BAB II
KAJIAN LITERATUR

2.1 Segmentasi Konsumen

Segmentasi konsumen merupakan langkah krusial dalam strategi pemasaran, yaitu proses membagi pasar atau kumpulan konsumen menjadi kelompok-kelompok yang lebih kecil berdasarkan kesamaan atribut, kebutuhan, atau perilaku [13]. Dalam konteks *e-commerce*, segmentasi membantu memahami keberagaman perilaku konsumen dan merancang pendekatan yang lebih personal [14]. Segmentasi yang efektif terbukti dapat meningkatkan efektivitas pemasaran dan memperkuat hubungan jangka panjang antara bisnis dan konsumen [15].

Secara umum, metode segmentasi dibagi menjadi tiga kategori utama, meskipun dalam praktiknya sering digunakan secara terintegrasi: [14].

1. Segmentasi Demografis, Metode ini membagi konsumen berdasarkan karakteristik demografis seperti usia, jenis kelamin, pendapatan, pendidikan, dan status pernikahan. Segmentasi ini memberikan gambaran awal yang berguna tentang karakteristik konsumen , tetapi sering kali tidak cukup untuk menggali motivasi dan preferensi yang lebih dalam [14]. Misalnya, dua konsumen dengan usia dan pendapatan yang sama mungkin memiliki perilaku pembelian yang sangat berbeda.
2. Segmentasi Psikografis, Pendekatan ini mengelompokkan konsumen berdasarkan faktor psikologis seperti nilai, sikap, minat, dan gaya hidup. Segmentasi psikografis memberikan wawasan yang lebih mendalam tentang motivasi dan preferensi konsumen , memungkinkan bisnis untuk merancang strategi pemasaran yang lebih sesuai dengan kepribadian dan kebutuhan emosional mereka [14] [16]. Namun, pengumpulan data psikografis seringkali lebih sulit dan memerlukan metode penelitian yang lebih kompleks.
3. Segmentasi Perilaku, Metode ini berfokus pada pola pembelian dan interaksi konsumen , seperti frekuensi pembelian, loyalitas, dan respons terhadap promosi [14] [16]. Segmentasi perilaku memungkinkan bisnis untuk mengidentifikasi pola yang mungkin tidak terlihat dalam segmentasi demografis atau psikografis. Misalnya, konsumen dengan karakteristik demografis yang sama dapat memiliki perilaku pembelian yang sangat berbeda, sehingga pendekatan berbasis perilaku dapat menghasilkan segmentasi yang lebih relevan dan efektif.

Segmentasi perilaku adalah dasar bagi taktik *e-commerce* modern, seperti *email marketing* dan rekomendasi produk yang dipersonalisasi berdasarkan riwayat transaksi konsumen [15]. Meskipun demikian, tantangan besar tetap ada dalam mengelola data transaksi yang besar dan kompleksitas perilaku konsumen yang terus berubah [17]. Oleh karena itu, penelitian ini berfokus pada pendekatan segmentasi perilaku dengan mengintegrasikan model RFM dan algoritma *clustering* untuk menghasilkan segmentasi yang optimal dan mendukung pengambilan keputusan berbasis data.

2.2 *Knowledge Discovery in Database* (KDD)

Metodologi *Knowledge Discovery in Database* (KDD) pertama kali diperkenalkan untuk menggambarkan proses pencarian pengetahuan dalam data yang melibatkan penerapan algoritma data mining secara spesifik [16]. Secara teoretis, KDD didefinisikan sebagai proses non-trivial untuk mengidentifikasi pola yang valid, baru, berpotensi berguna, dan pada akhirnya dapat dipahami dalam kumpulan data yang besar dan kompleks [18].

KDD bukan sekadar aktivitas tunggal, melainkan sebuah siklus berulang (iteratif) yang bertujuan untuk mengekstraksi informasi berharga [18]. Hubungan antara KDD dan data mining seringkali disalahpahami dalam kerangka teoretis ini, data mining merupakan salah satu tahapan krusial di dalam proses KDD, namun bukan merupakan keseluruhan proses itu sendiri. Kerangka kerja KDD menyediakan struktur yang memastikan data diolah dengan benar sebelum model diterapkan, dan hasil model dievaluasi sebelum dijadikan dasar pengambilan keputusan. Proses KDD terdiri dari lima tahapan utama yang dijelaskan secara mendetail sebagai berikut:

2.2.1 Data *Selection* (Seleksi Data)

Tahap pertama, *Selection*, merupakan proses pemilihan data yang relevan dari sumber data yang tersedia. Dalam penelitian ini, data yang digunakan berasal dari transaksi konsumen pada platform *e-commerce*, yang mencakup informasi seperti tanggal pembelian, jumlah transaksi, total nilai pembelanjaan, serta jenis produk yang dibeli. Tahap ini bertujuan memastikan bahwa data yang dipilih sesuai dengan tujuan analisis, yaitu untuk mengeksplorasi pola segmentasi konsumen menggunakan model RFM (*Recency*, *Frequency*, *Monetary*) yang akan diolah lebih lanjut dengan algoritma *clustering* [19].

