harus dilakukan secara efektif dan efisien guna mendapatkan hasil yang baik. Dalam bidang *retail* sudah banyak studi melakukan segmentasi pelanggan yang berfokus pada beberapa teknik klastering sumber data (Afzal et al., 2024).

2.5 Klastering

Klastering adalah proses mengelompokkan *item-item* yang mirip dan memisahkan *item* menjadi kategori yang berbeda (Li et al., 2021). Klastering banyak digunakan untuk

membuat analisa pengelompokkan terutama untuk bidang pembelajaran. Klastering sendiri terbagi menjadi *Hard-Klastering* dan *Soft-Klastering*, dimana *Hard-Klastering* mengelompokkan data berdasarkan angka, seperti *K-Means*, sedangkan *Soft-Klastering* mengelompokkan data berdasarkan probabilitas, sehingga 1 data bisa berada pada 2 klaster atau lebih (Uddin et al., 2024).

2.6 *K-Means*

K-Means adalah sebuah algoritma klastering yang mengelompokkan data pada *dataset* menjadi jumlah K klaster untuk mendefinisikan data (Rajput & Singh, 2023). Setiap klaster memiliki titik pusat yang dapat didefinisikan dengan berbagai cara, seperti nilai rata-rata dari setiap objek-objek dalam klaster (Li et al., 2021). *K-Means* memiliki rumus perhitungan sebagai berikut:

$$f = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

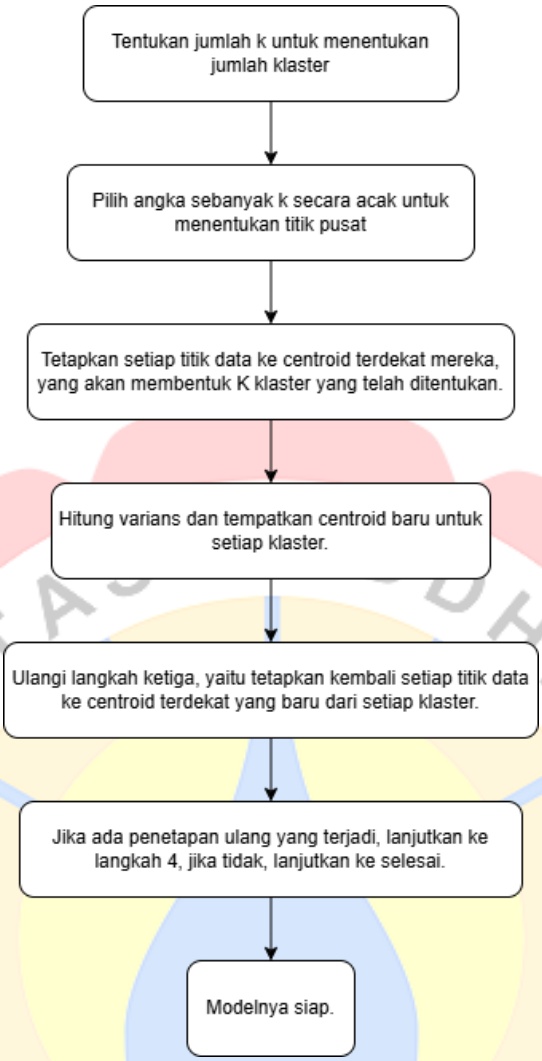
Dimana,

$x_i^{(j)}$ adalah titik data yang dipilih

c_j adalah pusat klaster data

$\left\| x_i^{(j)} - c_j \right\|^2$ adalah jarak antara pusat klaster dan titik data.

Untuk melakukan *K-Means* ada langkah-langkah algoritma yang terbagi menjadi beberapa langkah:



Gambar 2. 1 Langkah *K-Means*
Sumber: (Rajput & Singh, 2023)

2.7 *Possibilistic Fuzzy C-Means*

Possibilistic Fuzzy C-Means (PFCM) adalah algoritma kombinasi dari *Fuzzy C-Means* dan *Possibilistic C-Means*, dimana PFCM memiliki 2 jenis anggota, anggota *Possibilistic* mengukur derajat absolut dari tipikalitas suatu titik dalam sebuah klaster tertentu, dan anggota *Fuzzy* yang mengukur derajat relatif dalam berbagi titik di antara klaster-klaster (Kuo et al., 2023).

Langkah 1: Inisialisasi matriks partisi menggunakan Persamaan 2

$$u_{ik} = \begin{bmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nm} \end{bmatrix} \quad (2).$$

Langkah 2: Hitung matriks tipikalitas. Selain derajat keanggotaan atau matriks partisi, matriks tipikalitas juga diperlukan dalam perhitungan algoritma PFCM. Matriks tipikalitas merupakan nilai tipikalitas yang digunakan untuk menentukan titik data mana yang termasuk ke dalam klaster tertentu. Matriks tipikalitas dihitung menggunakan Persamaan 3 sebagai berikut:

$$t_{ik} = \left[1 + \left(\frac{b\left[\sum_{j=1}^{m}\left(X_{ij} - V_{iuj}\right)^2\right]}{\gamma_k}\right)^{\frac{1}{\eta-1}}\right] \quad (3)$$

di mana γ_k dihitung sebagai berikut:

$$\gamma_k = K\frac{\sum_{i=1}^{n}\left((\mu_{ik})^w\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]\right)}{\sum_{i=1}^{n}(\mu_{ik})^w} \quad (4)$$

Langkah 3: Hitung pusat klaster. Setelah menentukan matriks partisi dan matriks tipikalitas, pusat klaster dihitung sebagai berikut:

$$V_k = \frac{\sum_{i=1}^{n}(a{\mu_{ik}}^w + b{t_{ik}}^{\eta})X_{ij}}{\sum_{i=1}^{n}(a{\mu_{ik}}^w + b{t_{ik}}^{\eta})} \quad (5)$$

Langkah 4: Hitung fungsi objektif. Fungsi objektif dari PFCM ditunjukkan pada Persamaan (6):

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}(a{\mu_{ik}}^w + b{t_{ik}}^{\eta})\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right] + \sum_{k=1}^{c}\gamma_k\sum_{i=1}^{n}(1 - t_{ik})^{\eta} \quad (6)$$

Konstanta **'a'** dan **'b'** didefinisikan sebagai tingkat kepentingan relatif keanggotaan fuzzy dan nilai tipikalitas dalam fungsi objektif.

Langkah 5: Perbarui matriks partisi dan matriks tipikalitas. Jika nilai fungsi objektif belum mencapai ambang yang diterima atau jumlah iterasi belum mencapai batas maksimum

yang ditentukan, pembaruan matriks partisi dilakukan menggunakan Persamaan (7), dan pembaruan matriks tipikalitas dilakukan menggunakan Persamaan (3). Selanjutnya, hitung kembali pusat klaster sebagaimana dijelaskan pada Langkah 3 dan seterusnya.

$$\mu_{ik} = \left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}} / \sum_{i=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}} \quad (7)$$

2.8 *Hierarchical Clustering*

Hierarchical Clustering adalah algoritma pendekatan dimana setiap pendekatan dianggap sebagai sebuah klaster tersendiri, lalu membentuk klaster yang lebih besar dengan menggabungkan pengamatan terdekat dan bergerak menuju klasifikasi (Uddin et al., 2024). Selain itu, *hierarchical clustering* juga membuat visualisasi untuk memudahkan pembaca dalam membaca dan mengetahui klaster dari data. Dalam segmentasi pelanggan, visualisasi data yang dihasilkan dapat membantu untuk merepresentasikan hubungan dari setiap data sehingga dapat menjadikan *hierarchical clustering* menjadi pendekatan yang menjanjikan (Afzal et al., 2024). *Hierarchical Clustering Agglomerative* melakukan pendekatan dengan menganggap setiap data sebagai satu klaster, lalu secara bertahap menggabungkan klaster-klaster yang paling mirip/mendekati menjadi 1 klaster, ulangi hingga semua data tergabung menjadi 1 klaster.

