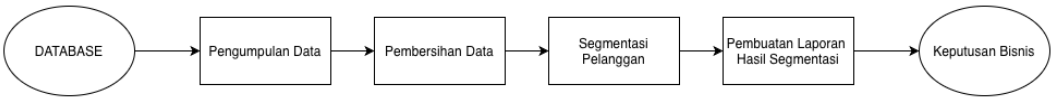
BAB III

METODE PENELITIAN

3.1 Flow Existing

Proses clustering pelanggan existing terdiri dari beberapa tahap untuk menghasilkan cluster-cluster pelanggan di masing-masing daerah hingga dapat menjadi informasi berguna yang dapat digunakan untuk keputusan bisnis seperti campaign marketing targeted sesuai dengan kebiasaan berbelanja pelanggan di daerah tersebut, diskon atau voucher targeted sesuai dengan kategori produk yang paling sering dibeli dan lainnya.



Gambar 3.1 Flow Existing

Proses clustering pelanggan dapat dilihat di gambar 3.1 yang menunjukkan proses dari pengumpulan data dari database yang berisi riwayat pembelian customer, proses pembersihan data untuk memastikan tidak ada missing values, proses mencari cluster-cluster pelanggan berdasarkan karakter kebiasaan berbelanja yang unik di masing-masing daerah, hingga pembuatan laporan hasil cluster yang ditemukan yang dapat digunakan sebagai informasi untuk digunakan oleh tim bisnis.

3.2 Gambaran Umum Metodologi

Penelitian ini merancang metodologi komprehensif untuk clustering pelanggan e-commerce PT. XYZ melalui lima tahap terintegrasi. Dengan mengkombinasikan K-Prototypes dan AdaBoost , penelitian ini mengoptimalkan identifikasi segmen pelanggan berdasarkan usia, gender, dan preferensi kategori produk. Pendekatan ini memproses data transaksi dari tiga bulan (Juni-Agustus 2025) secara paralel untuk memungkinkan analisis berkala dan perbandingan tren perilaku pembelian antar periode. Dari awal hingga akhir, setiap tahap penelitian

dirancang untuk memaksimalkan validitas hasil, interpretability insight, dan reliabilitas rekomendasi marketing.

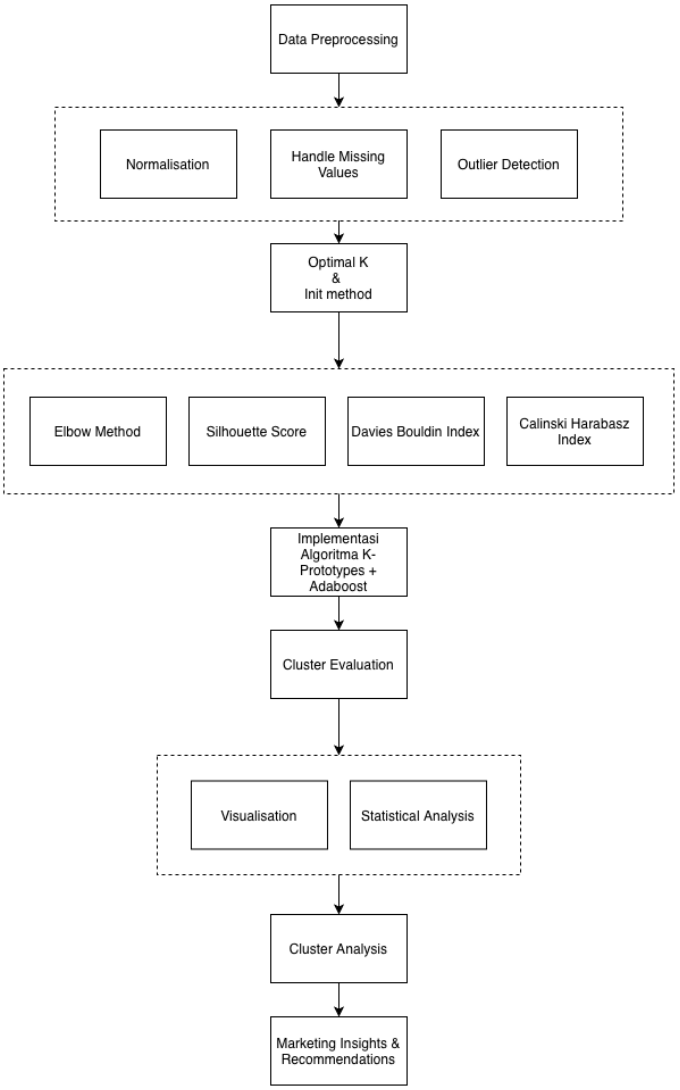
Tujuan proses clustering dalam penelitian ini adalah untuk menemukan kelompok pelanggan yang benar-benar "bermakna" dan bisa dimanfaatkan secara praktis, misalnya sebagai dasar penyusunan campaign marketing, pemberian voucher yang lebih targeted, dan inisiatif bisnis lain yang lebih efektif. Cluster tidak ditentukan atau disiapkan sejak awal, sehingga bentuk dan karakter tiap cluster sepenuhnya bergantung pada pola yang muncul dari data; algoritma akan mengelompokkan pelanggan berdasarkan kemiripan karakteristiknya tanpa label awal. Dengan demikian, drive utama dari clustering adalah mengidentifikasi kelompok pelanggan berdasarkan riwayat berbelanja pada suatu area tertentu, agar informasi mengenai kelompok-kelompok ini dapat dimanfaatkan untuk strategi bisnis yang lebih targeted dan diharapkan meningkatkan efektivitas marketing maupun pemanfaatan sumber daya perusahaan.