2.2.2 Data *Preprocessing* (Pra-pemrosesan Data)

Tahap berikutnya adalah *Preprocessing*, yaitu tahap pembersihan dan penyiapan data agar layak digunakan dalam proses analisis. Pada tahap ini dilakukan penghapusan data, koreksi terhadap data yang hilang atau tidak konsisten, serta strisasi format agar data

memiliki kualitas yang baik. Tahap ini penting karena hasil analisis yang akurat hanya dapat diperoleh dari data yang bersih dan valid. Dalam konteks *e-commerce*, proses ini memastikan bahwa setiap catatan transaksi konsumen dapat mencerminkan kondisi sebenarnya tanpa adanya kesalahan pencatatan atau duplikasi [19].

2.2.3 Data Transformation (Transformasi Data)

Selanjutnya, tahap Transformation dilakukan untuk mengubah data mentah menjadi format yang sesuai dengan kebutuhan analisis. Pada penelitian ini, transformasi dilakukan dengan menghitung nilai *Recency* (R), *Frequency* (F), dan *Monetary* (M) dari masing-masing konsumen . Ketiga nilai tersebut merupakan representasi kuantitatif dari perilaku konsumen dan menjadi variabel utama dalam proses pengelompokan. Melalui tahap ini, data transaksi yang sebelumnya belum terstruktur diubah menjadi informasi yang dapat digunakan untuk mengidentifikasi karakteristik dan perilaku konsumen secara lebih sistematis [19].

2.2.4 Data Mining

Tahap inti dari proses KDD adalah Data Mining, dimana berbagai teknik analisis diterapkan untuk menemukan pola atau struktur tersembunyi dalam data [19]. Dalam penelitian ini, tahap data mining dilakukan dengan menerapkan beberapa algoritma *clustering*, yaitu *K-Means*, DBSCAN, *Gaussian Mixture Model* (GMM), dan *Mean-Shift*, yang digunakan untuk mengelompokkan konsumen berdasarkan nilai RFM. Setiap algoritma memiliki karakteristik dan keunggulan tersendiri dalam membentuk kelompok data, sehingga hasil yang diperoleh dapat dibandingkan untuk menentukan model segmentasi konsumen yang paling optimal [19]. Penelitian sebelumnya menunjukkan bahwa penerapan algoritma *clustering* dalam data mining dapat meningkatkan pemahaman tentang perilaku konsumen secara signifikan [16]. Misalnya, penerapan *K-Means* untuk segmentasi konsumen dan menemukan bahwa hasilnya lebih baik dibandingkan dengan metode tradisional yang tidak menggunakan data mining [16]. Namun, penelitian tersebut juga mengakui bahwa *K-Means* memiliki keterbatasan dalam menangani data yang tidak terdistribusi secara normal.

Di sisi lain, penelitian lainnya menerapkan DBSCAN dan menemukan bahwa algoritma ini lebih efektif dalam mengidentifikasi kelompok konsumen dengan pola transaksi yang tidak teratur [20]. Meskipun demikian, penelitian ini juga menunjukkan bahwa DBSCAN memerlukan parameter yang tepat untuk menghasilkan hasil yang optimal, yang dapat menjadi tantangan tersendiri [20]. Penelitian lain yang menggunakan *Gaussian*

Mixture Model (GMM), menekankan bahwa model ini mampu menangkap kompleksitas data yang lebih tinggi, tetapi juga memerlukan waktu komputasi yang lebih lama [21].

2.2.5 Evaluation (Evaluasi)

Setelah menerapkan algoritma *clustering*, Tahap terakhir adalah Evaluation, yaitu proses penilaian terhadap hasil *clustering* yang telah dilakukan. Tujuannya untuk memastikan bahwa kelompok konsumen yang terbentuk benar-benar mencerminkan perbedaan perilaku yang bermakna [19]. Evaluasi dalam penelitian ini dilakukan menggunakan beberapa metrik kuantitatif, seperti *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Metrik-metrik tersebut digunakan untuk mengukur tingkat kekompakan antaranggota dalam satu kluster dan keterpisahan antar kluster, sehingga dapat diketahui algoritma mana yang memberikan hasil segmentasi terbaik [19].

2.3 Model Analisis RFM (*Recency*, *Frequency*, *Monetary*)

Model *Recency*, *Frequency*, *Monetary* (RFM) merupakan salah satu pendekatan analitis yang banyak digunakan dalam bidang pemasaran berbasis data untuk mengukur dan mengevaluasi nilai konsumen berdasarkan perilaku transaksionalnya [22]. Model ini menilai konsumen melalui tiga dimensi utama, yaitu *Recency* (R), *Frequency* (F), dan *Monetary* (M) yang bersama-sama memberikan gambaran mengenai seberapa aktif, sering, dan bernilai seorang konsumen bagi perusahaan.

2.3.1 Dimensi *Recency* (R)

Dimensi *Recency* mengukur seberapa baru konsumen melakukan transaksi terakhirnya. Konsumen yang baru saja bertransaksi memiliki tingkat keterlibatan dan potensi pembelian kembali yang lebih tinggi. Nilai R dihitung sebagai selisih waktu antara tanggal acuan (analisis) dan tanggal transaksi terakhir konsumen , di mana nilai yang kecil menunjukkan kondisi yang lebih baik [22].

$$R_i = T_{ref} - T_{last(i)} \quad \text{……} (1)$$

Dimana :

R_i = nilai *Recency* konsumen ke-i

T_{ref} = tanggal acuan atau tanggal analisis

$T_{last(i)}$ = tanggal transaksi terakhir konsumen ke-i

Penentuan nilai *Recency* didasarkan pada perbandingan relatif antar waktu transaksi konsumen dalam satu populasi data terhadap sebuah tanggal acuan tunggal (*reference date*).