2.9 *Adaptive Learning Particle Swarm Optimization*

Adaptive Learning Particle Swarm Optimization (ALPSO) adalah algoritma optimasi yang dikembangkan dari algoritma optimasi *Particle Swarm Optimization* (PSO) yang dirancang untuk dapat meningkatkan akurasi dari algoritma PSO, dimana dapat mengatasi masalah klasik seperti kecendurungan mengalami konvergensi prematur dan sensitivitas terhadap parameter tetap (Li et al., 2021). ALPSO memberikan kemampuan untuk menjaga keseimbangan antara eksploitasi dan eksplorasi secara dinamis, sehingga mampu

menemukan solusi optimal dengan efisiensi yang lebih tinggi dibandingkan dengan PSO (Wang et al., 2018).

2.10 *Improved Harris Hawks Optimization*

Improved Harris Hawks Optimization (IHHO) adalah algoritma optimasi yang dikembangkan dari *Harris Hawks Optimization* (HHO) dimana telah meningkatkan kualitas awal populasi, serta perubahan linier energi dan intensitas yang sering menyebabkan algoritma terjebak dengan solusi optimal lokal (Wu, 2024). IHHO telah mengatasi kelemahan HHO dengan memperkenalkan teknik inisialisasi populasi yang lebih baik dan memperbaiki aspek eksplorasi dan eksploitasi untuk mencegah terhentinya algoritma dalam solusi optimal lokal(Elgamal et al., 2020).

2.11 *Sine Cosine Algorithm*

Sine Cosine Algorithm (SCA) adalah pendekatan *meta-heuristik* yang menggunakan konsep sederhana tapi dapat mencapai hasil yang mengesankan dalam menyelesaikan masalah optimasi, dengan menggunakan fungsi *trigonometri*(Kuo et al., 2020). SCA dilakukan dengan 2 langkah, langkah pertama adalah menentukan solusi populasi secara acak, langkah kedua dilakukannya pembaruan posisi dengan menggunakan fungsi *sinus cosinus* (Kuo et al., 2021).

2.12 *Genetic Algorithm*

Genetic Algorithm (GA) adalah algoritma optimasi populer dalam evolusi alami dan genetika populasi yang diterapkan untuk menyelesaikan masalah optimasi yang kompleks (Kuo et al., 2020). GA secara bertahap mendekati solusi global yang optimal melalui proses iterasi berulang, sehingga mampu menghindari jebakan solusi lokal dan meningkatkan keakuratan klasifikasi (Zhang & Wu, 2024).

2.13 *Sillhouette Score*

Sillhouette Score adalah metode evaluasi *internal* yang mengukur seberapa baik sebuah titik data pada sebuah klaster dengan klaster lain. *Sillhouette Score* digunakan untuk menilai kualitas klastering dengan menganalisis kedekatan data terhadap klaster lain dengan klaster yang ditempatinya (Nowak-Brzezinska & Horyn, 2020).

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{8}$$

Di mana:

$a(i)$: Rata-rata jarak antara suatu data i dengan semua titik lain dalam klaster yang sama.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_k, j \neq i} distance(i, j) \tag{9}$$

Dimana:

C_i adalah tempat data i berada.

$b(i)$: Rata-rata jarak data i ke semua titik dalam klaster terdekat lainnya (tetangga terdekat).

$$b(i) = \min_{C_k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} distance(i, j) \tag{10}$$

Jika:

$s(i)$ memiliki nilai ≈ 1, menunjukkan jika data i terklaster dengan baik

$s(i)$ memiliki nilai ≈ 0, menunjukkan jika data i berada diantara 2 klaster

$s(i)$ memiliki nilai ≈ -1, menunjukkan jika data i berada di klaster yang salah

2.14 *Davies Bouldin Index*

Davies Bouldin Index (DBI) adalah metode evaluasi yang digunakan mengevaluasi kualitas klastering dengan cara memaksimalkan jarak antar-klaster dan meminimalkan jarak dalam klaster(Singh et al., 2020). *Davies Bouldin Index* memiliki penghitungan dengan persamaan 11:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{j \neq i}\left(\frac{s_i + s_i}{d(c_i, c_j)}\right) \quad (11)$$

Dimana:

n: Jumlah klaster.

s_i: Jarak rata-rata (atau kesamaan) antara setiap data di dalam klaster i dan centroid-nya.

$d(c_i, c_j)$: Jarak antara centroid klaster i dan klaster j

max: Mengambil nilai maksimum dari rasio untuk setiap pasangan klaster i dan j ($j \neq i$).

Dengan semakin rendah sebuah nilai DBI, menunjukkan semakin baik kualitas sebuah klastering yang dilakukan, mengindikasikan jika klaster terpisah dengan baik(Hosen et al., 2023).

2.15 *Calinski Harabasz Index*

Calinski Harabasz Index (CHI) adalah metode evaluasi yang digunakan mengevaluasi kualitas hasil klastering. CHI didefinisikan sebagai 2 karakteristik klastering, yaitu kompak dalam klaster dan terpisah antar klaster (Zhang et al., 2022). CHI mengukur seberapa dekat data dalam klaster dan seberapa jauh pusat klaster terhadap pusat klaster lainnya, CHI dirumuskan sebagai persamaan 12 :

$$CH(K) = \frac{Tr(K)}{Tc(K)} = \frac{\frac{n-K}{K-1}\sum_{i=1}^{K}|C_i|d(v_i, V)^2}{\frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{|C_i|}d(x, v_i)^2} \quad (12)$$

Di mana:

K: jumlah klaster.

n: total sampel.

$|C_i|$: jumlah sampel dalam klaster C_i

v_i: pusat klaster C_i.

V: pusat global data.

d: jarak *Euclidean*.

Nilai CH Index yang semakin tinggi menunjukkan semakin optimal hasil sebuah klastering (Zhang et al., 2022).

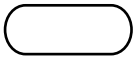
2.16 *Flowchart*

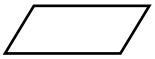
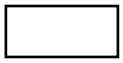
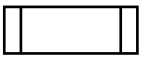
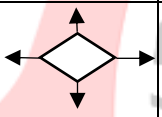
Flowchart adalah representasi diagramatis dari langkah-langkah untuk menyelesaikan suatu masalah yang sering digunakan dalam kursus Pemrograman Berorientasi Objek (OOP)(Armaya'u et al., 2022). Didalam *Flowchart*, terdapat kotak dengan bentuk yang berbeda-beda yang digunakan untuk menampilkan berbagai jenis operasi, yang dihubungkan dengan garis dengan panah yang menunjukkan arah yang harus dijalani untuk mengetahui langkah selanjutnya (Chaudhuri, 2020).

Flowchart dapat diklasifikasikan menjadi 2 jenis, *Program Flowchart* dan *System Flowchart*. *Program flowcharts* bertindak seperti cermin dari program komputer dalam hal simbol-simbol *flowcharting*. Mereka berisi langkah-langkah untuk menyelesaikan satu unit masalah untuk hasil tertentu, sedangkan *System flowcharts* berisi solusi dari banyak unit masalah yang saling terkait erat dan berinteraksi satu sama lain untuk mencapai tujuan (Chaudhuri, 2020).

Berikut Simbol standar yang biasanya digunakan dalam program *Flowchart*:

Tabel 2. 1 *Flowchart*

	*Terminal*	Digunakan untuk menunjukkan awal dan akhir dari serangkaian proses terkait komputer.