3.3 Dataset

Perancangan solusi customer segmentation dilakukan dengan menggunakan dataset yang merupakan data transaksi pada e-commerce PT. XYZ yang terdiri dari Gender, Usia, dan Kategori Produk yang dibeli oleh customer. Data Gender dan Kategori Produk yang dibeli merupakan data kategorikal, dan data Usia merupakan data numerik. Data ini dapat digunakan untuk mengenali pola belanja customer pada e-commerce dengan analisis yang lebih lanjut.

3.4 Perancangan Solusi

Solusi yang dirancang untuk PT. XYZ terdiri dari pipeline machine learning end-to-end yang mencakup preprocessing data, clustering menggunakan ensemble K-Prototypes, validasi menggunakan AdaBoost, analisis komparatif dengan baseline, dan interpretasi hasil untuk rekomendasi bisnis. Setiap tahapan dirancang untuk memenuhi kebutuhan spesifik PT. XYZ dalam hal efisiensi waktu, akurasi clustering, dan actionability hasil.



Gambar 3.2 Alur Penelitian

3.4.1 Data Preprocessing

Penelitian ini memanfaatkan dataset transaksi pelanggan PT. XYZ dari Kota Tangerang selama periode Juni-Agustus 2025, mencerminkan periode 3 bulan terkini yang menjadi basis clustering rutin perusahaan. Dataset keseluruhan sangat substansial dengan total 23.934 transaksi (Juni), 25.258 transaksi (Juli), dan 26.040 transaksi (Agustus), menunjukkan pertumbuhan konsisten dalam volume aktivitas pelanggan selama periode penelitian. Ketiga dataset ini dipilih secara strategis karena merepresentasikan kondisi pasar terkini dan memungkinkan validasi berkala dari clustering yang dilakukan. Setiap record dalam dataset

mengandung tiga fitur utama yang relevan untuk clustering: usia (numerik, kontinu, rentang 14-84 tahun), gender (kategorikal: Male/Female), dan kategori produk utama (kategorikal: 9 kategori termasuk Makanan, Minuman, Kebutuhan Dapur, Kebutuhan Rumah, Lifestyle, dan sejenisnya).

Kualitas data dalam penelitian ini dipertahankan pada standar yang sangat tinggi melalui beberapa protokol validasi. Tidak ada missing values yang terdeteksi di seluruh dataset, yang berarti setiap transaksi memiliki informasi lengkap untuk ketiga fitur utama. Distribusi data seimbang pada setiap fitur, menunjukkan representasi yang merata dari berbagai segmen demografis dan kategori produk. Pemeriksaan outlier telah dilakukan menggunakan metode Interquartile Range (IQR), sebuah pendekatan robust yang mengidentifikasi data points yang berada di luar rentang $Q1 - 1.5 \times IQR$ hingga $Q3 + 1.5 \times IQR$. Statistik deskriptif mengungkapkan bahwa rata-rata usia pelanggan adalah 45,2 tahun dengan standar deviasi 10,8 tahun, menunjukkan distribusi usia yang tersebar luas namun terkonsentrasi di sekitar kelompok umur pertengahan. Proporsi gender juga mencerminkan keseimbangan yang baik dengan 52,3% pelanggan perempuan dan 47,7% pelanggan laki-laki.

Karakteristik umur dan simpangannya diperiksa dahulu di karena analisis deskriptif (seperti rata-rata, minimum–maksimum, dan standar deviasi) memberikan gambaran awal tentang sebaran dan keragaman data umur sebelum masuk ke tahap clustering. Informasi ini penting untuk memastikan tidak ada pola yang janggal, rentang yang terlalu ekstrem, atau variasi yang terlalu kecil/besar yang bisa mengganggu proses normalisasi dan perhitungan jarak dalam algoritma K-Prototypes. Dengan memahami distribusi dan simpangan umur sejak awal, peneliti dapat memastikan bahwa variabel umur dalam kondisi “optimal” dan layak digunakan sehingga hasil cluster yang terbentuk menjadi lebih valid dan interpretatif.