Skala pengukuran dimensi ini menggunakan satuan waktu (hari), di mana secara metodologis, nilai yang lebih kecil mengindikasikan bahwa konsumen baru saja melakukan interaksi atau transaksi dengan platform. Kondisi ini dipandang lebih baik karena konsumen yang baru saja bertransaksi memiliki tingkat keterpautan (*engagement*) yang masih tinggi, sehingga probabilitas untuk melakukan pembelian kembali (*repurchase*) jauh lebih besar dibandingkan konsumen yang memiliki nilai *Recency* tinggi atau sudah lama tidak aktif.

2.3.2 Dimensi *Frequency* (F)

Dimensi *Frequency* menunjukkan seberapa sering konsumen melakukan transaksi dalam suatu periode tertentu. Nilai F mencerminkan tingkat keterlibatan dan loyalitas konsumen . Semakin tinggi nilai F, semakin kuat hubungan konsumen dengan *platform e-commerce* dan semakin besar potensi untuk dijadikan target program loyalitas [8].

$$F_i = N_{trans(i)} \quad \text{.....(2)}$$

Dimana :

F_i = nilai *Frequency* konsumen ke-i

$N_{trans(i)}$ = jumlah total transaksi konsumen ke-i dalam periode pengamatan

Skala pengukuran pada dimensi *Frequency* menggunakan rasio jumlah transaksi yang dilakukan konsumen selama periode observasi tertentu. Penentuan kategori tinggi atau rendahnya frekuensi konsumen dilakukan dengan membandingkan intensitas transaksi individu terhadap distribusi frekuensi rata-rata seluruh konsumen . Semakin besar nilai *Frequency* yang dimiliki seorang konsumen , maka semakin kuat loyalitas dan keterikatan konsumen tersebut terhadap bisnis. Konsumen dengan nilai frekuensi yang tinggi menunjukkan pola perilaku belanja yang rutin, yang menjadikannya aset penting bagi perusahaan untuk dipertahankan melalui program loyalitas.

2.3.3 Dimensi *Monetary* (M)

Dimensi *Monetary* mengukur total nilai uang yang dikeluarkan oleh konsumen selama periode pengamatan. Nilai M menggambarkan kontribusi finansial konsumen terhadap pendapatan perusahaan. Konsumen dengan nilai M tinggi dikategorikan sebagai *high-value customers* yang memberikan keuntungan signifikan dan membutuhkan strategi retensi khusus [8].

$$M_i = \sum_{j=1}^{n_i} P_{ij} \quad \text{.....(3)}$$

Dimana :

M_i = nilai *Monetary* konsumen ke-i

P_{ij} = nilai transaksi ke-j dari konsumen ke-i

n_i = jumlah transaksi konsumen ke-i

Dimensi *Monetary* diukur menggunakan skala rasio nilai mata uang yang mencerminkan akumulasi kontribusi finansial konsumen . Status tinggi atau rendahnya nilai *Monetary* ditentukan berdasarkan posisi nilai belanja individu terhadap sebaran statistik data, seperti nilai rata-rata (*mean*) atau nilai tengah (*median*) dari total belanja seluruh konsumen . Konsumen dikategorikan sebagai *high-value customers* apabila nilai akumulasi transaksinya berada secara signifikan di atas rata-rata populasi. Melalui dimensi ini, perusahaan dapat membedakan konsumen yang memberikan keuntungan besar secara finansial dibandingkan dengan konsumen yang hanya melakukan transaksi dalam nominal kecil.

Salah satu kelebihan utama dari model RFM adalah kemampuannya untuk mengidentifikasi konsumen loyal dengan lebih akurat dibandingkan dengan metode lain [22]. Dengan menganalisis ketiga dimensi ini, perusahaan dapat dengan cepat mengelompokkan konsumen ke dalam kategori yang berbeda, seperti konsumen setia, konsumen yang berisiko churn, dan konsumen baru. Sebuah penelitian menunjukkan bahwa konsumen yang memiliki nilai RFM tinggi cenderung lebih loyal dan lebih mungkin untuk melakukan pembelian ulang, sehingga memungkinkan perusahaan untuk merancang strategi pemasaran yang lebih terfokus dan efisien [10].

Namun, meskipun RFM memiliki banyak keunggulan, ada beberapa kelemahan yang perlu diperhatikan.

1. Statis: Model ini memberikan gambaran yang statis dan tidak secara alami mampu menangkap dinamika perubahan perilaku konsumen seiring waktu [16].
2. Tidak Holistik: RFM tidak mempertimbangkan faktor eksternal atau *Customer Lifetime Value* (CLV), sehingga tidak memberikan gambaran lengkap tentang potensi nilai jangka panjang konsumen [16] [23].

Keterbatasan ini menjadi dasar perlunya integrasi RFM dengan algoritma *clustering*, yang terbukti dapat meningkatkan akurasi segmentasi konsumen dengan mengenali pola tersembunyi yang kompleks dalam data [9].