	*Input/Output*	Digunakan untuk menunjukkan operasi *input/output* apa pun.
	*Computer Processing*	Digunakan untuk menunjukkan setiap pemrosesan yang dilakukan oleh sistem komputer.
	*Predefined Processing*	Digunakan untuk menunjukkan proses apa pun yang tidak secara khusus didefinisikan dalam *flowchart*.
	*Comment*	Digunakan untuk menulis pernyataan penjelasan apa pun yang diperlukan untuk mengklarifikasi sesuatu.
	Flow Line	Digunakan untuk menghubungkan simbol-simbol.
	*Document Input/Output*	Digunakan ketika input berasal dari dokumen dan output masuk ke dokumen.
	*Decision*	Digunakan untuk menunjukkan titik apa pun dalam proses di mana keputusan harus dibuat untuk menentukan tindakan selanjutnya.
	*On-Page Connector*	Digunakan untuk menghubungkan bagian-bagian *flowchart* yang dilanjutkan di halaman yang sama.
	*Off-Page Connector*	Digunakan untuk menghubungkan bagian-bagian *flowchart* yang dilanjutkan ke halaman yang terpisah.

Sumber: Diadaptasi dari (Chaudhuri, 2020).

2.17 *Python Language Programming*

Python adalah bahasa pemrograman yang digunakan untuk berbagai tugas seperti analisis data, *game*, situs web interaktif, dan otomatisasi. *Python* lebih mudah ditulis dan dibaca dibanding banyak bahasa lain, termasuk *assembly*. Komputer tidak menjalankan kode *Python* secara langsung, melainkan mengonversinya ke kode mesin (*binary*) sebelum dijalankan. Programmer saat ini menggunakan IDE (seperti *Visual Studio Code*) untuk

menulis, menjalankan, dan memperbaiki kode (Porter & Zingaro, 2024). Python terus dikembangkan agar dapat lebih serbaguna dan kuat. Python dapat menjalankan potongan kode saja di *terminal*, dimana *user* dapat menjalankan potongan kode tanpa harus menyimpan dan menjalankan seluruh program, dengan menggunakan tiga tanda panah (>>>) yang disebut *Python Prompt*(Matthes, 2023).

2.18 *Blackbox Testing*

Blackbox Testing adalah sebuah metode pengujian perangkat lunak yang digunakan untuk memeriksa fungsi atau fitur sebuah sistem berdasarkan input dan output tanpa melihat sumber kode sebuah perangkat lunak. *Blackbox Testing* bertujuan mengevaluasi sebuah sistem untuk memenuhi kebutuhan pengguna dan berfungsi sesuai dengan spesifikasi (Rahadi & Vikasari, 2020). Pengujian dengan menggunakan *Blackbox Testing* memberikan keuntungan dimana penguji dapat melakukan pengujian tanpa memahami atau memiliki pengetahuan tentang bahasa pemrograman tertentu, dimana pengujian dilakukan dalam sudut pandang pengguna sehingga dapat menangkap ambiguitas atau inkonsistensi spesifikasi. Salah satu teknik yang umum digunakan adalah *Equivalence Partitioning* yang dimana melakukan pembagian *domain input* kedalam kelas-kelas ekuivalen yang diasumsikan akan menghasilkan keluaran yang sama, sehingga pengujian dapat dilakukan secara efisien (Yulistyanti et al., 2022).

2.19 Tinjauan Studi

Penelitian yang berjudul "*Customer Segmentation of E-commerce data using K-means Klastering Algorithm*" melakukan segmentasi pelanggan dengan menggunakan algoritma *K-Means* pada dataset perusahaan E-Commerce untuk menemukan pola dan perilaku pelanggan untuk dilakukannya analisis kelompok pelanggan yang harus dijadikan fokus utama perusahaan. Penelitian yang berjudul "*Customer Segmentation Using Hierarchical Clustering*" melakukan segmentasi pelanggan dengan menggunakan algoritma klastering

Hierarchical Clustering untuk menemukan kelompok pelanggan pada dataset Mall. *Hierarchical Clustering* digunakan untuk memvisualisasikan pengelompokkan pelanggan serta memperlihatkan hubungan pelanggan dalam kelompok dan antar-kelompok agar dapat dijadikan landasan penting dalam kampanye pemasaran yang lebih terarah, menawarkan rekomendasi produk yang dipersonalisasi, serta meningkatkan alokasi sumber daya *mall*. Penelitian yang berjudul "*Customer Segmentation Using K-Means Klastering and The Adaptive Particle Swarm Optimization Algorithm*" melakukan segmentasi pelanggan dengan menggunakan algoritma klastering *K-Means* yang digabungkan dengan algoritma optimasi *Adaptive Learning Particle Swarm Optimization* (ALPSO), dimana ALPSO digunakan untuk mengoptimalkan titik pusat klaster yang selanjut akan digunakan untuk melakukan klastering *K-means*. ALPSO merupakan pengembangan dari algoritma optimasi *Particle Swarm Optimization* (PSO) dengan mendesain ulang bobot inersia, faktor pembelajaran dan metode pembaruan posisi pada PSO. Penelitian ini membuktikan performa KM-ALPSO lebih baik dibandingkan KM-PSO dan *K-Means* dengan menggunakan 5 dataset dari *UCI Machine Learning*. Penelitian yang berjudul "*Genetic Based Density Peak Possibillistic Fuzzy C-Means Algorithms to Cluster Analysis- A Case Study on Customer Segmentation*" melakukan penelitian segmentasi pelanggan dengan menggunakan Algoritma gabungan DP-GA-PFCM dimana algoritma DP digunakan untuk pencari titik pusat klaster yang kemudian di optimasi dengan menggunakan GA dan melakukan klasterisasi menggunakan PFCM, serta membandingkan dengan algoritma klastering dasar untuk membuktikan jika DP-GA-PFCM lebih baik dibandingkan algoritma klastering dasar. Penelitian yang berjudul "*Factor Influencing The Perceived Usability of Wearable Chair Excoskeleton with Market Segmentation: A Structural Equation Modeling and K-Means Klastering Approach*" melakukan segmentasi pasar pada kursi exoskeleton dengan menggunakan *K-Means* untuk dijadikan dasar teori untuk penelitian lebih lanjut dan untuk pengembangan serta pemasaran

teknologi secara global. Penelitian yang berjudul "*Customer Segmentation-Based Basket Analysis with Assosiation Rule Mining Combined with SVD and K-Means Algorithms*" melakukan segmentasi pelanggan yang dilanjutkan dengan asosiasi agar Perusahaan dapat memberikan rekomendasi yang menggunakan algoritma SVD dan *K-Means* kemudian dilanjutkan dengan FP-Growth untuk memberikan rekomendasi pada pelanggan. Penelitian yang berjudul "*Metaheuristic-based Possibilistic Multivariate Fuzzy Weighted C-Means Algorithms for Market Segmentation*" melakukan perbandingan algoritma optimasi yang digunakan, seperti Sine Cosine Algorithm (SCA), Genetic Algorithm (GA), dan Particle Swarm Optimization (PSO) yang digabungkan algoritma klastering *Possibilistic Multivariate Fuzzy Weighted C-Means Algorithms* (PMFWCM) dengan menggunakan dataset *UCI Machine Learning*, yang menunjukkan jika algoritma optimasi SCA lebih baik jika dibandingkan dengan PSO dan GA.