Pengecekan kualitas sumber data perlu dilakukan untuk memastikan bahwa data yang akan digunakan dalam proses clustering bersih dan bebas dari masalah seperti missing values, karena kondisi tersebut dapat mengganggu perhitungan

jarak, pembentukan centroid, dan akhirnya mempengaruhi kualitas pembentukan cluster secara keseluruhan. Jika missing values dibiarkan, algoritma clustering berpotensi menghasilkan kelompok yang tidak representatif atau bias, sehingga interpretasi segmen pelanggan yang dihasilkan menjadi kurang akurat dan keputusan bisnis yang diambil berdasarkan hasil tersebut bisa keliru.

Strategi preprocessing yang diterapkan dalam penelitian ini mengikuti prinsip fundamental dalam K-Prototypes: meminimalkan kehilangan informasi asli pada data sambil memastikan kompatibilitas algoritma. Untuk fitur numerik (usia), normalisasi dilakukan menggunakan MinMaxScaler, sebuah teknik yang mengubah nilai ke dalam rentang standar menggunakan formula Xscaled = (X−Xmin) / (Xmax−Xmin). Pendekatan ini dipilih karena MinMaxScaler menjaga distribusi data asli dan menghindari distorsi informasi yang sering terjadi dengan metode standarisasi lain. Fitur kategorikal (gender dan kategori produk) dibiarkan dalam bentuk string kategorikal tanpa encoding tambahan, sebuah keputusan desain yang mempertahankan struktur data campuran dan menjaga makna semantik dari kategori tersebut.

Normalisasi umur diperlukan agar kontribusi fitur umur tidak terlalu besar atau terlalu kecil dibanding variabel lain ketika dihitung dalam fungsi cost K-Prototypes. Dengan menormalkan umur ke rentang 0–1, skala fitur numerik (umur) disejajarkan dengan skala fitur kategorikal yang juga direpresentasikan dalam rentang 0–1 (misalnya melalui distance 0/1). Hal ini membuat bobot atau dampak setiap variabel terhadap perhitungan jarak dan penentuan cluster menjadi lebih seimbang, sehingga hasil pengelompokan tidak didominasi oleh variabel tertentu hanya karena skala angkanya lebih besar.

Parameter gamma (γ), yang merupakan faktor penyeimbang kontribusi fitur numerik terhadap kategorikal dalam distance metric K-Prototypes, dihitung secara sistematis menggunakan formula $\gamma = 0{,}5 \times SD_{numerik\_normalized}$. Dalam penelitian ini, nilai γ yang diperoleh adalah 0,0768, yang mencerminkan kontribusi proporsional dari dimensi numerik dan kategorikal dalam perhitungan jarak. Pemilihan formula ini didasarkan pada penelitian K-Prototypes yang

menunjukkan bahwa scaling gamma berdasarkan standar deviasi fitur numerik menghasilkan keseimbangan optimal dalam penilaian similarity. Hasil akhir dari preprocessing menghasilkan struktur array fitur yang terorganisir dengan Usia (normalized), Gender, dan Main_Category, di mana fitur kategorikal ditempatkan di posisi indeks ke-1 dan ke-2 untuk optimalisasi komputasi.

Keunggulan utama dari pendekatan preprocessing ini terletak pada minimnya kehilangan informasi; usia pelanggan tetap terekam secara penuh tanpa pengelompokan kasar, memungkinkan setiap variasi usia untuk membedakan profil pelanggan dengan detail yang optimal. Metode ini sangat tepat untuk pengolahan data e-commerce karena mampu memanfaatkan kekayaan data campuran tanpa risiko bias akibat konversi atau reduksi fitur yang tidak perlu. Dengan mempertahankan struktur data ini, stabilitas dan kualitas hasil clustering dapat dijaga di seluruh eksperimen dengan ketiga dataset dari tiga bulan berbeda, memastikan konsistensi dalam analisis.

3.4.2 Cari K terbaik, dari Elbow Method , Cari Metode Init terbaik

Penentuan jumlah cluster (K) yang optimal merupakan komponen kritis dalam penelitian ini, dan dilakukan melalui analisis komprehensif menggunakan multiple evaluation metrics. Pendekatan multi-metrik ini dipilih karena tidak ada satu metrik tunggal yang sempurna untuk semua situasi; setiap metrik menawarkan perspektif berbeda tentang kualitas clustering. Penelitian ini menguji rentang K dari 2 hingga 10 menggunakan algoritma K-Prototypes standar, kemudian menganalisis hasil menggunakan empat metrik evaluasi utama: Elbow Method, Silhouette Score, Davies Bouldin Index, dan Calinski Harabasz Index.

