BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terdahulu merupakan bagian yang sangat penting dalam penyusunan karya ilmiah, khususnya dalam penelitian berbasis *data mining*, karena berfungsi sebagai dasar dalam memahami perkembangan ilmu pengetahuan serta pendekatan yang telah dilakukan oleh peneliti sebelumnya[1]. Dengan mempelajari penelitian-penelitian yang relevan, peneliti dapat memperoleh gambaran mengenai metode yang digunakan, jenis data yang dianalisis, serta hasil yang telah dicapai dalam penelitian sejenis. Selain itu, kajian terhadap penelitian terdahulu juga membantu dalam mengidentifikasi kelebihan dan keterbatasan dari masing-masing penelitian, sehingga dapat menjadi bahan evaluasi untuk mengembangkan penelitian yang lebih baik. Dalam konteks penelitian ini, fokus utama kajian penelitian terdahulu adalah pada penerapan teknik *clustering* dalam mengelompokkan data, khususnya pada data yang berkaitan dengan transaksi atau aktivitas bisnis.

Lebih lanjut, penelitian terdahulu juga memiliki peran penting dalam mengidentifikasi kesenjangan penelitian (*research gap*) yang masih belum banyak dibahas atau belum diselesaikan secara optimal oleh penelitian sebelumnya. Dengan mengetahui kesenjangan tersebut, peneliti dapat menentukan kontribusi yang akan diberikan melalui penelitian yang dilakukan, baik dalam bentuk penggunaan metode yang berbeda, pengembangan model, maupun penerapan pada jenis data yang belum banyak diteliti. Dalam penelitian ini, kajian terhadap penelitian terdahulu difokuskan pada perbandingan berbagai algoritma *clustering* seperti *K-means*, *DBSCAN*, *Agglomerative Clustering*, dan *Gaussian Mixture Model* yang telah digunakan dalam berbagai studi sebelumnya. Hal ini bertujuan untuk memahami karakteristik masing-masing algoritma serta mengetahui performa dan keunggulannya dalam berbagai kondisi data.

Selain itu, penelitian terdahulu juga memberikan referensi dalam menentukan metode evaluasi yang tepat untuk mengukur kualitas hasil *clustering*. Beberapa penelitian sebelumnya umumnya menggunakan metrik evaluasi seperti *Silhouette*

Score, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index* untuk menilai tingkat kesamaan dalam satu *cluster* dan perbedaan antar *cluster*. Dengan memahami metode evaluasi yang digunakan dalam penelitian sebelumnya, peneliti dapat memilih pendekatan evaluasi yang paling sesuai dengan tujuan penelitian. Oleh karena itu, melalui kajian penelitian terdahulu yang komprehensif, diharapkan penelitian ini dapat memberikan kontribusi yang lebih baik, khususnya dalam mengelompokkan data *invoice* perusahaan agar lebih terstruktur, mudah dipahami, serta dapat digunakan sebagai dasar dalam pengambilan keputusan.

Selain itu, kajian terhadap penelitian terdahulu juga membantu dalam menentukan arah metodologi penelitian yang akan digunakan serta memastikan bahwa penelitian yang dilakukan memiliki dasar ilmiah yang kuat. Dengan mengkaji berbagai penelitian yang telah dilakukan sebelumnya, peneliti dapat memahami bagaimana penerapan metode *clustering* dalam berbagai studi, termasuk bagaimana proses pengolahan data, pemilihan parameter, serta interpretasi hasil *clustering* dilakukan[2]. Hal ini menjadi penting agar penelitian yang dilakukan tidak hanya mengulang penelitian sebelumnya, tetapi juga mampu memberikan nilai tambah, baik dari segi metode, objek penelitian, maupun hasil analisis yang dihasilkan. Dalam penelitian ini, penggunaan dataset berupa data *invoice* perusahaan menjadi salah satu pembeda utama dibandingkan penelitian sebelumnya, karena data tersebut memiliki karakteristik yang spesifik dan berkaitan langsung dengan aktivitas operasional perusahaan. Oleh karena itu, melalui analisis terhadap penelitian terdahulu, diharapkan penelitian ini mampu memberikan kontribusi yang lebih relevan dan aplikatif dalam membantu perusahaan memahami struktur data yang dimiliki serta mendukung pengambilan keputusan berbasis data.

Table 2.1 Penelitian Terdahulu

Ref	[3]
Penulis	Mim S, Logofatu D
Judul	A Cluster-based Analysis for Targeting Potential Customers in a Real-world Marketing System
Algoritma	*K-means Clustering, Affinitty Propagation, DBSCAN, Hierarchical*

Dataset	Shopping Mall Customer Dataset (Age, Annual Income, Spending Score)
Hasil Penelitian	Clustering berhasil mengidentifikasi pola perilaku pelanggan berdasarkan usia, pendapatan, dan spending score sehingga mendukung segmentasi pelanggan serta strategi pemasaran yang lebih efektif.
Ref	[4]
Penulis	X.Li, Y.Lee
Judul	*Customer Segmentation Marketing Strategy Based on Big data Analysis and Clustering Algorithm*
Algoritma	Support Vector Machine (SVM), K-Means Algorithm (KMA)
Dataset	Customer Dataset
Hasil Penelitian	Integrasi SVM dan K-Means berhasil menghasilkan segmentasi pelanggan yang akurat dengan akurasi rata-rata 97,51%, sehingga mendukung penyusunan strategi pemasaran dan pengambilan keputusan bisnis.
Ref	[5]
Penulis	Y.Li, X.Chu, D.Tian et al.
Judul	*Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm*
Algoritma	Improved K-Means, Adaptive Learning Particle Swarm Optimization (ALPSO), Improved K-Means Adaptive Learning PSO (IKM-ALPSO)
Dataset	Grape Customer Consumption Mixed Dataset dan lima UCI Datasets
Hasil Penelitian	Algoritma IKM-ALPSO menghasilkan segmentasi pelanggan yang lebih baik dibandingkan metode pembanding, sehingga terbukti efektif dan praktis untuk customer segmentation.
Ref	[6]
Penulis	A.Saxena, A.Agarwal, B.Pandey et al..
Judul	*Examination of the Criticality of Customer Segmentation Using Unsupervised Learning Methods*
Algoritma	*K-Means, Affinity Propagation, DBSCAN*
Dataset	*Shopping Mall Customer Dataset (Gender, Annual Income)*
Hasil Penelitian	Clustering berhasil mengelompokkan pelanggan berdasarkan gender dan pendapatan tahunan. Hasil penelitian menunjukkan bahwa 56% pelanggan merupakan perempuan, sehingga membantu perusahaan memahami karakteristik pelanggan untuk mendukung strategi pemasaran yang lebih efektif.