Penelitian yang berjudul "*Segmenting The Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering*" melakukan segmentasi pasar pada bidang pendidikan tinggi swasta dengan mempertimbangkan rekam akademis dan faktor geografis menggunakan *K-means*. Penelitian yang berjudul "*Data-Driven Strategies for Digital Native Market Segmentation using Clustering*" melakukan segmentasi pasar dengan menggunakan algoritma klastering konvensional seperti *K-Means, Mini-batch K-Means, Fuzzy C-Means*, dan *Hierarchical Clustering*, serta membandingkan akurasi dan ketepatan keempat algoritma konvensional. Penelitian yang berjudul "*E-Commerce Recommender System Based on Improved K-Means*" membuat sistem rekomendasi *e-commerce* yang memanfaatkan algoritma klastering seperti *K-Means* yang digabungkan dengan GA lalu dibandingkan dengan algoritma konvensional *K-Means* serta FCM dan mendapatkan hasil jika algoritma gabungan GA *K-Means* lebih baik dibandingkan algoritma konvensional. Penelitian yang berjudul "*Metaheuristic-Based Possibilistic Fuzzy K-Mode Algorithms for*

Categorical Data Clustering" mengusulkan Algoritma *Possibilistic Fuzzy K-Modes* (PFKM) dan algoritma *metaheuristik* seperti SCA, GA, dan PSO, sehingga didapatkan model SCA-PFKM, GA-PFKM, PSO-PFKM yang dibandingkan menggunakan indeks *Sum-of-Squared Error* (SSE) dan akurasi. Penelitian ini memberikan hasil jika PSO-PFKM dan SCA-PFKM memberikan kinerja yang lebih baik pada sebagian dataset. Penelitian yang berjudul "*Clustering Evaluation by Davies-Bouldin Index (DBI) in Cereal Data using K-Means*" bertujuan untuk mengetahui jumlah klaster optimal 3 atau 5 pada *dataset cereal* menggunakan *Davied-Bouldin Index* (DBI), sehingga didapatkan hasil jika jumlah klaster 5 lebih baik dibandingkan dengan jumlah klaster 3. Penelitian yang berjudul "*Clustering Models for Hospitals in Jakarta using Fuzzy C-Means and K-Means*" bertujuan untuk mengelompokkan rumah sakit berdasarkan distribusi layanan Kesehatan dan tenaga Kesehatan menggunakan Teknik klastering *K-Means* dan *Fuzzy C-Means* (FCM). Penelitian ini menghasilkan jika baik FCM dan KM menghasilkan jumlah klaster yang sama, namun distribusinya berbeda, yang diamati dengan menggunakan 3 metrik jarak, seperti *Hammiung* , *Euclidian*, dan *Manhattan*. Penelitian yang berjudul "*K-Means Clustering Approach for Intelligent Customer Segmentation Usiing Customer Purchase Behavior Data*" melakukan segmentasi pelanggan pada *e-commerce* guna meningkatkan strategi bisnis menggunakan Teknik klastering *K-Means* untuk menganalisis pola pembelian pelanggan untuk mengelompokkan pelanggan berdasarkan kesamaan perilaku.

Studi penelitian sebelumnya membahas algoritma-algoritma klastering seperti *K-Means, Possibilistic Fuzzy C-Means,* dan *Hirearchical Clustering* beserta algoritma Optimasi yang digunakan untuk membantu meningkatkan akurasi serta mengurangi nilai *error* dari algoritma klastering. Studi sebelumnya melakukan klastering pada segmentasi pelanggan, segmentasi pasar, dan pengelompokkan pengguna baik menggunakan algoritma klastering dasar atau menggunakan algoritma klastering yang dikombinasikan dengan

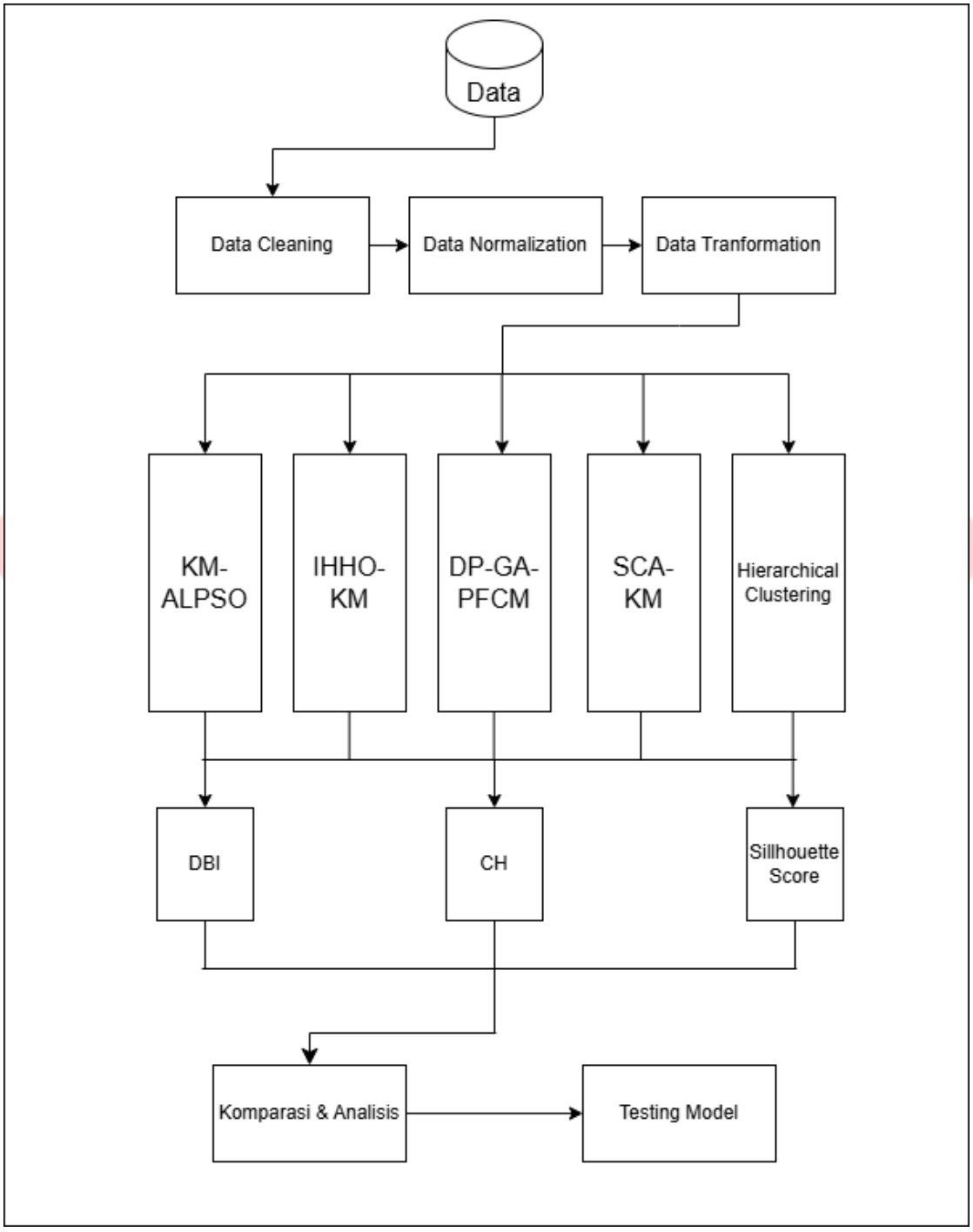
algoritma optimasi serta membandingkan hasil klastering dari beberapa algoritma, seperti algoritma klastering dasar dengan algoritma klastering yang sudah dikombinasikan algoritma optimasi. Sehingga menghasilkan algoritma-algoritma yang dianggap terbaik untuk dilakukannya klastering.

Penelitian ini bertujuan untuk membandingkan beberapa algoritma yang sudah digunakan pada studi sebelumnya pada dataset ”Ecommerce Dataset.csv” untuk menentukan manakah algoritma yang terbaik untuk melakukan klastering pada segmentasi pelanggan. Penelitian ini menggunakan algoritma klastering *K-Means* yang digabungkan dengan algoritma optimasi seperti ALPSO, dan IHHO yang telah digunakan pada penelitian sebelumnya serta SCA yang pada penelitian sebelumnya terbukti lebih baik jika dibandingkan dengan PSO dan GA dengan menggunakan algoritma klastering PMFWCM. Serta menggunakan algoritma DP-GA-PFCM dan *Hierarchical clustering* untuk mendapatkan algoritma yang efektif dan efisien dalam segmentasi pelanggan.