Aspek	Silhouette Score	Davies Bouldin Index	Calinski Harabazs Index
Tujuan	Mengukur seberapa baik tiap titik cocok dengan clusternya dibanding cluster	Mengukur rata-rata rasio dispersi intra-cluster terhadap jarak	Mengukur rasio antara dispersi antar-cluster dan intra-cluster secara global.

	lain.	antar-cluster terdekat.	
Rentang Nilai	(-1, 1)	[0, ∞)	(0, ∞)
Interpretasi Umum	Positif tinggi: cluster terpisah dan kompak; sekitar 0: overlap; negatif: banyak titik salah cluster.	Nilai rendah: cluster kompak dan saling berjauhan; nilai tinggi: cluster tumpang tindih/tidak jelas.	Nilai tinggi: variasi antar-cluster besar dan dalam-cluster kecil; nilai rendah: cluster kurang terpisah.

Tabel 3.1 Perbandingan metrik evaluasi clustering

Ketiga metrik evaluasi internal (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index) digunakan pada dua tahap dalam penelitian ini:

1. Tahap pemilihan metode inisialisasi: Metrik dihitung untuk membandingkan performa metode inisialisasi Huang vs Cao pada K optimal yang telah dipilih

2. Tahap evaluasi final clustering: Metrik dihitung pada hasil final ensemble clustering di ketiga dataset untuk memvalidasi kualitas clustering yang dihasilkan pendekatan evaluasi bertahap ini memastikan bahwa setiap keputusan parameter didasarkan pada bukti yang kuat dan hasil akhir clustering memiliki validitas statistik yang tinggi.

3.4.2.1 Elbow Method

Elbow Method bekerja dengan menghitung cost function (within-cluster sum of squares) untuk setiap nilai K dan mengidentifikasi titik "elbow" di mana penurunan cost mulai melambat. Dalam konteks K-Prototypes, cost function mengintegrasikan perhitungan untuk fitur numerik dan kategorikal, memberikan representasi keseluruhan dari kualitas kluster. Berikut adalah formula yang digunakan untuk menghitung cost pada jumlah cluster K yang digunakan untuk data numerik dan kategorikal :

Numerik

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \quad (1)$$

dimana :

- k adalah jumlah cluster
- C_i adalah anggota cluster ke-i
- x adalah data point
- μ_i adalah centroid cluster ke-i
- $||x - \mu_i||^2$ adalah jarak Euclidean secara kuadrat antara data point dan centroid cluster-nya

Kategorikal

$$Cost = \sum_{i=1}^{k} \sum_{x \in C_i} \sum_{j=1}^{m} \delta(x_j, y_{ij}) \quad (2)$$

dimana :

- k adalah jumlah cluster
- C_i adalah anggota dari cluster ke-i
- x_j adalah nilai atribut j pada data point x
- y_{ij} adalah mode atribut j pada cluster i
- $\delta(x_j, y_{ij})$ bernilai 0 jika atribut sama dan bernilai 1 jika atribut beda (fungsi dissimilarity kategorikal)

3.4.2.2 Silhouette Score

Silhouette Score adalah cara untuk mengukur seberapa baik kualitas hasil pengelompokan (clustering) pada sebuah dataset [2]. Skor ini membantu menilai seberapa tepat setiap titik data ditempatkan dalam clusternya. Jika nilainya mendekati 1, berarti sebuah titik data sangat cocok dengan clusternya dan jauh dari cluster lain. Jika nilainya mendekati 0, berarti titik data berada di perbatasan antara dua cluster. Jika nilainya mendekati -1, berarti titik data kemungkinan besar

berada di cluster yang kurang tepat (salah kelompok). Berikut adalah formula yang digunakan untuk menghitung Silhouette Score pada jumlah cluster K yang digunakan :

$$S(i) = \frac{b(i) - a(i)}{max(a(i),b(i))} \quad (3)$$

dimana :

- a(i) adalah rata-rata jarak dari i ke data points lain pada cluster yang sama
- b(i) adalah jarak dengan rata-rata terkecil dari i ke data points pada cluster yang berbeda

3.4.2.3 Davies Bouldin Index

Davies-Bouldin Index (DBI) digunakan untuk mengukur seberapa baik kualitas hasil clustering pada suatu dataset [2]. Indeks ini melihat dua hal utama: seberapa rapat (kompak) titik-titik dalam satu cluster, dan seberapa jauh jarak antar cluster satu dengan yang lain. DBI yang rendah berarti cluster yang terbentuk lebih baik dan lebih jelas (tidak saling tumpang tindih). DBI yang tinggi menunjukkan cluster cenderung berantakan dan saling tumpang tindih. Nilai yang lebih kecil lebih baik, karena titik-titik dalam satu cluster saling berdekatan dan jarak antar cluster saling berjauhan. Berikut adalah formula yang digunakan untuk menghitung Davies Bouldin Index pada jumlah cluster K yang digunakan :

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i}\left(\frac{R_{ii}+R_{jj}}{R_{ij}}\right) \quad (4)$$

di mana :

- k adalah jumlah total cluster.
- R_{ii} adalah tingkat kepadatan cluster i.
- R_{jj} adalah tingkat kepadatan cluster j.