2.4 Algoritma *K-Means*

Dalam analisis data modern, khususnya pada bidang pemasaran digital dan *e-commerce*, algoritma *K-Means* menjadi salah satu metode *clustering* yang paling banyak

digunakan karena kesederhanaan, efisiensi, dan kemampuannya dalam menangani data berskala besar [9]. Tujuan utama dari algoritma ini adalah mengelompokkan data ke dalam sejumlah kelompok (kluster) berdasarkan kesamaan karakteristik antar data, sehingga data dalam satu kluster memiliki tingkat kemiripan yang tinggi (*intra-cluster similarity*), sedangkan antar kluster memiliki perbedaan yang signifikan (inter-cluster dissimilarity) [9]. Dalam konteks penelitian ini, algoritma *K-Means* digunakan untuk mengelompokkan konsumen *e-commerce* berdasarkan hasil analisis *Recency*, *Frequency*, dan *Monetary* (RFM), agar diperoleh gambaran yang lebih jelas mengenai pola perilaku konsumen dan nilai ekonomi masing-masing segmen.

K-Means bekerja dengan prinsip minimisasi jarak antara titik data dan pusat kelompoknya (centroid) [24]. Secara matematis, algoritma ini berusaha meminimalkan fungsi objektif yang disebut *Within-Cluster Sum of Squares* (WCSS), yang dihitung berdasarkan jarak kuadrat antara setiap titik data dan centroid klusternya [24]. Fungsi objektif tersebut dinyatakan sebagai:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \quad \text{.....(4)}$$

Dimana :

J = total variasi dalam seluruh cluster (fungsi objektif yang diminimalkan)

k = jumlah cluster yang ditentukan

x_j = data ke-j

C_i = himpunan data dalam cluster ke-i

μ_i = centroid (titik pusat) dari cluster ke-i

$\left\| x_j - \mu_i \right\|^2$ = jarak kuadrat antara data x_j dan centroid μ_i

Selain memiliki keunggulan dalam efisiensi dan kemudahan implementasi, *K-Means* juga memiliki beberapa keterbatasan. Salah satunya adalah ketergantungan terhadap nilai awal centroid dan jumlah kluster yang ditentukan [18]. Pemilihan nilai k yang kurang tepat dapat menyebabkan hasil pengelompokan menjadi tidak optimal. Selain itu, algoritma ini sensitif terhadap data yang mengandung outlier karena perhitungan jarak Euclidean dapat memperbesar pengaruh data ekstrem [25].

2.5 Algoritma DBSCAN

Algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) merupakan salah satu metode *clustering* berbasis kepadatan (*density-based clustering*) yang digunakan untuk mengelompokkan data berdasarkan tingkat kerapatan antar titik dalam

ruang fitur [20]. Berbeda dengan algoritma *K-Means* yang memerlukan penentuan jumlah kluster di awal, DBSCAN mampu menentukan jumlah kluster secara otomatis berdasarkan distribusi dan kedekatan antar data. Pendekatan ini menjadikan DBSCAN lebih fleksibel dan efektif, terutama untuk dataset yang memiliki bentuk kluster tidak beraturan atau ukuran kluster yang berbeda-beda [26]. Selain itu, DBSCAN memiliki kemampuan untuk mendeteksi outlier atau data anomali secara alami, sehingga cocok digunakan dalam analisis data transaksi konsumen yang kompleks dan tidak selalu terdistribusi secara seragam.

Dalam prinsip kerjanya, DBSCAN membedakan setiap titik data ke dalam tiga kategori utama berdasarkan parameter radius (epsilon, disimbolkan sebagai ϵ) dan jumlah minimum tetangga yang diperlukan (MinPts) [20]:

1. *Core Point*, yaitu titik yang memiliki setidaknya MinPts tetangga dalam radius ϵ. Titik ini dianggap sebagai pusat dari suatu kluster karena memiliki kepadatan tinggi.
2. *Border Point*, yaitu titik yang berada dalam radius ϵ dari sebuah *Core Point*, namun tidak memiliki cukup tetangga untuk menjadi *Core Point*. Titik ini berada di tepi kluster dan masih dianggap sebagai bagian dari kluster tersebut.
3. *Noise Point* (Outlier), yaitu titik yang tidak termasuk dalam radius ϵ dari *Core Point* mana pun. Titik ini dianggap sebagai data yang terisolasi atau tidak termasuk dalam kluster mana pun.

Kelebihan utama DBSCAN terletak pada kemampuannya menangkap struktur alami data tanpa asumsi bentuk kluster tertentu [20]. Hal ini berbeda dengan *K-Means* yang hanya efektif untuk kluster berbentuk bulat dan ukuran relatif seragam. Dalam analisis segmentasi konsumen *e-commerce*, kondisi data sering kali bersifat tidak linier dan memiliki variasi nilai yang signifikan antar konsumen , sehingga DBSCAN menjadi alternatif yang kuat untuk menggambarkan distribusi perilaku konsumen yang lebih realistis [20].