BAB III

METODE PENELITIAN

3.1 Kerangka Pemikiran



Gambar 3. 1 Kerangka Pemikiran

Pada Gambar 3.1 menunjukkan alur sistematis pengolahan data hingga *testing model* untuk menemukan algoritma klastering terbaik untuk segmentasi pelangan pada Dataset *'ECommerce Customer.csv'*. Pada tahap *Data Cleaning Dataset* yang didapat dilakukan penghapusan data-data yang tidak digunakan dan data yang tidak berpengaruh pada penghitungan algoritma yang akan digunakan. Pada tahap *Data Normalization,* dilakukannya normalisasi agar rentang data tidak terlalu besar dari data satu ke data lainnya. Kemudian dilakukannya *Data Tranformation* dengan mengubah nilai data yang dapat membantu pengolahan data menjadi lebih mudah dan lebih cepat. Data yang sudah selesai dilakukan *Data Cleaning, Data Tranformation*, dan *Data Normalization*, *dataset* dilakukan pengolahan data dengan mengaplikasikan pada 5 algoritma klastering beserta algoritma optimasinya seperti KM-ALPSO, IHHO-KM, SCA-KM, DP-GA-PFCM, dan *Hierarchical clustering*. Setelah Pemodelan berhasil dilakukan, maka setiap *model training* dilakukannya evaluasi menggunakan metode evaluasi seperti *Sillhouette Score*, *Davies-Boulding Index* (DBI), dan *Calinski–Harabasz Index* (CHI). Kemudian hasil penghitungan evaluasi dibandingkan serta dianalisis untuk mendapatkan algoritma klastering yang lebih baik untuk segmentasi pelanggan. Setelah komparasi dan analisis dilakukannya *testing model* dengan menggunakan data baru untuk melihat hasil *model training.*

3.2 *Data Understanding*

Penelitian menggunakan Dataset yang berjudul "*Ecommerce Customer*" berisi 500 data dan memiliki 8 atribut yaitu:

a. *Email* : Alamat *Email* yang digunakan untuk mengidentifikasi pelanggan dalam dataset.

b. *Address* : Alamat tempat tinggal pelanggan yang memuat nomor jalan,nama jalan, kode negara bagian, dan kode pos

c. *Avatar* : warna atau representasi *avatar* yang dipilih oleh pelanggan yang digunakan untuk personalisasi antar muka pengguna

d. *Avg. Session Length* : Rata-rata durasi sesi pelanggan dalam satuan menit yang berinteraksi dengan *platform* perusahaan

e. *Time on App* : Rata-rata waktu yang digunakan oleh pelanggan untuk berinteraksi dengan *platform* perusahaan berupa aplikasi dalam satuan menit

f. *Time on Website* : Rata-rata waktu yang digunakan oleh pelanggan untuk berinteraksi dengan *platform* perusahaan berupa website dalam satuan menit

g. *Length of Membership* : Durasi sejak pelanggan menjadi anggota perusahaan dalam satuan tahun

h. *Yearly Amount Spent* : Jumlah total uang yang dikeluarkan pelanggan untuk berbelanja selama satu tahun terakhir dalam satuan *dollar*

Tabel 3.1 Berisi informasi dataset yang digunakan pada penelitian, terdapat 8 kolom masing-masing kolom terdiri dari 500 data. *Email, Address,* dan *Avatar* memiliki tipe data *String* sedangkan *Avg. Session Length, Time on App, Time on Website, Length of Membership,* dan *Yearly Amount Spent* memiliki tipe data berupa *integer*.

Tabel 3. 1 Informasi *Dataset*

No.	Nama Kolom	Jumlah Data	Tipe Data	Contoh Data
1	*Email*	500	*String*	mstephenson@fernandez.com
2	*Address*	500	*String*	*Unit 6538 Box 8980DPO AP 09026-4941*
3	*Avatar*	500	*String*	*Aqua*
4	*Avg. Session Length*	500	*Integer*	31,936548618448914
5	*Time on App*	500	*Integer*	11,814128294972196
6	*Time on Website*	500	*Integer*	37,14516822352819
7	*Length of Membership*	500	*Integer*	3,202806071553459
8	*Yearly Amount Spent*	500	*Integer*	427,19938489532814

3.3 *Data Preprocessing*

3.3.1 *Data Cleaning*

Berdasarkan penelitian sebelumnya yang berjudul "*Customer Segmentation of E-commerce data using K-means Clustering Algorithm*" yang menggunakan *dataset* yang sama, tidak ditemukannya data yang hilang atau data yang kosong, tetapi terdapat atribut yang tidak digunakan dalam penelitian seperti *Email*, *Address*, dan *Avatar* dalam pembuatan model klastering sehingga atribut tersebut akan dihapus (Rajput & Singh, 2023).

Tabel 3. 2 Informasi *Dataset* sebelum *Data Cleaning*

No.	Nama Kolom	Tipe Data

1	*Email*	*String*
2	*Address*	*String*
3	*Avatar*	*String*
4	*Avg. Session Length*	*Integer*
5	*Time on App*	*Integer*
6	*Time on Website*	*Integer*
7	*Length of Membership*	*Integer*
8	*Yearly Amount Spent*	*Integer*

Tabel 3.2 menunjukkan informasi *dataset* sebelum dilakukannya pembersihan data, dimana terdapat kolom *Email, Address, Avatar, Avg. Session Length, Time on App, Time on Website, Length of Membership,* dan *Yearly Amount Spent.* Sedangkan pada Tabel 3.3 menunjukkan bahwa kolom *Email, Address,* dan *Avatar* telah dihapus, sehingga hanya tersisa 5 kolom, kolom *Avg. Session Length, Time on App, Time on Website, Length of Membership,* dan *Yearly Amount Spent*.

Tabel 3. 3 Informasi *Dataset* setelah *Data Cleaning*

No.	Nama Kolom	Tipe Data
1	*Avg. Session Length*	*Integer*
2	*Time on App*	*Integer*
3	*Time on Website*	*Integer*

4	*Length of Membership*	*Integer*
5	*Yearly Amount Spent*	*Integer*

3.3.2 *Data Normalization*

Berdasarkan pemeriksaan *dataset* ini, rentang data pada dataset ini cukup beragam dan berbeda cukup jauh, sehingga dilakukan normalisasi data agar rentang data memiliki ukuran yang mirip dan seragam:

Tabel 3. 4 Data sebelum dilakukan Data Normalization

Avg. Session Length	*Time on App*	*Time on Website*	*Length of Membership*	*Yearly Amount Spent*
34,4972677251122	12,6556511491667	39,5776680195261	4,08262063295296	587,9510539684
31,9262720263601	11,1094607286825	37,2689588682977	2,66403418213262	392,204933444326
33,0009147556426	11,3302780577775	37,1105974421208	4,10454320237642	487,547504867472
34,3055566297555	13,7175136651425	36,7212826779031	3,12017878274809	581,852344035217
33,3306725236463	12,7951885510781	37,5366533005947	4,44630831835143	599,406092045763

Tabel 3. 5 Data setelah dilakukan Normalisasi Data

Avg. Session Length	*Time on App*	*Time on Website*	*Length of Membership*	*Yearly Amount Spent*
1,456351173	0,607280027	2,493588585	0,550106512	1,118653852
-1,136502147	-0,949463723	0,20655573	-0,87092735	-1,351783019

-0,052723217	-0,727139232	0,049681148	0,572066903	-0,148500915
1,263010223	1,676390152	-0,33597837	-0,413995788	1,041684368
0,279838026	0,747769824	0,4717368	0,914421648	1,263223504

Tabel 3.4 menunjukkan informasi *dataset* sebelum dilakukannya normalisasi, dimana setiap kolom memiliki rentang data yang berbeda cukup jauh. Sehingga dilakukannya normalisasi untuk menyama-ratakan rentang data yang dimiliki setiap atribut. Normalisasi yang dilakukan adalah normalisasi *Z-Score*, yang dapat membantu memberikan bobot yang sama pada setiap data yang dapat memberikan kestabilan akurasi pada data (Henderi et al., 2021). Tabel 3.5 menunjukkan informasi dataset setelah dilakukannya *Z-Score Normalization* dimana setiap data memiliki rentang yang sama antara -4 hinga 4.