- R_{ij} adalah dissimilaritas antar cluster i dan j

3.4.2.4 Calinski Harabazs Index

Calinski-Harabasz Index digunakan untuk mengukur seberapa baik kualitas cluster pada suatu dataset. Indeks ini melihat dua hal utama: seberapa rapat titik-titik di dalam setiap cluster, dan seberapa jauh jarak antar cluster [2]. Semakin tinggi nilai Calinski-Harabasz, semakin baik kualitas cluster nya, karena cluster menjadi lebih kompak dan lebih terpisah satu sama lain. Indeks ini sering digunakan untuk membantu menentukan jumlah cluster yang paling ideal, dengan memilih jumlah cluster yang memberikan nilai indeks tertinggi. Berikut adalah formula yang digunakan untuk menghitung Calinski-Harabasz Index pada jumlah cluster K yang digunakan :

$$CH = \frac{B}{W} \times \frac{N-K}{K-1} \quad (5)$$

dimana :

- B adalah jumlah kuadrat antar cluster.
- W adalah jumlah kuadrat dalam cluster.
- N adalah jumlah total data points
- K adalah jumlah total cluster

$$B = \sum_{k=1}^{K} n_k \times ||C_k - C||^2 \quad (6)$$

dimana :

- n_k adalah jumlah observasi dari cluster 'k'
- C_k adalah centroid dari cluster 'k'
- C adalah centroid dari dataset
- K adalah jumlah cluster

$$W = \sum_{i=1}^{n_k} ||X_{ik} - C_k||^2 \quad (7)$$

dimana :

- n_k adalah jumlah observasi di cluster 'k'
- X_{ik} adalah jumlah observasi ke-i di cluster 'k'
- C_k adalah centroid dari cluster 'k'

3.4.3 Implementasi Algoritma Ensemble K-Prototypes

Setelah penentuan nilai K dan metode inisialisasi yang optimal, penelitian melanjutkan dengan implementasi algoritma K-Prototypes yang diintegrasikan dengan AdaBoost Classification Layer untuk meningkatkan akurasi dan konsistensi hasil clustering. K-Prototypes adalah algoritma clustering yang menggabungkan K-Means (untuk fitur numerik) dan K-Modes (untuk fitur kategorikal), menghasilkan metode yang robust untuk data campuran.

3.4.4 Evaluasi Konsistensi Label Cluster Antar Periode

Karena penelitian ini mengimplementasikan clustering secara independen pada tiga dataset bulanan (Juni, Juli, Agustus 2025), terdapat potensi terjadinya label switching problem—yaitu situasi di mana cluster dengan karakteristik serupa dapat memiliki label numerik berbeda di periode berbeda. Misalnya, segmen pelanggan perempuan muda yang fokus pada produk makanan dapat terlabel sebagai Cluster 0 di bulan Juni tetapi menjadi Cluster 2 di bulan Juli, meskipun karakteristik demografis dan preferensi produknya identik.

Untuk mengatasi potensi masalah ini dan memastikan konsistensi interpretasi hasil antar periode, penelitian menerapkan pendekatan validasi konsistensi berbasis profiling karakteristik cluster. Setelah proses clustering pada ketiga periode selesai, dilakukan analisis perbandingan mendalam terhadap profil setiap cluster menggunakan tiga dimensi utama:

1. Distribusi Usia: Analisis box plot untuk membandingkan median, quartile range, dan outliers distribusi usia setiap cluster di ketiga periode

2. Komposisi Gender: Analisis proporsi gender (Male/Female) di setiap cluster untuk mengidentifikasi pola dominasi gender

3. Preferensi Kategori Produk: Analisis distribusi kategori produk utama untuk mengidentifikasi pola pembelian yang konsisten

Jika profil karakteristik suatu cluster di periode berikutnya sesuai dengan profil cluster di periode sebelumnya berdasarkan ketiga dimensi tersebut, maka cluster tersebut dianggap merepresentasikan segment pelanggan yang sama. Pendekatan ini memastikan bahwa ketika analisis merujuk pada "Cluster 0" dalam pembahasan lintas waktu, yang dimaksud adalah segmen pelanggan dengan karakteristik konsisten di ketiga periode, bukan hanya label numerik yang arbitrary.