Ref	[7]
Penulis	N. Hicham and S. Karim
Judul	*Analysis of Unsupervised Machine learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering*
Algoritma	*DBSCAN, K-Means, Mini Batch K-Means, Mean Shift, Spectral Clustering (Clustering Ensemble)*
Dataset	*Customer Dataset dari masyarakat Maroko (data kuesioner melalui media sosial dan email)*
Hasil Penelitian	Pendekatan clustering ensemble menghasilkan kualitas segmentasi pelanggan yang lebih baik dibandingkan algoritma individual dengan ARI 70,14%, NMI 71,75%, Dunn Index 75,15, dan Silhouette Coefficient 72,89%.
Ref	[8]
Penulis	Y. Zheng
Judul	*Customer segmentation research in marketing through clustering algorithm analysis*
Algoritma	*DBSCAN, K-Means, Mini Batch K-Means, Mean Shift, Spectral Clustering (Clustering Ensemble)*
Dataset	*Customer Dataset dari masyarakat Maroko (data kuesioner melalui media sosial dan email)*
Hasil Penelitian	Pendekatan clustering ensemble menghasilkan kualitas segmentasi pelanggan yang lebih baik dibandingkan algoritma individual dengan ARI 70,14%, NMI 71,75%, Dunn Index 75,15, dan Silhouette Coefficient 72,89%.
Ref	[9]
Penulis	John, J. M., Shobayo, O., & Ogunleye, B.
Judul	*An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. Analytics*
Algoritma	*DBSCAN, K-Means, Mini Batch K-Means, Mean Shift, Spectral Clustering (Clustering Ensemble)*
Dataset	*UK Online Retail Dataset (UCI Machine Learning Repository, 541.909 transaksi, 8 atribut)*
Hasil Penelitian	Perbandingan beberapa algoritma clustering menunjukkan bahwa GMM menghasilkan performa terbaik dengan Silhouette Score sebesar 0,80, sehingga mampu meningkatkan segmentasi pelanggan dan mendukung pengambilan keputusan berbasis data pada industri ritel.

Ref	[10]
Penulis	Y. Liu
Judul	*Customer Segmentation in User Behavior Analysis: A Comparative Study of Clustering Algorithms*
Algoritma	*K-Means, Hierarchical Clustering, DBSCAN*
Dataset	*Mall Customer Segmentation Dataset (Age, Annual Income, Spending Score)*
Hasil Penelitian	Perbandingan algoritma clustering menunjukkan bahwa DBSCAN menghasilkan performa terbaik pada dataset dengan kepadatan data yang tidak merata karena mampu mengidentifikasi cluster dan noise secara otomatis, sehingga memberikan segmentasi pelanggan yang lebih efektif.
Ref	[11]
Penulis	K. Tabianan, S. Velu, and V. Ravi
Judul	*K-means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data,*
Algoritma	*K-Means Clustering)*
Dataset	*E-commerce Customer Behavior Dataset (Event Type, Product, Category, Purchase History)*
Hasil Penelitian	K-Means berhasil mengelompokkan pelanggan berdasarkan perilaku pembelian sehingga membantu mengidentifikasi segmen pelanggan dengan tingkat profitabilitas tinggi maupun rendah untuk mendukung strategi pemasaran yang lebih efektif..
Ref	[12]
Penulis	Ling, S. S., Too, C. W., Wong, W. Y., & Hoo, M. H.
Judul	*An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. Analytics*
Algoritma	*DBSCAN, K-Means, Mini Batch K-Means, Mean Shift, Spectral Clustering (Clustering Ensemble)*
Dataset	*UK Online Retail Dataset (UCI Machine Learning Repository, 541.909 transaksi, 8 atribut)*
Hasil Penelitian	Perbandingan beberapa algoritma clustering menunjukkan bahwa GMM menghasilkan performa terbaik dengan Silhouette Score sebesar 0,80, sehingga mampu meningkatkan segmentasi pelanggan dan mendukung pengambilan keputusan berbasis data pada industri ritel.
Ref	[13]

Penulis	A. Ullah, M. I. Mohmand, H. Hussain, S. Johar, and I. Khan
Judul	*Customer Analysis Using Machine learning-Based*
Algoritma	*Agglomerative Clustering, K-Means, Gaussian Mixture Model (GMM), DBSCAN*
Dataset	*Pakistan E-commerce Dataset berbasis RFMT (Recency, Frequency, Monetary, Time)*
Hasil Penelitian	Perbandingan beberapa algoritma clustering menghasilkan tiga cluster yang stabil berdasarkan majority voting dan berbagai metrik evaluasi (Silhouette, Calinski-Harabasz, Davies-Bouldin, dan Dunn Index), sehingga membantu meningkatkan strategi pemasaran dan hubungan pelanggan..
Ref	[14]
Penulis	G. Wang.
Judul	*Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering*
Algoritma	*K-Means, Q-Learning Differential Evolution (QLDE), Principal Component Analysis (PCA)*
Dataset	*Kaggle Customer Dataset*
Hasil Penelitian	Integrasi K-Means dengan QLDE dan PCA berhasil menghasilkan enam cluster pelanggan dengan akurasi segmentasi di atas 95%, sehingga meningkatkan efektivitas strategi pemasaran digital dan identifikasi karakteristik pelanggan.
Ref	[15]
Penulis	C. P. Ufeli, M. U. Sattar, R. Hasan, and S. Mahmood
Judul	*Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-means and Hierarchical Clustering Algorithms*
Algoritma	*K-Means, Agglomerative Clustering, Factor Analysis of Mixed Data (FAMD)*
Dataset	*Kaggle Retail Dataset (3.900 pelanggan, data numerik dan kategorikal)*
Hasil Penelitian	Integrasi FAMD dengan Agglomerative Clustering menghasilkan performa terbaik dengan Silhouette Score sebesar 0,52, sehingga mampu mengidentifikasi segmen pelanggan dan mendukung strategi pemasaran yang lebih terarah..
Ref	[16]
Penulis	B. Turkmen.

Judul	*CUSTOMER SEGMENTATION WITH MACHINE," EjSBS - The European Journal of Social & Behavioural Sciences*
Algoritma	*K-Means, Hierarchical Clustering, DBSCAN, RFM Clustering*
Dataset	*Online Retail Dataset*
Hasil Penelitian	Perbandingan beberapa metode segmentasi menunjukkan bahwa K-Means menghasilkan segmentasi pelanggan yang paling intuitif dan sesuai untuk kebutuhan bisnis, sehingga lebih efektif dalam mendukung strategi pemasaran dibandingkan metode lainnya.
Ref	[17]
Penulis	U. I. Hartanto, I. G. P. A. Buditjahjanto, and W. Yustanti
Judul	*Hybrid Clustering and Classification of At-Risk Customer Segments in Network Marketing*
Algoritma	*K-Means Clustering, Decision Tree (C4.5)*
Dataset	*Historical Transaction Dataset CV Jowon Solusindo (576 data transaksi)*
Hasil Penelitian	K-Means berhasil mengelompokkan pelanggan ke dalam tiga segmen berdasarkan karakteristik transaksi, sedangkan Decision Tree memperoleh akurasi 67% dalam menghasilkan aturan keputusan untuk mendukung strategi pemasaran yang lebih personal..
Ref	[18]
Penulis	S. P. G, V. S. K, H. B, K. R, and K. S. B.
Judul	*Customer Segmentation using Clustering Techniques: Data-Driven Approach to Enhance Marketing Strategy*
Algoritma	*K-Means, Hierarchical Clustering, DBSCAN*
Dataset	*Mall Customer Segmentation Dataset (Age, Gender, Annual Income, Spending Score)*
Hasil Penelitian	Perbandingan beberapa algoritma clustering berhasil mengidentifikasi karakteristik segmen pelanggan berdasarkan atribut demografi dan perilaku, sehingga mendukung penyusunan strategi pemasaran dan pengambilan keputusan berbasis data.
Ref	[19]
Penulis	L. S. Ling and C. T. Weiling
Judul	*An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. Analytics*
Algoritma	*K-Means, K-Means++, K-Medoids, Agglomerative Clustering, DBSCAN, Fuzzy C-Means, Mini Batch K-Means, Mean Shift, Gaussian Mixture Model (GMM)*