Namun demikian, DBSCAN juga memiliki keterbatasan, yaitu sensitivitas terhadap pemilihan nilai parameter ϵ dan *MinPts* [20]. Nilai parameter yang terlalu kecil dapat menyebabkan banyak data dianggap sebagai *noise*, sementara nilai yang terlalu besar dapat membuat kluster menjadi tumpang tindih [20]. Dengan menerapkan DBSCAN pada data hasil analisis RFM, penelitian ini bertujuan untuk mengeksplorasi pola segmentasi konsumen yang tidak dapat diidentifikasi dengan algoritma berbasis centroid seperti *K-Means*. Hasil pengelompokan menggunakan DBSCAN kemudian akan dibandingkan dengan hasil dari algoritma lainnya seperti *Gaussian Mixture Model* (GMM) dan *Mean-Shift*, sehingga dapat diketahui algoritma mana yang memberikan hasil segmentasi paling akurat, stabil, dan sesuai dengan karakteristik data *e-commerce* yang dianalisis.

2.6 Algoritma *Gaussian Mixture Model* (GMM)

Algoritma *Gaussian Mixture Model* (GMM) merupakan salah satu metode *clustering* berbasis probabilitas yang berasumsi bahwa sekumpulan data berasal dari kombinasi beberapa distribusi normal Gaussian [21]. GMM menggunakan pendekatan soft *clustering*, yaitu setiap data memiliki probabilitas keanggotaan pada lebih dari satu kluster. Pendekatan ini menjadikan GMM lebih fleksibel dalam menangani data dengan struktur kompleks, tumpang tindih, atau memiliki bentuk kluster yang tidak bulat sempurna [21].

Secara matematis, GMM memodelkan distribusi data x sebagai kombinasi dari G distribusi Gaussian yang dihitung melalui algoritma *Expectation-Maximization* (EM), yang dinyatakan dengan persamaan:

$p(x) = \sum_{G=1}^{G} \pi_G \, \mathcal{N}\,(x|\mu_G, \sum_G)$……...…(5)

Dimana :

G = jumlah komponen Gaussian (cluster)

π_G = bobot komponen k, $\sum_{k=1}^{K} \pi_k = 1$

$N\,(x|\mu_G, \sum_G)$ = distribusi Gaussian dengan rata-rata μ_G dan kovarians $\sum_G$

x = data konsumen , misalnya vektor RFM [R,F,M]

Kelebihan utama GMM terletak pada kemampuannya menangani data yang heterogen dan menghasilkan pembagian kluster yang lebih halus karena bersifat probabilistik [27]. Dengan pendekatan *soft clustering*, konsumen yang memiliki karakteristik di antara dua segmen dapat tetap terwakili secara proporsional berdasarkan tingkat probabilitasnya. Namun demikian, GMM juga memiliki keterbatasan, seperti kebutuhan untuk menentukan jumlah kluster (K) di awal serta sensitif terhadap inisialisasi parameter awal [28]. Oleh karena itu, pemilihan nilai K pada penelitian ini ditentukan berdasarkan hasil evaluasi menggunakan metrik seperti *Bayesian Information Criterion* (BIC) dan *Silhouette Score*, agar diperoleh jumlah kluster yang paling optimal dan stabil [27].

2.7 Algoritma *Mean-Shift*

Algoritma *Mean-Shift* merupakan salah satu metode *clustering* berbasis estimasi kepadatan (*density estimation*) yang tidak memerlukan penentuan jumlah kluster (k) di awal proses [29]. *Mean-Shift* bekerja dengan cara mengidentifikasi area dengan kepadatan data tinggi (*density peaks*) dan secara iteratif menggeser titik data menuju pusat kepadatan tertinggi [29]. Pendekatan ini memungkinkan pembentukan kluster secara alami berdasarkan

distribusi data, sehingga hasil pengelompokan yang diperoleh lebih adaptif terhadap struktur data sebenarnya [29].

Algoritma *Mean-Shift* didasarkan pada konsep *Kernel Density Estimation* (KDE), yang digunakan untuk memperkirakan fungsi kepadatan data dalam ruang multidimensi [29]. *Mean-Shift* menghitung vektor perpindahan (*shift vector*) yang menunjukkan arah pergerakan setiap titik menuju pusat kepadatan maksimum. Persamaan vektor perpindahan dinyatakan sebagai berikut [29]:

$$m(x) = \frac{\sum_{i-1}^{n} K(x_i - x) x_i}{\sum_{i-1}^{n} K(x_i - x)} - x \quad \text{.....(6)}$$

Dimana :

x = posisi titik data saat ini (misal vektor RFM [R,F,M])

x_i = titik data tetangga dari x

K(·) = fungsi kernel, biasanya Gaussian, yang memberi bobot lebih tinggi pada titik dekat pusat

m(x) = vektor perpindahan yang menunjukkan arah pergerakan titik menuju pusat kepadatan

Keunggulan utama algoritma *Mean-Shift* terletak pada kemampuannya dalam mengungkap struktur kluster yang kompleks dan tidak simetris, serta kemampuannya menangkap variasi perilaku konsumen tanpa memerlukan asumsi distribusi awal [25]. Hal ini sangat sesuai dengan karakteristik data *e-commerce* yang umumnya dinamis, tidak linier, dan memiliki perbedaan intensitas antar konsumen . Selain itu, *Mean-Shift* juga mampu mengatasi permasalahan yang sering muncul pada algoritma lain seperti *K-Means* dan GMM, di mana hasil kluster cenderung dipengaruhi oleh inisialisasi centroid dan asumsi bentuk distribusi yang terbatas [21].