Validasi konsistensi dilakukan melalui overlay visualisasi side-by-side dari ketiga periode, di mana konsistensi pola demografis dan preferensi produk dapat diverifikasi secara visual dan statistik. Pendekatan ini dipilih karena transparansi dan kemudahan interpretasinya dalam konteks bisnis e-commerce, di mana stakeholder perlu memahami evolusi segment pelanggan dari waktu ke waktu untuk pengambilan keputusan marketing yang efektif.

3.4.5 Evaluasi Hasil Clustering dan Customer Segment Profiling

Setelah implementasi algoritma pada ketiga dataset, proses evaluasi dilakukan secara komprehensif untuk menginterpretasikan hasil clustering yang diperoleh. Setiap cluster yang terbentuk dianalisis berdasarkan karakteristik unik yang mencerminkan profil pelanggan yang berbeda-beda. Evaluasi ini dilaksanakan melalui dua dimensi utama, yaitu evaluasi kuantitatif menggunakan metrik statistik yang telah diidentifikasi sebelumnya serta visualisasi kualitatif untuk mendukung interpretasi intuitif struktur cluster. Salah satu visualisasi kunci adalah bar plot, yang digunakan untuk menampilkan distribusi fitur kategorikal seperti gender dan kategori produk di masing-masing cluster. Visualisasi tersebut sangat efektif dalam memberikan gambaran cepat mengenai komposisi demografis setiap segmen dan memudahkan perbandingan proporsi antar kelompok. Misalnya, bar plot dapat memperlihatkan perbedaan proporsi

pelanggan perempuan dan laki-laki secara jelas di antara cluster, sehingga mengungkapkan adanya preferensi demografis yang signifikan pada segmen tertentu.

Selain bar plot, box plot digunakan untuk menunjukkan distribusi usia di setiap cluster, termasuk informasi median, kuartil, serta keberadaan outlier. Dengan analisis box plot, dapat dipahami kecenderungan sentral, sebaran, dan potensi outlier pada fitur numerik usia, sehingga memudahkan identifikasi cluster dengan rentang usia luas maupun segmen yang menargetkan kelompok usia tertentu. Analisis lebih lanjut dilakukan secara statistik untuk menguji perbedaan signifikan karakteristik pada setiap cluster. Salah satu fokus utama adalah mengevaluasi perkembangan atau penurunan ukuran cluster secara berkala, yaitu dengan membandingkan ukuran cluster dan volume transaksi dari bulan ke bulan. Pengecekan ini dihitung menggunakan growth rate dan trend analysis seperti simple moving average atau linear regression, dengan tujuan mengidentifikasi segmen yang mengalami pertumbuhan, penurunan, atau potensi shifting ke segmen lain.

3.4.6 Cluster Analysis

Analisis karakteristik cluster pada hasil clustering berbasis K-Prototypes dan AdaBoost memungkinkan perusahaan memahami perbedaan mendasar antara kelompok konsumen berdasarkan ciri demografis dan perilaku belanja. Setiap cluster diprofilkan secara mendetail dengan melihat kombinasi fitur numerik (seperti usia) dan kategorikal (seperti gender dan preferensi kategori produk). Hal ini menghasilkan deskripsi segmen yang tegas—misalnya, cluster yang mayoritasnya perempuan muda dengan preferensi lifestyle dapat diidentifikasi sebagai basis calon pelanggan fashion, sementara cluster laki-laki paruh baya dapat dikenali sebagai pelanggan dengan pola belanja lebih praktis pada produk kebutuhan rumah.

Selanjutnya, penelitian mendalami pola perilaku (behavioral patterns) di setiap segmen. Ini mencakup pengamatan atas preferensi musiman, kecenderungan membeli lintas kategori produk, serta waktu pembelian utama

yang membedakan tiap kelompok. Dengan pemahaman tersebut, perusahaan bisa menyesuaikan stok barang, mengatur peluncuran promosi berbasis waktu, serta menyusun rekomendasi produk yang lebih relevan untuk masing-masing segmen. Penilaian clustering berdasarkan value juga dilakukan untuk menandai cluster yang paling berkontribusi terhadap pendapatan atau berpotensi memiliki lifetime value tinggi, sekaligus mengenali segmen dengan risiko churn sehingga intervensi lebih awal dapat diterapkan.