Dataset	*Kaggle Customer Dataset berbasis RFM (Recency, Frequency, Monetary)*
Hasil Penelitian	Perbandingan beberapa algoritma clustering menunjukkan bahwa K-Means++ menghasilkan performa segmentasi terbaik, sehingga meningkatkan prediksi Customer Lifetime Value (CLV) dan mendukung strategi pemasaran yang lebih efektif.
Ref	[20]
Penulis	J. Homepage, M. F. Fadhillah, A. Lovely, A. Suyoso, and I. Puspitasari
Judul	*MALCOM: Indonesian Journal of Machine learning and Computer Science Customer Segmentation with Clustering Algorithm Based on Recency, Frequency, and Monetary (RFM) Attributes*
Algoritma	*K-Means, Agglomerative Clustering, DBSCAN*
Dataset	*Data Riwayat Transaksi Penjualan PT SID Region Jawa Timur (118.314 transaksi, 1.570 pelanggan) berbasis RFM*
Hasil Penelitian	Perbandingan algoritma clustering menunjukkan bahwa K-Means menghasilkan performa terbaik dengan Silhouette Score 0,364, DBI 0,93, dan Calinski-Harabasz Index 1303,6, serta berhasil mengelompokkan pelanggan menjadi loyal customer, adequate customer, dan churn customer.
Ref	[21]
Penulis	A. Välimäki
Judul	*Customer Segmentation using Two-Stage Clustering*
Algoritma	*Self-Organizing Map (SOM), K-Means*
Dataset	*Dataset Nasabah Digital Bank Nordnet (5.000 pelanggan anonim; atribut demografi, riwayat transaksi, dan pola penggunaan platform)*
Hasil Penelitian	Integrasi SOM dan K-Means menghasilkan performa clustering yang lebih baik dibandingkan K-Means saja, serta mampu mengidentifikasi segmen pelanggan yang mudah diinterpretasikan berdasarkan aktivitas transaksi, penggunaan aplikasi, dan masa keanggotaan pelanggan.
Ref	[22]
Penulis	Y. N. Kholisho and Z. Amri
Judul	*E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary Model*
Algoritma	*K-Means Clustering*

Dataset	*Dataset Riwayat Transaksi Pelanggan E-commerce (3.606 pelanggan) berbasis LRFM (Length, Recency, Frequency, Monetary)*
Hasil Penelitian	Algoritma K-Means berhasil menghasilkan tiga segmen pelanggan (New Customer, Lost Customer, dan Core Customer) berdasarkan model LRFM, sehingga membantu perusahaan mengenali loyalitas pelanggan dan menyusun strategi pemasaran yang lebih efektif.

Penelitian terdahulu merupakan salah satu landasan penting dalam mengidentifikasi perkembangan penelitian terkait customer segmentation serta menentukan posisi penelitian yang dilakukan. Berdasarkan hasil studi literatur terhadap 20 penelitian terdahulu, diketahui bahwa penerapan metode clustering telah banyak digunakan untuk mengelompokkan pelanggan berdasarkan karakteristik transaksi, perilaku pembelian, maupun atribut demografis. Berbagai algoritma clustering seperti K-Means, Agglomerative Clustering, DBSCAN, Gaussian Mixture Model (GMM), Hierarchical Clustering, hingga berbagai metode hybrid dan ensemble telah diterapkan pada beragam jenis dataset, seperti data transaksi perusahaan, data e-commerce, data retail, data CRM, maupun data perilaku pelanggan. Secara umum, penelitian-penelitian tersebut menunjukkan bahwa customer segmentation berbasis clustering mampu membantu perusahaan dalam memahami karakteristik pelanggan, mendukung penyusunan strategi pemasaran, meningkatkan customer relationship management, serta mendukung proses pengambilan keputusan berbasis data. Ringkasan penelitian terdahulu yang menjadi acuan dalam penelitian ini disajikan pada Tabel 2.1.

Berdasarkan penelitian terdahulu yang telah dikaji, dapat diketahui bahwa penerapan teknik *clustering* dalam segmentasi pelanggan telah banyak dilakukan dengan berbagai pendekatan dan algoritma yang terus berkembang. Penelitian oleh Mim S dan Logofatu D menunjukkan bahwa algoritma *K-means* mampu mengelompokkan data pelanggan secara efektif berdasarkan tingkat kemiripan, sehingga memudahkan perusahaan dalam memahami pola perilaku pelanggan dan mendukung pengambilan keputusan strategis. Selanjutnya, penelitian oleh X. Li

dan Y. Lee mengembangkan pendekatan yang lebih kompleks dengan mengintegrasikan *Support vector machine* (SVM) dan *clustering*, di mana hasilnya menunjukkan peningkatan performa segmentasi dengan tingkat error yang rendah, recall yang tinggi, serta prediksi keuntungan yang mendekati nilai aktual, sehingga membuktikan bahwa kombinasi metode supervised dan unsupervised learning dapat meningkatkan kualitas segmentasi pelanggan secara signifikan. Pengembangan metode juga dilakukan melalui pendekatan optimasi, seperti pada penelitian yang menggabungkan *K-means* dengan *Adaptive Learning Particle Swarm Optimization* (ALPSO), yang mampu mengatasi kelemahan *K-means* dalam pemilihan centroid awal dan menghasilkan performa *clustering* yang lebih optimal serta efektif dalam menangani data campuran.

Pendekatan lain kemudian dikembangkan melalui perbandingan berbagai algoritma *clustering* untuk memahami karakteristik masing-masing metode. Penelitian oleh Saxena et al. membandingkan *K-means*, *Affinity Propagation*, dan *DBSCAN*, yang menunjukkan bahwa setiap algoritma memiliki keunggulan berbeda tergantung pada distribusi data serta mampu memberikan insight tambahan terkait perilaku pelanggan berdasarkan faktor demografis. Selanjutnya, pendekatan *clustering ensemble* yang dikembangkan oleh Hicham dan Karim dengan menggabungkan beberapa algoritma seperti *DBSCAN*, *K-means*, *Mini Batch K-means*, dan *Mean Shift* yang diintegrasikan menggunakan *Spectral Clustering* terbukti mampu meningkatkan kualitas *clustering* secara signifikan dengan nilai evaluasi yang tinggi. Hal ini menunjukkan bahwa kombinasi beberapa algoritma dapat menghasilkan segmentasi yang lebih stabil dan akurat dibandingkan metode tunggal.