Meskipun memiliki banyak keunggulan, *Mean-Shift* juga memiliki keterbatasan, seperti kebutuhan komputasi yang relatif tinggi terutama pada dataset berukuran besar, serta sensitivitas terhadap pemilihan parameter bandwidth pada fungsi kernel [29]. Nilai bandwidth yang terlalu kecil dapat menghasilkan banyak kluster kecil, sedangkan nilai yang terlalu besar dapat menyebabkan beberapa kluster yang berbeda bergabung menjadi satu [29].

2.8 *Silhouette Score*

Silhouette Score merupakan salah satu metrik evaluasi yang digunakan untuk mengukur kualitas hasil *clustering* dengan memperhatikan dua aspek utama, yaitu kedekatan

data dalam satu kluster (*intra-cluster distance*) dan jarak antar kluster (*inter-cluster distance*) [28]. Metrik ini membantu menilai seberapa baik setiap titik data ditempatkan dalam kluster yang sesuai, serta seberapa jelas pemisahan antar kluster yang terbentuk. Nilai *Silhouette Score* berada pada rentang antara −1 hingga +1, di mana nilai yang lebih tinggi menunjukkan hasil pengelompokan yang lebih baik dan lebih terpisah secara alami [28]. Secara umum, interpretasi nilai *Silhouette Score* dapat dijelaskan sebagai berikut:

1. Nilai mendekati +1 menunjukkan bahwa data tersebut sangat sesuai dengan klusternya sendiri dan jauh dari kluster lainnya [30].
2. Nilai mendekati 0 menunjukkan bahwa data berada di area perbatasan antara dua kluster [30].
3. Nilai negatif (<0) menunjukkan bahwa data mungkin ditempatkan pada kluster yang tidak tepat karena lebih dekat dengan kluster lain dibandingkan klusternya sendiri [30].

Secara matematis, *Silhouette Score* untuk setiap titik data i dihitung menggunakan persamaan berikut [31]:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad \text{.....(7)}$$

Dimana :

a(i) = rata-rata jarak antara titik i dengan semua titik lain dalam cluster yang sama (*intra-cluster distance*).

b(i) = rata-rata jarak antara titik i dengan semua titik di cluster terdekat berikutnya (*inter-cluster distance*).

s(i) = skor S*ilhouette* untuk titik data i.

2.9 *Davies-Bouldin Index* (DBI)

Davies-Bouldin Index (DBI) merupakan salah satu metrik evaluasi yang digunakan untuk menilai kualitas hasil *clustering* dengan mempertimbangkan tingkat keseragaman antar data dalam satu kluster (*intra-cluster similarity*) serta jarak pemisahan antar kluster (*inter-cluster separation*) [32]. Nilai DBI menunjukkan seberapa baik struktur kluster yang terbentuk. Semakin kecil nilai DBI, semakin baik kualitas pengelompokan, karena hal tersebut menunjukkan bahwa setiap kluster bersifat kompak dan memiliki jarak yang jelas dengan kluster lainnya.

$$DBI = \frac{1}{k} \sum_{i=j}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{M_{ij}} \right) \quad \text{.....(8)}$$

Dimana :

K = jumlah cluster.

s_i = ukuran penyebaran (scatter) cluster ke-i

μ_i = centroid cluster ke-i.

M_{ij} = jarak antara centroid cluster i dan j:

Nilai DBI yang rendah akan menunjukkan bahwa konsumen dengan karakteristik transaksi yang mirip telah dikelompokkan dalam kluster yang sama, sementara konsumen dengan perilaku berbeda ditempatkan pada kluster yang terpisah dengan jelas [30].

2.10 *Calinski-Harabasz Index* (CHI)

Calinski–Harabasz Index (CHI) merupakan metrik evaluasi yang digunakan untuk menilai kualitas hasil *clustering* dengan membandingkan rasio antara variasi antar kluster (*between-cluster dispersion*) dan variasi di dalam kluster (*within-cluster dispersion*) [30]. Nilai CHI menggambarkan seberapa baik data dikelompokkan ke dalam kluster yang kompak dan saling terpisah. dalam penggunaan metrik *Calinski-Harabasz Index*, tidak terdapat batasan nilai absolut atau ambang batas (*threshold*) angka tertentu untuk dikategorikan sebagai nilai tinggi secara universal. Hal ini dikarenakan besaran nilai CHI sangat dipengaruhi oleh jumlah total observasi (N) dan jumlah kluster (k) yang digunakan dalam pemodelan. Oleh karena itu, penilaian terhadap metrik ini dilakukan secara relatif dengan membandingkan nilai CHI antar algoritma yang berbeda atau antar variasi jumlah kluster pada dataset yang sama.

Semakin tinggi nilai CHI, semakin baik kualitas pengelompokan, karena menunjukkan bahwa kluster yang terbentuk memiliki jarak antar pusat kluster yang besar serta penyebaran data dalam kluster yang relatif kecil. Secara matematis, *Calinski–Harabasz Index* dirumuskan sebagai berikut [33]:

$$CHI = \frac{Tr\,(B_k)}{Tr\,(W_k)}\ x\ \frac{N-k}{k-1} \quad \text{.....(9)}$$

Dimana :

N = total jumlah data.

k = jumlah cluster.