3.3.3 *Data Transformation*

Pada dataset terdapat kolom *Avg. Session Length, Time on App, Time on Website, Length of Membership,* dan *Yearly Amount Spent* yang memiliki nilai desimal yang cukup panjang, sehingga dilakukannya pemotongan nilai desimal dibelakang koma hingga 5 angka dibelakang koma dengan pembulatan keatas untuk mengurangi nilai yang terlalu panjang. Tabel 3.6 menunjukkan data sebelum dilakukannya *Data Transformation Rounding* atau pembulatan keatas, sedangkan Tabel 3.7 menunjukkan data setelah transformasi data dilakukan, dimana hanya menyisakan 5 angka di belakang koma.

Tabel 3. 6 Data sebelum dilakukan *Data Transformation*

Avg. Session Length	*Time on App*	*Time on Website*	*Length of Membership*	*Yearly Amount Spent*

1,456351173	0,607280027	2,493588585	0,550106512	1,118653852
-1,136502147	-0,949463723	0,20655573	-0,87092735	-1,351783019
-0,052723217	-0,727139232	0,049681148	0,572066903	-0,148500915
1,263010223	1,676390152	-0,33597837	-0,413995788	1,041684368
0,279838026	0,747769824	0,4717368	0,914421648	1,263223504

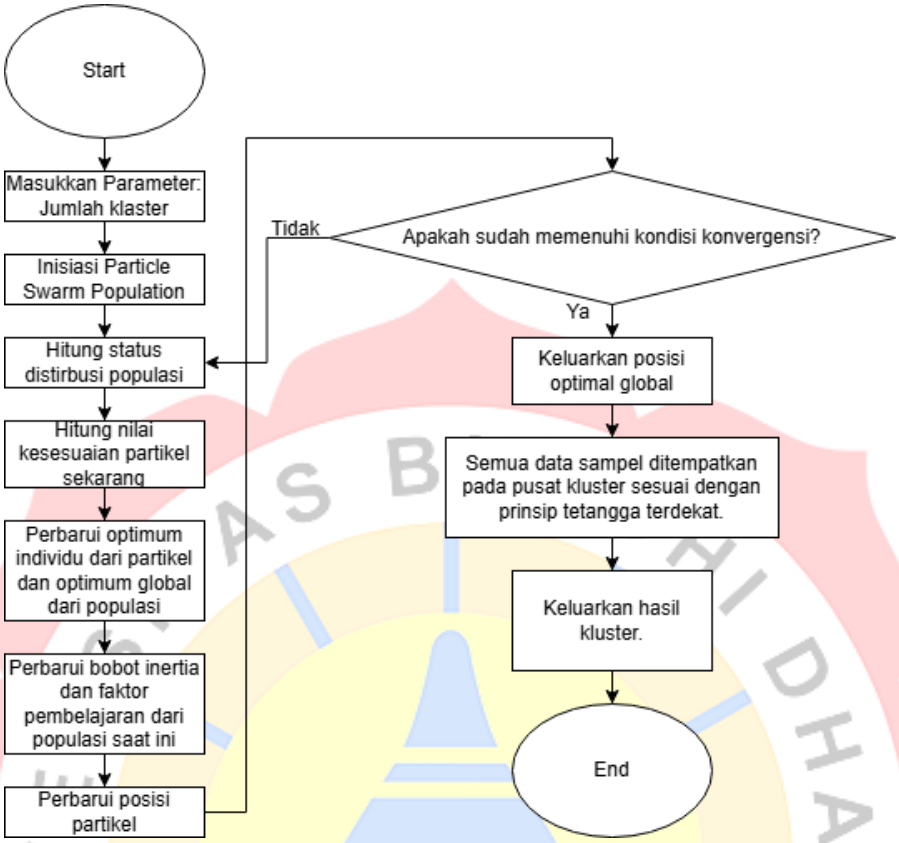
Tabel 3. 7 Data setelah dilakukannya *Data Transformation*

Avg. Session Length	*Time on App*	*Time on Website*	*Length of Membership*	*Yearly Amount Spent*
1,45635	0,60728	2,49359	0,55011	1,11865
-1,13650	-0,94946	0,20656	-0,87093	-1,35178
-0,05272	-0,72714	0,04968	0,57207	-0,14850
1,26301	1,67639	-0,33598	-0,41400	1,04168
0,27984	0,74777	0,47174	0,91442	1,26322

3.4 *Model Training*

Penelitian ini menggunakan beberapa algoritma klastering yang digabungkan dengan algoritma optimasi untuk membandingkan setiap algoritma klastering untuk mencari algoritma yang terbaik.

3.4.1 KM-ALPSO

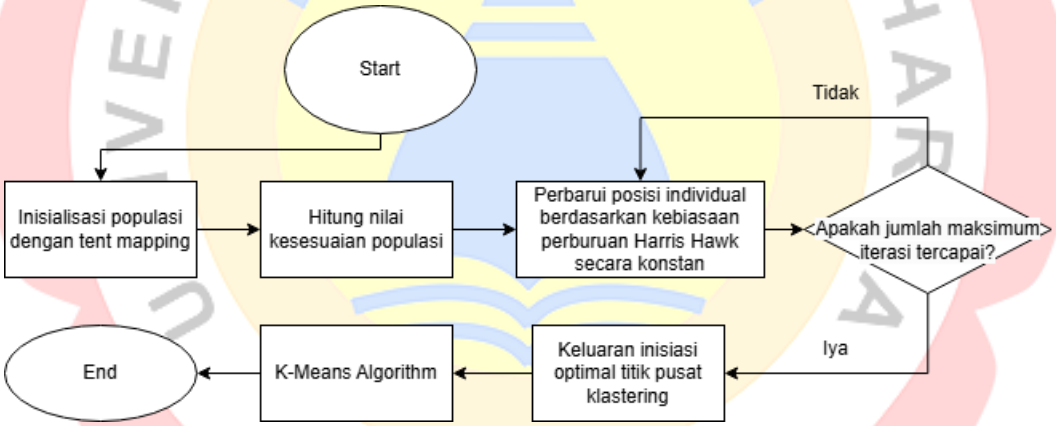


Gambar 3. 2 *Flowchart* KM-ALPSO
Sumber: (Li et al., 2021)

KM-ALPSO adalah algoritma *K-Means* yang dioptimasi titik pusat klaster menggunakan ALPSO, sehingga ALPSO dilakukan sebelum *K-Means* untuk menentukkan titik pusat klaster terbaik(Li et al., 2021). Gambar 3.2 menunjukkan alur dari algoritma KM-ALPSO (Li et al., 2021), langkah pertama yang dilakukan adalah memasukan parameter seperti jumlah klaster, parameter PSO seperti bobot inersia, jumlah iterasi maksimum dan sebagainya. Langkah kedua adalah inisiasi *particle swarm population* seperti menentukan titik pusat awal klaster. Langkah ketiga adalah menghitung distribusi awal partikel dengan menghitung jarak *euclidian* setiap data dengan pusat klaster. Lalu menghitung nilai kesesuaian berdasarkan variansi dalam klaster, seperti jumlah jarak *kuardrat* antara data dan *centroid cluster*. Langkah kelima adalah memperbarui nilai Optimum lokal dan global untuk setiap partikel dan antara

semua partikel. Langkah keenam adalah memperbarui bobot *inersia* dan parameter pembelajaran PSO untuk disesuaikan untuk mempercepat konvergensi partikel ke solusi optimal. Langkah ketujuh adalah memperbarui posisi partikel dengan menghitung bobot *inersia* dengan konstanta pembelajaran. Langkah kedelapan adalah mengecek kondisi konvergensi, dimana memeriksa apakah nilai kesesuaian tidak mengalami perubahan secara signifikan atau jumlah iterasi maksimum sudah tercapai. Jika belum tercapai, akan kembali ke langkah ketiga, jika sudah tercapai maka akan mendapatkan titik pusat klaster terbaik. Setelah titik pusat klaster didapatkan, maka dilakukan penentuan klaster menggunakan algoritma *K-Means*. Setelah hasil klaster *K-Means* didapatkan, proses KM-ALPSO telah selesai dilaksanakan.