Rangkaian insight tersebut digunakan untuk mengembangkan rekomendasi pemasaran spesifik berbasis karakteristik clustering multi-dimensi. Setiap strategi mempertimbangkan demografi (seperti penyusunan pesan dan pemilihan kanal komunikasi berbeda antara pelanggan muda dan dewasa), perilaku pembelian (frekuensi, loyalitas, atau respons terhadap promosi), hingga kebutuhan unik tiap segmen (misalnya, clustering untuk pelanggan berbasis fungsi praktis vs. gaya hidup). Tindakan lanjutan berupa personalisasi produk, kustomisasi promosi, serta program retensi dan akuisisi pelanggan didesain khusus menyesuaikan profil tiap cluster untuk mengoptimalkan engagement dan konversi.

3.5 Metode Pengujian

3.5.1 Pengujian Efisiensi Waktu

Pengukuran efisiensi waktu untuk menjawab pertanyaan penelitian pertama dilakukan dengan mencatat durasi eksekusi pada setiap tahap utama proses clustering dan ensemble menggunakan library Python `time.time()`. Proses pengukuran waktu mencakup lima komponen: durasi preprocessing data, waktu eksekusi bootstrap ensemble (menggunakan 10 subset data), waktu meta-clustering, waktu training AdaBoost layer, serta waktu total end-to-end dari awal hingga akhir. Pengukuran waktu dilakukan dengan mencatat timestamp sebelum dan sesudah setiap tahap utama proses, menggunakan fungsi time.time() dari library Python. Durasi setiap tahap kemudian dihitung sebagai selisih timestamp dan disimpan dalam struktur data timing_breakdown untuk analisis efisiensi komputasi. Pengukuran dilakukan pada satu complete pipeline execution

untuk menghindari overhead waktu yang tidak perlu dan mencerminkan kondisi operasional aktual.

Setelah visualisasi Elbow Method dihasilkan dan peneliti memilih nilai K optimal berdasarkan identifikasi visual elbow point, pipeline secara otomatis mengeksekusi seluruh tahap berikutnya (evaluasi init method, clustering ensemble, validasi AdaBoost, visualisasi, dan statistik deskriptif) tanpa intervensi manual lebih lanjut. Pengukuran waktu mencakup seluruh tahap otomatis ini untuk mencerminkan efisiensi komputasi end-to-end dari pipeline.

Target keberhasilan utama adalah tercapainya waktu total proses kurang dari satu hari kerja penuh, yaitu 8 jam (setara dengan 28.800 detik). Untuk menilai efisiensi, dilakukan perbandingan antara waktu eksekusi metode ensemble dengan metode K-Prototypes standar, sehingga dapat dihitung rasio waktu sebagai ukuran trade-off. Analisis lebih lanjut juga menyoroti aspek justifikasi: tambahan waktu komputasi yang muncul akibat penggunaan ensemble dibandingkan metode standar harus dibenarkan oleh adanya peningkatan substansial pada kualitas hasil clustering, kestabilan clustering, dan konsistensi model.

Target maksimal 8 jam (setara dengan satu hari kerja penuh) dipilih berdasarkan definisi standar jam kerja operasional PT. XYZ dan kebutuhan bisnis untuk mendapatkan hasil clustering dalam waktu yang responsif. Dengan target ini, proses clustering dapat diselesaikan dalam satu shift kerja, memungkinkan tim marketing untuk segera mengimplementasikan strategi berbasis hasil clustering pada hari yang sama.

Approach yang digunakan dalam pelaporan efisiensi waktu sangat sesuai untuk penelitian berbasis data besar pada domain e-commerce, sebab transparansi durasi proses dan penegakan target waktu memungkinkan penelitian tetap relevan secara praktis untuk implementasi bisnis. Selain itu, pengulangan eksekusi dan pelaporan nilai rata-rata hasil pengukuran memberikan validasi yang baik terhadap reliability hasil eksperimen, meminimalisir dampak outliers dan noise pada performa komputasi. Analisis rasio dan trade-off antara kualitas dan waktu

komputasi menjadi landasan penting dalam pengambilan keputusan adopsi algoritma untuk implementasi industri atau operasional harian.

3.5.2 Pengujian Kualitas Clustering

Evaluasi kualitas clustering dalam penelitian ini dilakukan menggunakan tiga metrik internal yang tidak memerlukan ground truth labels, sehingga sangat cocok untuk data e-commerce yang memang umumnya tidak memiliki label kelas bawaan. Silhouette Score digunakan untuk mengukur seberapa baik setiap data point telah terkluster, dengan rentang nilai antara -1 hingga 1 di mana nilai di atas 0,5 menandakan clustering yang sangat baik; metrik ini menjadi indikator utama untuk menilai struktur cluster yang terbentuk. Davies-Bouldin Index, dengan rentang nilai mulai dari 0 ke atas (semakin kecil semakin baik), mengukur rasio antara scatter dalam cluster dan pemisahan antar cluster—nilai di bawah 1,0 menandakan cluster yang terpisah dengan baik, sehingga model dianggap berhasil jika threshold ini terpenuhi. Sementara itu, Calinski-Harabasz Index memfokuskan pada rasio antara variansi antar cluster dengan variansi dalam cluster; semakin tinggi nilainya, semakin baik clustering yang dihasilkan.