B_k = *between-cluster dispersion matrix* (variasi antar cluster)

W_k = *within-cluster dispersion matrix* (variasi dalam cluster)

Dari rumus tersebut, nilai CHI akan meningkat ketika data dalam satu kluster semakin homogen (nilai W_k kecil) dan jarak antar kluster semakin jauh (nilai B_k besar).

Dengan demikian, nilai CHI yang tinggi menekankan struktur kluster yang kompak dan terpisah dengan jelas, sedangkan nilai CHI yang rendah menunjukkan kluster yang saling tumpang tindih atau tidak terdefinisi dengan baik [34].

2.11 Penelitian Terdahulu

Dalam penyusunan tugas akhir ini, penelitian terdahulu dilakukan tentang penerapan analisis RFM (*Recency*, *Frequency*, and *Monetary*) serta berbagai algoritma *clustering*, seperti *K-Means*, DBSCAN, *Gaussian Mixture Model* (GMM), dan *Mean-Shift*. Tujuan dari penelitian ini adalah untuk mendapatkan pemahaman tentang metode, model, dan hasil dari peneliti sebelumnya sehingga dapat digunakan sebagai dasar dan perbandingan dalam pengembangan model analisis pada dasar RFM. Studi sebelumnya yang ditinjau menunjukkan bagaimana metode data mining dan *clustering* digunakan untuk mengklasifikasikan konsumen berdasarkan perilaku transaksi, khususnya dalam hal *e-commerce*. Literatur ini membantu mengidentifikasi kelebihan dan kekurangan masing-masing model, serta bagaimana masing-masing model berkontribusi pada peningkatan akurasi dan efektivitas strategi pemasaran berbasis data.

Tabel 2. 1 Literatur Review

No	Sumber	Judul Penelitian	Model	Metriks	Kontribusi	Area Aplikasi	Hasil	Kekurangan (Riset GAP)	Relevansi dengan Penelitian
1	[9]	Customer Segmentation through RFM Analysis and *K-Means Clustering*: Leveraging Data-Driven Insights for Effective Marketing Strategy	RFM + *K-Means*	*Silhouette, DBI*	Strategi pemasaran berbasis segmen	Retail / *E-commerce*	Segmentasi valid, segmen Platidbl–Gold teridentifikasi	Penelitian hanya menggunakan satu algoritma (*K-Means*) tanpa perbandingan	Menjadi baseline penting untuk segmentasi konsumen , studi yang akan dilakukan menggunakan multi-algorithm comparison (*K-Means*, DBSCAN, GMM, *Mean-Shift*) sehingga menghasilkan segmentasi lebih komprehensif pada data *e-commerce*.
2	[10]	Customer Relationship Man*Age*ment System for Retail Stores Using Unsupervised *Clustering*	RFM + *Clustering*	Validasi internal	Implementasi CRM berbasis segmentasi	Retail	Sistem CRM berhasil diterapkan	Tidak membandingkan algoritma dan tidak menggunakan metrik evaluasi yang beragam.	Masih relevan karena memperkuat pentingnya RFM, studi yang akan dilakukan melengkapi kekurangannya

		Algorithms with RFM Modeling for Customer Segmentation							dengan evaluasi metodologis, memakai beberapa algoritma dan metrik (Silhouette, DBI, CHI).
3	[11]	Smart Next-Generation Revenue Growth: A Methodology for Partitioning Customers Utilizing the *K-Means* Algorithm and RFM Model	RFM + *K-Means*	Validasi internal	Identifikasi konsumen risiko churn	Bisnis umum	Pemisahan konsumen bernilai tinggi dan risiko churn	Hanya menggunakan *K-Means*, belum menguji stabilitas metode lain pada variasi data.	Studi yang akan dilakukan mengisi gap dengan menguji banyak algoritma untuk menentukan model paling stabil dan akurat dalam domain *e-commerce*.
4	[8]	A multi layer *Recency Frequency Monetary* method for customer priority segmentation in online transaction	Multi-layer RFM	Akurasi	Pendalaman RFM granular	Online Transaction	Segmentasi lebih detail	Penelitian fokus pada modifikasi RFM, bukan perbandingan model *clustering* lain.	Studi yang akan dilakukan tidak memodifikasi RFM, namun memperdalam aspek pemodelan dengan perbandingan berbagai algoritma untuk

									mengukur performa *clustering* pada transaksi *e-commerce*.
5	[26]	Comparison of the DBSCAN and *K-MEANS* Algorithms in Segmenting Customers Using Public Transportation of Transjakarta Using the RFM Method	RFM + *K-Means* + DBSCAN	Silhouette, DBI	Perbandingan dua algoritma	Transportasi	Evaluasi *K-Means* vs DBSCAN	Hanya membandingkan dua algoritma dan tidak pada domain *e-commerce* sehingga generalisasi terbatas.	Studi yang akan dilakukan memperluas cakupan dengan menambahkan GMM & *Mean-Shift*, serta diaplikasikan pada data *e-commerce* sehingga meningkatkan kontribusi praktis.
6	[16]	A review of data mining methods in RFM-based customer segmentation	RFM + 7 Metode	Silhouette, DBI	Review menyeluruh	Online	Menunjukkan dominasi *K-Means*	Penelitian ini tidak melakukan uji coba algoritma apapun, hanya menyatakan bahwa studi sebelumnya banyak memakai *K-Means* dan	Studi yang akan dilakukan yakni menguji algoritma yang jarang digunakan seperti GMM dan *Mean-Shift* pada data *e-commerce*.