Selain itu, pengembangan algoritma juga dilakukan melalui pendekatan optimasi yang lebih lanjut. Penelitian oleh Zheng menggunakan *Improved Lion Swarm Optimization* (ILSO) pada *K-means* yang mampu meningkatkan akurasi *clustering* hingga di atas 90% serta menghasilkan segmentasi pelanggan yang lebih jelas. Kemudian, penelitian oleh Kucharska E mengembangkan algoritma *K-means* dengan pendekatan penentuan jumlah *cluster* secara otomatis serta penggabungan

dengan *hierarchical agglomeration* dan metode AHP, yang terbukti meningkatkan efisiensi dan akurasi *clustering* serta menghasilkan segmentasi yang lebih relevan dengan kebutuhan bisnis. Hal ini menunjukkan bahwa modifikasi algoritma dapat memberikan peningkatan performa yang signifikan dalam proses *clustering*.

Selanjutnya, penelitian oleh Liu Y menunjukkan bahwa *DBSCAN* memiliki performa terbaik pada dataset dengan kepadatan yang tidak merata dibandingkan *K-means* dan Hierarchical *Clustering*, sehingga menegaskan bahwa pemilihan algoritma sangat bergantung pada karakteristik data. Penelitian oleh Tabianan et al. juga menunjukkan bahwa *K-means* efektif dalam segmentasi pelanggan berbasis perilaku pada *e-commerce*, di mana hasilnya mampu membantu perusahaan dalam mengidentifikasi pelanggan potensial serta meningkatkan strategi pemasaran dan profit. Penelitian lain oleh Hicham dan Karim kembali menegaskan bahwa pendekatan *clustering ensemble* mampu menghasilkan segmentasi yang lebih stabil dan berkualitas tinggi dibandingkan metode tunggal.

Perkembangan penelitian selanjutnya menunjukkan adanya integrasi metode yang lebih kompleks dalam *clustering*. Penelitian oleh Ullah et al. mengusulkan model segmentasi berbasis RFMT dengan mengombinasikan beberapa algoritma *clustering* serta menggunakan berbagai metode evaluasi untuk menghasilkan *cluster* yang stabil melalui teknik *majority voting.* Selanjutnya, penelitian oleh Wang mengintegrasikan *K-means* dengan Differential Evolution berbasis Reinforcement Learning dan PCA, yang mampu meningkatkan performa *clustering* secara signifikan serta menghasilkan akurasi klasifikasi di atas 95%, sehingga menunjukkan bahwa integrasi metode optimasi dan reduksi dimensi dapat meningkatkan kemampuan model dalam mengenali karakteristik pelanggan.

Pendekatan lain juga dikembangkan untuk menangani kompleksitas data yang lebih tinggi. Penelitian oleh Ufeli et al. memperkenalkan metode FAMD untuk menangani data campuran yang dikombinasikan dengan *K-means* dan *Agglomerative Clustering*, di mana hasilnya menunjukkan peningkatan kualitas *clustering* dengan mempertahankan hubungan antar fitur serta menghasilkan segmentasi pelanggan yang lebih jelas. Selain itu, penelitian oleh Türkmen

menunjukkan bahwa tidak ada satu algoritma yang selalu terbaik, karena setiap metode seperti *K-means*, Hierarchical, *DBSCAN*, dan RFM memiliki kelebihan dan kekurangan masing-masing, sehingga pemilihan algoritma harus mempertimbangkan kebutuhan bisnis.

Lebih lanjut, penelitian oleh Hartanto et al. mengusulkan pendekatan hybrid dengan menggabungkan *clustering* dan klasifikasi, di mana *K-means* digunakan untuk segmentasi awal dan SVM memberikan performa klasifikasi terbaik dengan akurasi di atas 92%, sehingga mampu meningkatkan ketepatan segmentasi serta mendeteksi pelanggan yang berpotensi *churn*. Penelitian oleh G S et al. juga menunjukkan bahwa penggunaan beberapa algoritma *clustering* yang dikombinasikan dengan metode evaluasi seperti *Elbow* dan *Silhouette Score* dapat membantu dalam menentukan jumlah *cluster* yang optimal serta menghasilkan insight yang lebih jelas terhadap karakteristik pelanggan.

Selanjutnya, penelitian oleh Ling dan Weiling menunjukkan bahwa *K-means*++ memberikan performa terbaik dibandingkan algoritma lain seperti *K-Medoids*, *DBSCAN*, *Fuzzy C-Means*, dan GMM dalam hal akurasi segmentasi, sehingga menegaskan pentingnya eksplorasi berbagai algoritma *clustering*. Penelitian oleh Fikri et al. menggunakan model RFM menunjukkan bahwa *K-means* memberikan performa terbaik dalam membentuk *cluster* pelanggan seperti pelanggan loyal, cukup, dan *churn*, sehingga efektif dalam segmentasi berbasis transaksi.

Terakhir, penelitian oleh Valimaki A mengusulkan metode two-stage *clustering* dengan menggabungkan Self-Organizing Maps (SOM) dan *K-means* yang mampu meningkatkan kualitas *clustering* pada data berdimensi tinggi serta menghasilkan segmentasi yang lebih mudah diinterpretasikan. Hal ini kemudian diperkuat oleh penelitian Kholisho dan Amri yang menunjukkan bahwa penggunaan *K-means* dengan model LRFM mampu menghasilkan segmentasi pelanggan yang akurat menjadi kelompok pelanggan baru, pelanggan inti, dan pelanggan yang hilang, sehingga dapat membantu perusahaan dalam memahami perilaku pelanggan dan menyusun strategi bisnis yang lebih efektif. Secara keseluruhan, penelitian-penelitian tersebut menunjukkan bahwa teknik *clustering* memiliki peran yang

sangat penting dalam segmentasi pelanggan, dengan berbagai pengembangan metode seperti optimasi, *ensemble, hybrid*, serta integrasi dengan teknik lain yang mampu meningkatkan kualitas hasil *clustering* dan mendukung pengambilan keputusan berbasis data.

Berdasarkan keseluruhan dua puluh penelitian yang telah dikaji, dapat disimpulkan bahwa algoritma *K-means* merupakan metode yang paling dominan dan paling sering digunakan dalam segmentasi pelanggan karena kesederhanaannya, kemudahan implementasi, serta kemampuannya dalam menghasilkan *cluster* yang mudah diinterpretasikan. Namun, berbagai penelitian juga menunjukkan bahwa *K-means* memiliki keterbatasan, terutama dalam penentuan centroid awal dan dalam menangani data dengan distribusi yang kompleks atau tidak merata, sehingga banyak penelitian mengembangkan pendekatan lanjutan seperti *K-means*++, *Improved K-means*, serta kombinasi dengan algoritma optimasi seperti PSO dan ILSO untuk meningkatkan performanya. Selain itu, algoritma lain seperti *DBSCAN* dan *Agglomerative Clustering* juga cukup sering digunakan, di mana *DBSCAN* terbukti lebih unggul dalam menangani data dengan kepadatan yang tidak merata, sementara *Agglomerative* memberikan hasil yang lebih baik dalam beberapa kasus data campuran. Beberapa penelitian juga menunjukkan tren penggunaan metode yang lebih kompleks seperti *clustering* ensemble, *hybrid clustering*-classification, serta integrasi dengan teknik lain seperti PCA, FAMD, dan SOM, yang secara umum mampu meningkatkan kualitas hasil *clustering* dibandingkan metode tunggal.