3.4.2 IHHO-KM

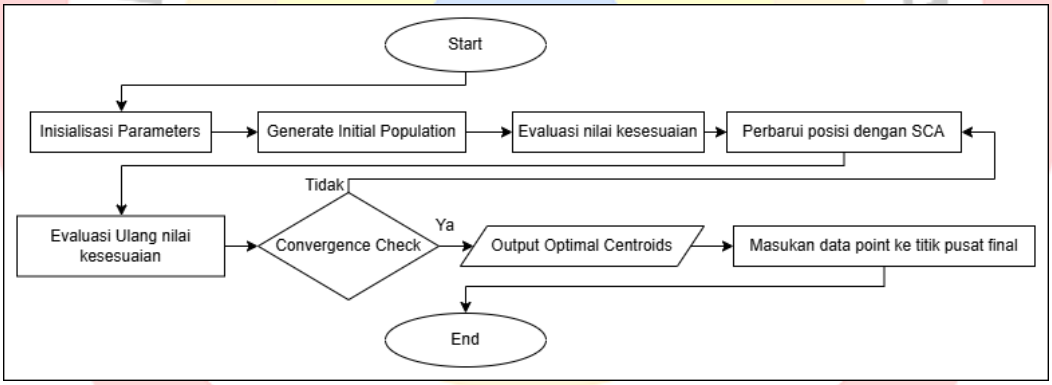


Gambar 3. 3 *Flowchart* IHHO-KM
Sumber: (Wu, 2024)

IHHO-KM adalah algoritma *K-Means* yang dioptimasi untuk mendapatkan akurasi dan efisiensi konvergensi terbaik, sehingga IHHO-KM memiliki kinerja unggul dalam memilih pusat klaster awal (Wu, 2024). Gambar 3.3 menunjukkan alur Algoritma IHHO-KM dimana pada tahap awal melakukan inisiasi *centroid* awal menggunakan *tent mapping* dimana dapat memastikan distribusi *centroid* awal lebih baik dan lebih menyebar secara merata, serta inisiasi jumlah *iterasi* maksimum. Langkah kedua adalah mengevaluasi setiap individu dalam populasi menggunakan

fungsi objektif. Langkah ketiga adalah memperbarui posisi individual menggunakan kebiasaan *Harris-Hawk*. Langkah keempat adalah pengecekan kondisi apakah sudah mencapai *iterasi* maksimum atau posisi individual sudah tidak berubah secara signifikan. Jika sudah mencapai iterasi maksimum atau sudah tidak ada perubahan secara signifikan, maka titik pusat optimal klastering dapat digunakan sebagai titik pusat awal untuk *K-Means*, jika belum mencapai titik pusat maksimum iterasi atau masih ada perubahan secara signifikan maka proses dilanjutkan pada perbarui posisi individual berdasarkan kebiasaan HHO. Setelah didapatkan titik pusat awal yang sudah di optimasi dengan IHHO, maka titik pusat dapat dilakukan klastering dengan *K-Means*.

3.4.3 SCA-KM

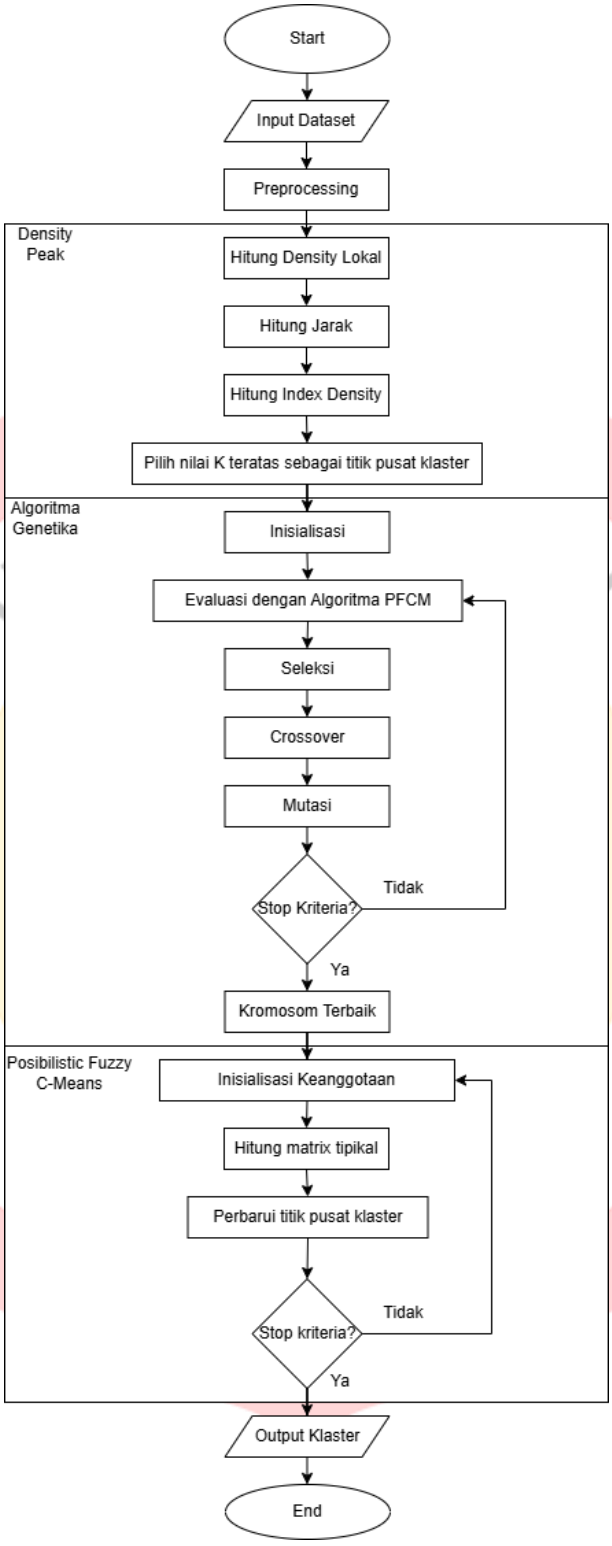


Gambar 3. 4 *Flowchart* SCA-KM
Sumber: Diadaptasi dari (Kuo et al., 2020)

SCA-KM adalah pendekatan klastering yang dioptimasi titik pusat klaster menggunakan *Sine Cosine Algorithm*, dimana memperbarui titik pusat dengan menggunakan persamaan *trigonometri* (Kuo et al., 2020). Gambar 3.4 menunjukkan alur algoritma SCA-KM dimana pada awalnya dilakukannya inisialisasi parameter seperti jumlah individu dalam klaster, *iterasi* maksimum, dan parameter algoritma SCA. Langkah kedua adalah membentuk populasi awal dimana setiap individu merepresentasikan sebuah solusi berupa koordinat *centroid* awal untuk algoritma

klastering. Langkah ketiga adalah mengevaluasi kesesuaian untuk setiap individu menggunakan WCSS. Dimana semakin kecil nilai WCSS maka semakin baik kualitasnya. Langkah keempat adalah dengan memperbarui posisi *centroid* dengan SCA yang menggunakan fungsi sinus cosinus. Langkah kelima adalah mengevaluasi nilai kesesuaian ulang menggunakan WCSS. Langkah keenam adalah mengecek apakah *iterasi* sudah mencapai maksimum, atau apakah perubahan yang terjadi sangat kecil, jika belum tercapai maka kembali pada langkah memperbarui posisi dengan SCA. Jika sudah tercapai maka nilai posisi titik pusat *centroid* dapat digunakan untuk melakukan klastering *K-Means*.