								belum mencoba metode lain.	
7	[33]	*Clustering* Analysis of PLN XYZ Customers Using *K-Means* and *Gaussian Mixture Model* (GMM) for Marketing Strategy	*K-Means* + GMM	Validasi internal	Perbandingan centroid vs probabilistik	Public Utility	Segmentasi konsumen sukses	Hanya membandingkan dua algoritma dan tidak pada domain *e-commerce* dan metrik evaluasi terbatas.	Studi yang akan dilakukan menambah kedalaman dengan membandingkan lebih banyak algoritma (*K-Means*, DBSCAN, GMM, *Mean-Shift*) dan memakai multi-metrik pada data *e-commerce* yang lebih kompleks.
8	[34]	*K-Means Clustering* Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data	*K-Means* + SAPK	CHI	Pengembangan *K-Means*	General	Segmentasi efisien	Tidak membandingkan dengan algoritma non-*K-Means* dan hanya memakai satu metrik evaluasi.	Studi yang akan dilakukan dengan multi-metrik evaluasi (Silhouette, DBI, CHI) dan membandingkan beragam algoritma secara objektif.

9	[29]	Faster *Mean-Shift*: GPU-accelerated *clustering* for cosine embedding-based cell segmentation and tracking	*Mean-Shift* (GPU)	Akurasi & kecepatan	Pengembangan algoritma	Biologi	Kluster adaptif tanpa parameter k	belum diketahui performanya untuk data transaksi konsumen .	Studi yang akan dilakukan berkontribusi dengan mengadaptasi *Mean-Shift* ke domain *e-commerce* sehingga memberi nilai kebaruan metodologis.
10	[35]	Optimizing Marketing Strategies with RFM Method and *K-Means Clustering*-Based AI Customer Segmentation Analysis	RFM + *K-Means*	Purity	Efisiensi segmentasi	Consumer	Purity 0.95	Menggunakan metrik Purity yang membutuhkan label ground truth, padahal data konsumen biasanya tidak memiliki label serta hanya memakai satu algoritma.	Studi yang akan dilakukan mengatasi keterbatasan ini dengan memakai unsupervised metrics dan membandingkan berbagai algoritma, sehingga lebih sesuai untuk data konsumen *e-commerce* tanpa label.

Berdasarkan analisis terhadap sepuluh penelitian terdahulu, terlihat bahwa penelitian-penelitian tersebut masih memiliki beberapa kelemahan yang cukup konsisten dan menjadi dasar penting bagi penelitian ini.

1. Minimnya perbandingan algoritma

   Sebagian besar penelitian hanya memakai satu algoritma biasanya algoritma *K-Means* atau membandingkan dua algoritma saja. Selain itu, belum banyak studi yang menguji alternatif lain seperti DBSCAN, GMM, dan *Mean-Shift* yang sebenarnya punya cara kerja berbeda dalam menangani bentuk cluster dan data noisy (data dengan pola yang berbeda-beda, outlier, dan mengganggu pembentukan cluster)

2. Evaluasi model masih terbatas

   Rata-rata penelitian terdahulu hanya menggunakan satu atau dua metrik, seperti Silhouette atau DBI, sehingga hasil penilaian kualitas cluster belum benar-benar komprehensif. Penelitian dengan multi-metrik seperti CHI masih jarang digunakan.

3. Banyak studi di luar konteks *e-commerce*

   Penelitian sebelumnya lebih banyak mengambil kasus ritel umum, transportasi, utilitas, atau biologi. Akibatnya, hasilnya kurang cocok jika langsung diterapkan pada data *e-commerce* yang jauh lebih kompleks dan berubah cepat.

4. Kurangnya eksplorasi algoritma non-parametrik dan probabilistik

   Metode seperti GMM dan *Mean-Shift* hampir tidak digunakan dalam segmentasi konsumen , padahal keduanya bisa menghasilkan cluster yang lebih fleksibel tanpa asumsi bentuk tertentu.

5. Algoritma belum banyak diuji pada data *e-commerce* sesungguhnya

   Beberapa algoritma contohnya *Mean-Shift*, banyak digunakan pada bidang lain seperti pemrosesan citra, tetapi belum diadaptasi untuk transaksi konsumen di *e-commerce*. Hal inilah yang menjadi Gap pada penelitian ini yakni untuk membuka peluang kontribusi baru.

6. RFM belum divalidasi dengan berbagai algoritma

   Walaupun RFM sering dipakai, banyak penelitian hanya memodifikasi komponennya tanpa mengecek performa *clustering* lintas metode. Akibatnya, belum diketahui algoritma mana yang paling cocok untuk data konsumen yang tidak berlabel.

Keseluruhan temuan ini menunjukkan adanya kebutuhan untuk melakukan penelitian yang tidak hanya membandingkan berbagai algoritma, tetapi juga mengevaluasinya dengan beberapa metrik sekaligus agar dapat menghasilkan model segmentasi konsumen yang lebih akurat dan sesuai dengan karakteristik data *e-commerce*.