Dari sisi evaluasi, metrik yang paling sering digunakan adalah *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*, yang digunakan secara bersamaan untuk menilai kualitas *cluster* dari aspek kohesi dan separasi. Beberapa penelitian menunjukkan bahwa kombinasi metrik ini memberikan hasil evaluasi yang lebih komprehensif dibandingkan hanya menggunakan satu metrik saja. Selain itu, dalam beberapa studi yang lebih kompleks, digunakan pula metrik tambahan seperti ARI, NMI, dan *Dunn Index* untuk meningkatkan validitas hasil evaluasi. Secara umum, metode yang menggabungkan beberapa algoritma atau

menggunakan teknik optimasi cenderung menghasilkan nilai evaluasi yang lebih baik, seperti peningkatan akurasi di atas 90% atau performa metrik yang lebih tinggi dibandingkan metode konvensional. Oleh karena itu, dapat disimpulkan bahwa tidak terdapat satu algoritma yang selalu unggul dalam semua kondisi, melainkan efektivitas metode sangat bergantung pada karakteristik data dan tujuan analisis. Namun demikian, tren penelitian menunjukkan bahwa pendekatan kombinasi, optimasi, serta penggunaan multi-metrik evaluasi merupakan arah yang paling efektif dalam menghasilkan segmentasi pelanggan yang akurat, stabil, dan relevan untuk mendukung pengambilan keputusan bisnis.

Berdasarkan studi literatur yang telah dilakukan, sebagian besar penelitian segmentasi *customer* berfokus pada perbandingan performa algoritma *clustering* berdasarkan metrik evaluasi teknis seperti *Silhouette Score*, *Davies-Bouldin Index*, maupun *Calinski-Harabasz Index*. Namun, masih terbatas penelitian yang menghubungkan hasil evaluasi tersebut dengan interpretasi bisnis dari *cluster* yang dihasilkan. Selain itu, mayoritas penelitian menggunakan dataset *retail*, *e-commerce*, dan marketing, serta berhenti pada tahap *modeling* dan evaluasi tanpa implementasi lebih lanjut dalam bentuk sistem yang dapat digunakan oleh pengguna bisnis. Oleh karena itu, penelitian ini melakukan perbandingan *K-means*, *DBSCAN*, *Gaussian Mixture Model*, dan *Agglomerative Clustering* pada data *invoice* proyek PT XYZ serta mengevaluasi hasilnya tidak hanya berdasarkan metrik *clustering*, tetapi juga berdasarkan stabilitas *cluster* dan kemudahan interpretasi bisnis. Hasil penelitian kemudian diimplementasikan ke dalam dashboard *Customer* Segmentation berbasis *Streamlit* dan divalidasi melalui *User acceptance test* (UAT) untuk memastikan manfaat praktis bagi perusahaan.

2.2 Teori yang berkaitan

2.2.1 Sistem Distribusi dan Logistik

Sistem distribusi dan logistik merupakan bagian penting dalam aktivitas operasional perusahaan yang bergerak di bidang perdagangan, manufaktur, maupun distribusi barang[23]. Distribusi dapat diartikan sebagai proses penyaluran barang dari produsen kepada pelanggan atau konsumen akhir

melalui berbagai jalur distribusi yang telah ditentukan[24]. Sementara itu, logistik merupakan kegiatan yang berkaitan dengan perencanaan, pengelolaan, penyimpanan, serta pengiriman barang secara efektif dan efisien agar produk dapat diterima oleh pelanggan pada waktu dan lokasi yang tepat[25]. Dalam lingkungan bisnis modern, sistem distribusi dan logistik tidak hanya berfokus pada pengiriman barang, tetapi juga mencakup pengelolaan informasi, pengendalian persediaan, pengelolaan transaksi, serta pelayanan pelanggan.

Perkembangan teknologi informasi menyebabkan sistem distribusi dan logistik mengalami transformasi yang signifikan. Perusahaan kini memanfaatkan sistem digital untuk mencatat aktivitas transaksi, pengiriman, hingga pengelolaan pelanggan secara terintegrasi[26]. Aktivitas seperti pencatatan *invoice*, pengelolaan data pelanggan, data produk, serta histori transaksi menjadi bagian penting dalam operasional perusahaan distribusi. Data-data tersebut terus bertambah seiring meningkatnya aktivitas bisnis perusahaan, sehingga perusahaan memerlukan sistem pengelolaan data yang baik agar informasi yang dimiliki dapat dimanfaatkan secara optimal. Dalam praktiknya, perusahaan sering menghadapi permasalahan berupa jumlah data yang sangat besar dan belum terorganisir dengan baik, sehingga menyulitkan proses analisis maupun pengambilan Keputusan.

Selain itu, sistem distribusi yang baik juga berpengaruh terhadap efisiensi operasional perusahaan. Pengelolaan distribusi yang tidak optimal dapat menyebabkan keterlambatan pengiriman, ketidaksesuaian stok, kesalahan pencatatan transaksi, hingga menurunnya kualitas pelayanan kepada pelanggan. Oleh karena itu, perusahaan perlu memahami pola transaksi, perilaku pelanggan, serta karakteristik produk yang dimiliki. Salah satu cara yang dapat dilakukan adalah dengan memanfaatkan analisis data untuk mengelompokkan data transaksi berdasarkan kemiripan tertentu. Dengan adanya pengelompokan data tersebut, perusahaan dapat lebih mudah memahami kategori pelanggan, produk unggulan, maupun pola transaksi yang sering terjadi dalam perusahaan.

Dalam perusahaan distribusi dan logistik, data transaksi seperti *invoice* memiliki peran yang sangat penting karena mencerminkan aktivitas bisnis perusahaan secara langsung. Data *invoice* biasanya memuat informasi seperti nama pelanggan, produk yang dibeli, jumlah transaksi, waktu transaksi, serta nilai pembelian. Jika data tersebut dianalisis dengan baik, perusahaan dapat memperoleh insight yang bermanfaat untuk mendukung strategi bisnis dan pengambilan keputusan. Namun, karena jumlah data yang terus meningkat setiap waktu, proses analisis secara manual menjadi kurang efektif dan membutuhkan waktu yang lama. Oleh sebab itu, diperlukan pendekatan berbasis *data mining* dan *clustering* untuk membantu perusahaan dalam mengelompokkan dan memahami data transaksi secara lebih efisien.

Penerapan teknologi analisis data pada sistem distribusi dan logistik juga dapat membantu perusahaan dalam meningkatkan daya saing bisnis. Dengan memanfaatkan teknik analisis data, perusahaan dapat mengidentifikasi pelanggan potensial, memahami pola pembelian pelanggan, menentukan kategori produk yang paling sering digunakan, serta meningkatkan efektivitas strategi pemasaran dan distribusi. Selain itu, hasil analisis data juga dapat digunakan untuk membantu karyawan perusahaan, khususnya karyawan baru, dalam memahami karakteristik pelanggan dan produk perusahaan tanpa harus bergantung sepenuhnya pada pengalaman karyawan lama. Oleh karena itu, integrasi antara sistem distribusi dan analisis data menjadi salah satu faktor penting dalam mendukung perkembangan perusahaan di era digital saat ini.