3.4.4 DP-GA-PFCM

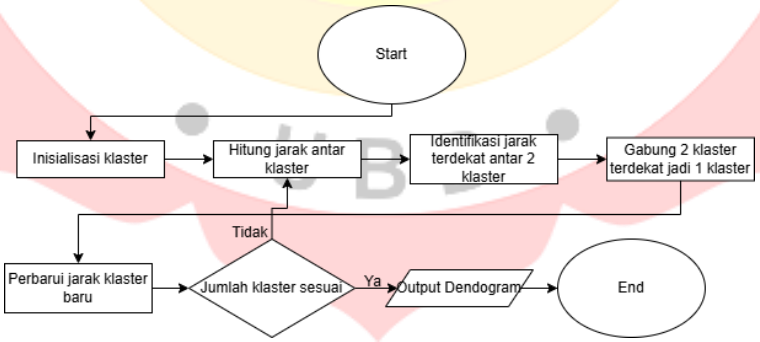


Gambar 3. 5 *Flowchart* DP-GA-PFCM
Sumber:(Kuo et al., 2023)

DP-GA-PFCM adalah pendekatan klastering yang menggabungkan 3 metode *Density Peak*, *Genetic Algorithm*, dan *Possibillistic Fuzzy C-Means* untuk

meningkatkan kinerja klastering. Gambar 3.5 menunjukkan alur algoritma *Density Peak, Genetic Algorithm,* dan *Prossibilistic Fuzzy C-Means*. Pada tahap *Density Peak*, menghitung berapa banyak tetangga dalam radius R untuk setiap data, lalu menghitung jarak antar titik data menggunakan jarak *Euclidean*. Lalu menghitung tingkat kepadatan relatif setiap titik data, terakhir memilih nilai k teratas sebagai pusat klaster. Lalu pada bagian *Genetic Algorithm* dilakukan inisialisasi populasi awal, lalu mengevaluasi setiap populasi dengan algoritma PFCM untuk menghitung kesesuaiannya. Melakukan Seleksi, *Crossover*, dan Mutasi untuk menghasilkan generasi baru. Lalu pengecekan kriteria, apakah iterasi tercapai atau tidak, jika tidak maka dilakukan pengulangan pada evaluasi menggunakan PFCM, jika tercapai maka dipilih kromosom terbaik. Pada PFCM dilakukan inisialisasi matriks keanggotaan, lalu dilakukan penghitungan matriks tipikal berdasarkan pusat klaster. Selanjutnya adalah memperbarui titik pusat klaster berdasarkan matriks keanggotaan dan matrik tipikal, kemudian dilakukan pengecekan apakah ada perbedaan hasil antara iterasi. Lalu akan menghasilkan klaster yang terbentuk dari Proses PFCM.

3.4.5 *Hierarchical Clustering*



Gambar 3. 6 *Flowchart Hierarchical clustering*
Sumber: Diadaptasi dari (Uddin et al., 2024)

Hierarchical Clustering adalah algoritma pendekatan dasar klastering yang dapat memisahkan klaster serta memberikan visualisasi berupa *dendogram* tentang hubungan dari setiap klaster. Gambar 3.6 menunjukkan alur *Hierarchical clustering*,

pertama dilakukan inisialisasi setiap data dianggap sebagai sebuah klaster dan jumlah klaster yang ingin dicapai. Setelah inisialisasi dilakukan penghitungan jarak antar klaster. Setelah penghitungan jarak antar klaster didapatkan, maka identifikasi jarak dilakukan dengan mencari jarak terpendek dari 2 buah klaster, 2 buah klaster yang saling berdekatan dilakukan penggabungan, sehingga 1 klaster terdiri dari 2 buah data terdekat. Setelah penggabungan klaster dilakukan, maka penghitungan jarak antar klaster baru. Setelah itu dilakukan pengecekan kondisi, apakah jumlah klaster sudah sesuai dengan yang diinginkan atau belum, jika belum maka akan dilakukan pengulangan pada penghitungan jarak antar klaster, selanjutnya akan dilakukan output berupa *dendogram*.

3.5 Metode Evaluasi

Untuk mendapatkan hasil algoritma terbaik, dilakukannya evaluasi untuk diidentifikasi kemampuan algoritma dalam melakukan klastering dengan baik. Evaluasi yang digunakan adalah *Sillhouette Score*, *Davies Bouldin Index* (DBI), dan *Calinski Harabasz Index* (CHI), dimana setiap evaluasi ini dapat memberikan perpektif berbeda dalam mengevaluasi kualitas klastering.

3.5.1 *Sillhouette Score*

Sillhouette Score adalah salah satu metode evaluasi yang digunakan untuk menilai kualitas hasil klastering. Metode ini mengukur seberapa baik objek-objek dalam satu klaster memiliki kemiripan internal (*intra-cluster similarity*) dibandingkan dengan kemiripan terhadap klaster lain (*inter-cluster dissimilarity*)(Papadogiannis et al., 2024). *Sillhouette Score* memberikan nilai yang berkisar antar -1 sampai 1, dimana 1 menunjukkan bahwa objek berada didalam klasternya dengan baik dan jauh dari klaster lain, nilai 0 menunjukkan bahwa objek berada dibatas antara dua klaster,

sedangkan nilai -1 menunjukkan jika objek lebih mendekati klaster lain dibandingkan klasternya sendiri, dimana menunjukkan pembagian klaster yang buruk.

Sillhouette Score menghitung jarak *intra-cluster* dimana menghitung rata-rata jarak antara data *i* dengan seluruh data lain dalam klaster yang sama dan menghitung jarak *inter-cluster* dimana menghitung rata-rata jarak antara data *i* dan semua data di klaster lain. Setelah jarak *intra-cluster* dan *inter-cluster* didapatkan, maka dilakukan penghitungan *Sillhouette Score* pada data *i*. Setelah didapatkan untuk semua data, dilakukan penghitungan rata-rata dari semua *Sillhouette Score* pada data *i*. Semakin tinggi *Sillhouette Score*, maka semakin baik pula hasil klasteringnya.

3.5.2 *Davies Bouldin Index*

Davies Bouldin Index (DBI) adalah salah satu metrik evaluasi yang digunakan untuk menilai kualitas klastering. DBI mengukur tingkat kesamaan antar klaster berdasarkan rasio jarak *intra-klaster* terhadap jarak *antar-klaster* (Papadogiannis et al., 2024). DBI menghasilkan nilai dimana nilai yang lebih kecil menunjukkan hasil klaster yang lebih baik dimana klaster memiliki kepadatan tinggi secara internal dan terpisah dengan jelas satu sama lain. DBI dikatakan sebagai rata-rata dari nilai maksimum kesamaan antar klaster untuk semua klaster (Hosen et al., 2023). DBI bertujuan untuk memaksimalkan jarak inter-klaster dan pada waktu yang sama mengurangi jarak antara titik-titik di dalam klaster untuk mengukur kualitas klastering.

Penghitungan DBI dimulai dengan menghitung rata-rata penyebaran dalam setiap klaster, lalu menghitung jarak antar pusat klaster, lalu menghitung rasio *similarity* antar klaster. Kemudian menghitung DBI untuk setiap klaster. Hitung total DBI seluruh klaster dengan mengambil nilai rata-rata dari nilai DBI setiap klaster.

3.5.3 *Calinski–Harabasz Index*

Calinski-Harabasz Index (CHI) adalah metrik evaluasi yang digunakan untuk menilai kualitas hasil klastering dengan mengukur rasio antar *variabilitas antar-klaster* dan *variabilitas intra-klaster*. CHI mengukur sejauh mana data dalam klaster yang sama saling berdekatan dan sejauh mana data dalam klaster yang berbeda saling berjauhan (Zhang et al., 2022). CHI digunakan sebagai metode evaluasi untuk membandingkan hasil klastering yang dihasilkan oleh beberapa algoritma, sehingga membantu menentukan algoritma terbaik berdasarkan tingkat kedekatan dan pemisahan klaster.

Calinski-Harabasz Index pertama menghitung rata-rata semua data dalam *dataset*, lalu menghitung penyebaran antar klaster dengan menghitung jumlah *kuardrat* jarak antara pusat setiap klaster degnan pusat global dikalikan jumlah data dalam klaster. Lalu hitung penyebaran dalam klaster dengan menghitung jumlah *kuardrat* jarak antar data dengan pusat klaster, tahap terakhir menghitung *Calinski-Harabasz Index* total.