Ketiga metrik ini dihitung secara spesifik untuk hasil Ensemble K-Prototypes, memberikan gambaran menyeluruh mengenai kualitas hasil clustering pelanggan. Dengan mengaplikasikan ketiga metode ini secara bersamaan, peneliti dapat memvalidasi pemisahan antar segmen secara statistik sekaligus mendapatkan justifikasi ilmiah bahwa cluster yang terbentuk memiliki validitas dan interpretabilitas yang tinggi, bahkan tanpa keberadaan label referensi. Pendekatan ini memperkuat argumen bahwa clustering yang dihasilkan efektif digunakan sebagai dasar analisis bisnis dan rekomendasi marketing selanjutnya.

Metrik-metrik internal clustering ini (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index) memerlukan perhitungan jarak (distance) antar data points dalam ruang numerik. Karena dataset penelitian mengandung fitur kategorikal (Gender dan Main_Category), diperlukan preprocessing tambahan berupa encoding fitur kategorikal ke bentuk numerik menggunakan LabelEncoder

sebelum perhitungan metrik. Encoding ini hanya digunakan untuk tujuan evaluasi metrik dan tidak mengubah proses clustering itu sendiri, karena K-Prototypes mampu menangani data campuran secara native melalui dissimilarity measure yang menggabungkan Euclidean distance (untuk fitur numerik) dan simple matching distance (untuk fitur kategorikal).

3.5.3 Pengujian Validasi Cluster (AdaBoost Classification)

Pengujian ini dilakukan untuk memastikan bahwa cluster yang dihasilkan benar-benar terpisah dengan jelas dan dapat digunakan sebagai basis pengambilan keputusan bisnis yang actionable. Prosedur yang digunakan adalah mengambil label cluster hasil clustering sebagai target variabel klasifikasi, kemudian membagi data menjadi 70% untuk pelatihan dan 30% untuk pengujian. Model AdaBoost classifier dengan n_estimators sebanyak 300 dan base classifier Decision Tree dilatih pada data ini, lalu dievaluasi menggunakan empat metrik utama yaitu accuracy, precision (weighted average), recall (weighted average), dan F1-score (weighted average).

Parameter n_estimators = 300 dipilih berdasarkan praktik umum dalam implementasi AdaBoost untuk dataset berukuran medium (>20,000 samples), di mana jumlah estimator dalam rentang 100-500 umumnya menghasilkan convergence yang baik tanpa overfitting signifikan. Nilai 300 memberikan keseimbangan antara akurasi prediksi dan waktu training, memastikan bahwa evaluasi separability cluster dapat dilakukan dengan reliable namun tetap efisien.

Interpretasi hasil evaluasi berfokus pada nilai accuracy, di mana accuracy di atas 85% menunjukkan cluster yang well-separated sebagai segmen bisnis; accuracy antara 75-85% mengindikasikan cluster yang cukup baik namun ada kemungkinan overlap, sementara accuracy di bawah 75% menandakan clustering yang kurang optimal dan cluster cenderung overlapping. Target keberhasilan utama adalah tercapainya accuracy di atas 85% pada model ensemble K-Prototypes, dengan harapan nilai ini signifikan lebih tinggi dibanding metode standar yang tidak menggunakan ensemble, sehingga membuktikan manfaat tambahan dari pendekatan ensemble dalam clustering pelanggan e-commerce.

Dalam penelitian ini, proses perhitungan akurasi clustering dilakukan dengan memanfaatkan label cluster yang dihasilkan oleh ensemble K-Prototypes sebagai "kelas" pada tahap berikutnya. Data yang sudah memiliki label cluster tersebut kemudian digunakan sebagai data latih untuk membangun model AdaBoost, sehingga model belajar memetakan atribut-atribut pelanggan terhadap label cluster hasil ensemble. Setelah model terlatih, sebagian data yang disisihkan sebagai data uji digunakan untuk mengukur seberapa baik AdaBoost mampu memprediksi kembali label cluster tersebut, dan nilai akurasi prediksi inilah yang dijadikan ukuran akurasi dari hasil clustering ensemble K-Prototypes.

Evaluasi berbasis metrik classification ini sangat relevan karena memberikan bukti kuantitatif bahwa hasil clustering memang dapat diprediksi dan dipisahkan secara konsisten oleh classifier yang robust. Dengan demikian, clustering yang dihasilkan tidak hanya valid secara statistik, tetapi juga praktis dan efektif untuk strategi pemasaran, retensi, serta penargetan spesifik pada bisnis e-commerce.