Berdasarkan hasil kajian terhadap penelitian-penelitian terdahulu, dapat disimpulkan bahwa metode clustering merupakan pendekatan yang efektif untuk melakukan segmentasi pelanggan pada berbagai domain bisnis, seperti retail, *e-commerce*, perbankan, maupun industri jasa. Sebagian besar penelitian memanfaatkan data transaksi pelanggan dengan pendekatan *feature engineering* seperti RFM maupun LRFM, kemudian melakukan evaluasi menggunakan metrik seperti *Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Index* untuk menentukan kualitas cluster yang dihasilkan.

Selain itu, berbagai penelitian juga menunjukkan bahwa hasil segmentasi pelanggan mampu memberikan manfaat praktis dalam mendukung strategi pemasaran, customer relationship management, identifikasi pelanggan prioritas, serta pengambilan keputusan bisnis berbasis data.

Meskipun demikian, masih terdapat beberapa keterbatasan pada penelitian terdahulu. Sebagian besar penelitian hanya berfokus pada perbandingan performa algoritma berdasarkan metrik evaluasi *clustering* tanpa mengaitkannya secara mendalam dengan interpretasi bisnis dari *cluster* yang dihasilkan. Selain itu, mayoritas penelitian menggunakan dataset *retail*, *e-commerce*, atau CRM, sementara penelitian yang memanfaatkan data invoice proyek perusahaan masih relatif terbatas. Beberapa penelitian juga berhenti pada tahap evaluasi model tanpa mengimplementasikan hasil segmentasi ke dalam sistem yang dapat digunakan secara langsung oleh pengguna bisnis maupun melakukan validasi melalui *User Acceptance Test (UAT).* Oleh karena itu, penelitian ini berupaya mengisi kesenjangan tersebut dengan membandingkan algoritma *K-Means, DBSCAN, Gaussian Mixture Model, dan Agglomerative Clustering* pada data *invoice* proyek PT XYZ, kemudian mengimplementasikan hasil segmentasi ke dalam dashboard berbasis *Streamlit* serta memvalidasi kegunaannya melalui *User Acceptance Test* (UAT).

2.2.2 Data Transaksi Perusahaan

Data transaksi perusahaan merupakan kumpulan data yang dihasilkan dari berbagai aktivitas operasional bisnis yang dilakukan perusahaan dalam periode tertentu. Data transaksi biasanya mencatat seluruh aktivitas yang berkaitan dengan proses penjualan, pembelian, distribusi barang, pembayaran, maupun hubungan perusahaan dengan pelanggan[27]. Dalam perusahaan distribusi dan logistik, data transaksi memiliki peran yang sangat penting karena menjadi sumber utama informasi yang menggambarkan kondisi operasional perusahaan secara langsung. Data tersebut umumnya tersimpan dalam bentuk digital melalui sistem informasi perusahaan dan terus bertambah seiring meningkatnya aktivitas bisnis yang dilakukan setiap harinya.

Salah satu bentuk data transaksi yang paling umum digunakan dalam perusahaan adalah data *invoice*. *Invoice* merupakan dokumen transaksi yang berisi rincian aktivitas penjualan atau pembelian yang dilakukan perusahaan kepada pelanggan maupun mitra bisnis. Data *invoice* biasanya memuat berbagai informasi penting seperti nomor *invoice*, tanggal transaksi, nama pelanggan, jenis produk, jumlah barang, harga barang, total transaksi, serta informasi pembayaran. Informasi tersebut dapat digunakan perusahaan untuk memantau aktivitas penjualan, mengevaluasi performa bisnis, serta memahami pola transaksi pelanggan. Oleh karena itu, data *invoice* tidak hanya berfungsi sebagai dokumen administrasi, tetapi juga sebagai sumber data yang memiliki nilai strategis bagi perusahaan.

Seiring berkembangnya aktivitas bisnis perusahaan, jumlah data transaksi yang dimiliki perusahaan akan terus meningkat dan menjadi semakin kompleks. Data transaksi yang besar dan tidak terorganisir dengan baik dapat menimbulkan berbagai permasalahan, seperti kesulitan dalam pencarian data, proses analisis yang lambat, hingga kesalahan dalam pengambilan keputusan. Dalam banyak kasus, pemahaman terhadap data perusahaan hanya dimiliki oleh karyawan tertentu yang telah lama bekerja dan memahami pola transaksi perusahaan secara manual. Hal tersebut menyebabkan karyawan baru sering mengalami kesulitan dalam memahami karakteristik pelanggan, produk unggulan, maupun pola transaksi yang dimiliki perusahaan. Oleh karena itu, diperlukan suatu metode analisis data yang mampu membantu perusahaan dalam mengelompokkan dan menyusun data transaksi secara lebih terstruktur.

Data transaksi perusahaan juga memiliki karakteristik yang beragam karena terdiri dari berbagai jenis atribut, baik numerik maupun kategorikal. Contoh atribut numerik meliputi jumlah transaksi, harga produk, dan total pembelian, sedangkan atribut kategorikal meliputi nama pelanggan, kategori produk, dan wilayah distribusi. Keberagaman atribut tersebut menyebabkan proses analisis data menjadi lebih kompleks, terutama jika dilakukan secara manual. Oleh sebab itu, diperlukan proses pengolahan data seperti data preprocessing,

normalisasi, dan *clustering* agar data transaksi dapat dianalisis secara lebih efektif dan menghasilkan informasi yang mudah dipahami.

Selain digunakan untuk kebutuhan administrasi dan pencatatan transaksi, data transaksi perusahaan juga dapat dimanfaatkan sebagai dasar dalam pengambilan keputusan bisnis. Melalui analisis data transaksi, perusahaan dapat mengetahui pelanggan yang paling aktif melakukan transaksi, produk yang paling sering dibeli, pola pembelian pelanggan, hingga potensi pelanggan yang memberikan kontribusi terbesar terhadap perusahaan. Informasi tersebut dapat membantu perusahaan dalam menentukan strategi pemasaran, meningkatkan kualitas layanan, mengoptimalkan distribusi produk, serta meningkatkan efisiensi operasional perusahaan secara keseluruhan.

Dalam penelitian ini, data transaksi perusahaan yang digunakan berupa data *invoice* dari perusahaan distribusi yang telah disamarkan menjadi PT XYZ. Data tersebut akan dianalisis menggunakan metode *clustering* untuk mengelompokkan data berdasarkan karakteristik tertentu sehingga dapat menghasilkan informasi yang lebih terstruktur dan mudah dipahami. Dengan adanya proses *clustering*, perusahaan diharapkan dapat lebih mudah memahami pola data transaksi, mengidentifikasi pelanggan maupun produk unggulan, serta membantu proses pengambilan keputusan secara lebih efektif dan berbasis data.

2.2.3 *Invoice* Perusahaan

Invoice perusahaan merupakan dokumen transaksi yang digunakan sebagai bukti terjadinya aktivitas penjualan atau pembelian barang maupun jasa antara perusahaan dengan pelanggan atau mitra bisnis[28]. *Invoice* biasanya diterbitkan oleh perusahaan kepada pelanggan setelah proses transaksi dilakukan dan berisi rincian informasi terkait transaksi tersebut. Dalam kegiatan operasional perusahaan, *invoice* memiliki peran yang sangat penting karena tidak hanya digunakan sebagai dokumen administrasi dan penagihan, tetapi juga sebagai sumber informasi bisnis yang dapat digunakan untuk melakukan analisis terhadap aktivitas perusahaan.

Secara umum, *invoice* memuat berbagai informasi penting seperti nomor *invoice*, tanggal transaksi, nama pelanggan, nama produk, jumlah barang, harga satuan, total pembayaran, serta informasi pembayaran lainnya. Informasi tersebut menggambarkan aktivitas transaksi perusahaan secara rinci dan dapat digunakan untuk memantau performa penjualan maupun distribusi barang. Dalam perusahaan distribusi dan logistik, jumlah *invoice* yang dihasilkan setiap hari dapat sangat besar karena tingginya aktivitas transaksi yang terjadi. Oleh karena itu, pengelolaan data *invoice* menjadi salah satu aspek penting dalam menjaga efisiensi operasional perusahaan.

Selain berfungsi sebagai dokumen administrasi, *invoice* juga memiliki nilai strategis karena menyimpan data historis transaksi perusahaan. Data *invoice* dapat digunakan untuk mengetahui pola pembelian pelanggan, produk yang paling sering dibeli, pelanggan dengan tingkat transaksi tertinggi, hingga tren penjualan perusahaan dalam periode tertentu. Dengan melakukan analisis terhadap data *invoice*, perusahaan dapat memperoleh insight yang bermanfaat untuk mendukung pengambilan keputusan bisnis. Misalnya, perusahaan dapat menentukan produk unggulan, memahami karakteristik pelanggan, serta menyusun strategi pemasaran berdasarkan pola transaksi yang dimiliki pelanggan.

Namun, dalam praktiknya, banyak perusahaan masih menghadapi kendala dalam pengelolaan data *invoice*, terutama ketika jumlah data yang dimiliki sangat besar dan belum terorganisir dengan baik. Data *invoice* yang terus bertambah setiap waktu dapat menyebabkan proses pencarian dan analisis data menjadi lebih sulit jika dilakukan secara manual. Selain itu, pemahaman terhadap data transaksi perusahaan sering kali hanya dimiliki oleh karyawan tertentu yang telah lama bekerja dan memahami pola transaksi perusahaan secara langsung. Kondisi tersebut dapat menyebabkan terjadinya ketergantungan terhadap pengalaman individu dan menyulitkan karyawan baru dalam memahami data perusahaan secara menyeluruh.

Data *invoice* juga memiliki karakteristik yang kompleks karena terdiri dari berbagai atribut dengan tipe data yang berbeda, seperti data numerik dan kategorikal. Sebagai contoh, total transaksi dan jumlah barang termasuk atribut numerik, sedangkan nama pelanggan, kategori produk, dan wilayah distribusi termasuk atribut kategorikal. Kompleksitas tersebut menyebabkan perusahaan memerlukan metode pengolahan data yang tepat agar informasi yang tersimpan dalam *invoice* dapat dimanfaatkan secara optimal. Oleh karena itu, diperlukan proses analisis data seperti data preprocessing dan *clustering* untuk membantu perusahaan dalam mengelompokkan serta memahami data *invoice* secara lebih efektif.

Dalam penelitian ini, data *invoice* perusahaan digunakan sebagai sumber utama data penelitian karena dianggap mampu merepresentasikan aktivitas bisnis perusahaan secara langsung. Melalui analisis terhadap data *invoice*, perusahaan dapat memperoleh informasi terkait pola transaksi, segmentasi pelanggan, serta kategori produk yang paling dominan dalam aktivitas bisnis perusahaan. Hasil analisis tersebut diharapkan dapat membantu perusahaan dalam meningkatkan efektivitas pengelolaan data, mendukung pengambilan keputusan berbasis data, serta membantu karyawan perusahaan dalam memahami informasi bisnis secara lebih terstruktur dan mudah dipahami.

2.2.4 Segmentasi Pelanggan

Segmentasi pelanggan merupakan proses pengelompokan pelanggan ke dalam beberapa kelompok berdasarkan karakteristik, kebutuhan, perilaku, atau pola transaksi yang memiliki kesamaan tertentu[29]. Segmentasi pelanggan dilakukan agar perusahaan dapat memahami karakteristik masing-masing kelompok pelanggan secara lebih spesifik sehingga strategi bisnis yang diterapkan dapat disesuaikan dengan kebutuhan setiap kelompok pelanggan. Dalam dunia bisnis modern, segmentasi pelanggan menjadi salah satu strategi penting karena setiap pelanggan memiliki perilaku dan preferensi yang berbeda-beda dalam melakukan transaksi.

Pada dasarnya, tujuan utama segmentasi pelanggan adalah membantu perusahaan dalam mengidentifikasi pelanggan yang memberikan kontribusi terbesar terhadap perusahaan, memahami kebutuhan pelanggan, serta meningkatkan efektivitas strategi pemasaran dan pelayanan. Dengan adanya segmentasi pelanggan, perusahaan dapat menentukan pendekatan yang lebih tepat dalam memberikan layanan maupun promosi kepada pelanggan. Sebagai contoh, pelanggan dengan tingkat transaksi tinggi dapat diberikan layanan prioritas atau program loyalitas khusus, sedangkan pelanggan dengan aktivitas transaksi rendah dapat diberikan strategi promosi tertentu untuk meningkatkan minat pembelian.

Segmentasi pelanggan juga membantu perusahaan dalam memahami pola perilaku pelanggan berdasarkan data transaksi yang dimiliki perusahaan. Data transaksi seperti *invoice* dapat digunakan untuk mengetahui frekuensi pembelian pelanggan, jenis produk yang sering dibeli, jumlah transaksi, hingga nilai pembelian pelanggan dalam periode tertentu. Informasi tersebut dapat digunakan untuk mengidentifikasi kelompok pelanggan yang aktif, pelanggan potensial, maupun pelanggan yang mulai mengalami penurunan aktivitas transaksi. Dengan demikian, perusahaan dapat mengambil langkah yang lebih tepat dalam mempertahankan hubungan dengan pelanggan.

Dalam praktiknya, segmentasi pelanggan dapat dilakukan dengan berbagai pendekatan, seperti segmentasi geografis, demografis, psikografis, maupun perilaku. Namun, dalam konteks analisis data transaksi perusahaan, segmentasi perilaku menjadi salah satu pendekatan yang paling relevan karena berfokus pada aktivitas transaksi pelanggan. Segmentasi perilaku biasanya dilakukan berdasarkan pola pembelian, tingkat loyalitas pelanggan, intensitas transaksi, serta nilai transaksi yang dihasilkan pelanggan. Pendekatan ini memungkinkan perusahaan untuk memperoleh insight yang lebih mendalam terkait perilaku pelanggan dalam melakukan transaksi.

Perkembangan teknologi informasi dan meningkatnya jumlah data transaksi menyebabkan proses segmentasi pelanggan secara manual menjadi

kurang efektif. Dalam perusahaan dengan jumlah pelanggan yang besar, proses analisis secara manual membutuhkan waktu yang lama dan berpotensi menghasilkan kesalahan dalam pengelompokan data. Oleh karena itu, perusahaan mulai memanfaatkan teknik *data mining* dan machine learning untuk melakukan segmentasi pelanggan secara otomatis dan lebih akurat. Salah satu metode yang paling sering digunakan dalam segmentasi pelanggan adalah *clustering*, yaitu metode pengelompokan data berdasarkan tingkat kemiripan antar data

Penerapan segmentasi pelanggan memberikan berbagai manfaat bagi perusahaan, terutama dalam mendukung pengambilan keputusan bisnis. Dengan memahami karakteristik setiap kelompok pelanggan, perusahaan dapat meningkatkan kualitas pelayanan, menentukan strategi pemasaran yang lebih tepat sasaran, meningkatkan loyalitas pelanggan, serta mengoptimalkan penjualan produk. Selain itu, hasil segmentasi pelanggan juga dapat membantu perusahaan dalam mengidentifikasi peluang bisnis baru serta meningkatkan efisiensi operasional perusahaan secara keseluruhan.

Dalam penelitian ini, segmentasi pelanggan dilakukan menggunakan data *invoice* perusahaan yang memuat informasi transaksi pelanggan. Melalui proses segmentasi, data pelanggan akan dikelompokkan berdasarkan karakteristik transaksi tertentu sehingga perusahaan dapat memahami pola pelanggan secara lebih jelas dan terstruktur. Hasil segmentasi tersebut diharapkan dapat membantu perusahaan dalam mengenali pelanggan unggulan, memahami pola transaksi pelanggan, serta mendukung pengambilan keputusan bisnis yang lebih efektif dan berbasis data.

2.2.5 Business Intellegence

Business Intelligence (BI) merupakan suatu konsep dan teknologi yang digunakan untuk mengumpulkan, mengolah, menganalisis, serta menyajikan data menjadi informasi yang berguna dalam mendukung pengambilan keputusan bisnis[30]. Business Intelligence membantu perusahaan dalam memahami kondisi bisnis berdasarkan data yang dimiliki sehingga perusahaan

dapat mengambil keputusan secara lebih cepat, akurat, dan berbasis data. Dalam lingkungan bisnis modern yang menghasilkan data dalam jumlah besar setiap harinya, Business Intelligence menjadi salah satu komponen penting dalam meningkatkan efisiensi operasional dan daya saing perusahaan.

Secara umum, Business Intelligence mencakup berbagai proses seperti pengumpulan data, penyimpanan data, analisis data, visualisasi data, hingga penyajian laporan dalam bentuk dashboard atau grafik interaktif. Data yang digunakan dalam Business Intelligence dapat berasal dari berbagai sumber, seperti data transaksi penjualan, data pelanggan, data distribusi, data inventaris, maupun data operasional perusahaan lainnya. Seluruh data tersebut kemudian diolah untuk menghasilkan informasi yang dapat membantu perusahaan dalam memahami pola bisnis, tren penjualan, serta performa operasional perusahaan.

Dalam perusahaan distribusi dan logistik, Business Intelligence memiliki peran yang sangat penting karena perusahaan biasanya memiliki data transaksi dan data operasional dalam jumlah besar. Data tersebut terus bertambah setiap hari seiring meningkatnya aktivitas bisnis perusahaan. Jika data tersebut tidak dikelola dengan baik, perusahaan akan mengalami kesulitan dalam memahami kondisi bisnis secara menyeluruh. Oleh karena itu, Business Intelligence digunakan untuk membantu perusahaan mengubah data mentah menjadi informasi yang lebih terstruktur dan mudah dipahami. Dengan adanya Business Intelligence, perusahaan dapat mengetahui produk yang paling sering terjual, pelanggan yang paling aktif, wilayah distribusi dengan performa terbaik, hingga pola transaksi pelanggan dalam periode tertentu.

Salah satu tujuan utama Business Intelligence adalah mendukung proses data-driven decision making atau pengambilan keputusan berbasis data. Dalam pendekatan ini, keputusan bisnis tidak hanya didasarkan pada intuisi atau pengalaman individu, tetapi juga berdasarkan hasil analisis data yang objektif. Dengan memanfaatkan Business Intelligence, perusahaan dapat mengurangi risiko kesalahan dalam pengambilan keputusan serta meningkatkan efektivitas strategi bisnis yang diterapkan. Sebagai contoh, perusahaan dapat menentukan

strategi pemasaran berdasarkan pola pembelian pelanggan atau menentukan prioritas distribusi berdasarkan performa penjualan di wilayah tertentu.

Selain mendukung pengambilan keputusan, Business Intelligence juga membantu perusahaan dalam meningkatkan efisiensi operasional. Informasi yang dihasilkan dari proses analisis data dapat digunakan untuk mengidentifikasi permasalahan operasional, mengoptimalkan proses bisnis, serta meningkatkan kualitas layanan kepada pelanggan. Dalam konteks perusahaan distribusi, Business Intelligence dapat membantu perusahaan dalam memantau aktivitas distribusi barang, mengevaluasi performa pelanggan, serta mengidentifikasi produk yang memiliki tingkat penjualan tertinggi. Dengan demikian, perusahaan dapat mengambil langkah yang lebih cepat dan tepat dalam menghadapi perubahan kondisi bisnis.

Perkembangan teknologi juga menyebabkan Business Intelligence semakin terintegrasi dengan berbagai metode analisis data seperti *data mining*, machine learning, dan *clustering*. Teknik *clustering* sering digunakan dalam Business Intelligence untuk membantu perusahaan mengelompokkan data pelanggan atau data transaksi berdasarkan karakteristik tertentu. Hasil *clustering* tersebut kemudian dapat divisualisasikan dalam bentuk dashboard atau laporan interaktif sehingga lebih mudah dipahami oleh pengguna. Integrasi antara Business Intelligence dan analisis data memungkinkan perusahaan untuk memperoleh insight yang lebih mendalam terhadap data yang dimiliki.

Dalam penelitian ini, konsep Business Intelligence digunakan sebagai dasar dalam pengolahan dan pemanfaatan data transaksi perusahaan menjadi informasi yang lebih bermanfaat. Data *invoice* perusahaan yang sebelumnya hanya berupa data transaksi mentah akan dianalisis dan divisualisasikan menggunakan teknik *clustering* dan dashboard interaktif. Hasil analisis tersebut diharapkan dapat membantu perusahaan dalam memahami pola transaksi, mengenali pelanggan unggulan, serta mendukung pengambilan keputusan bisnis secara lebih efektif dan berbasis data.

2.2.6 Visualisasi Data

Visualisasi data merupakan proses penyajian data atau informasi dalam bentuk visual seperti grafik, diagram, tabel, maupun dashboard interaktif agar data lebih mudah dipahami oleh pengguna[31]. Dalam perkembangan teknologi informasi saat ini, visualisasi data menjadi salah satu komponen penting dalam proses analisis data karena mampu membantu pengguna memahami pola, tren, hubungan antar data, serta informasi penting yang sulit dipahami apabila hanya disajikan dalam bentuk angka atau teks. Dengan visualisasi yang baik, proses interpretasi data dapat dilakukan secara lebih cepat, efektif, dan intuitif.

Dalam lingkungan bisnis, visualisasi data memiliki peran yang sangat penting dalam mendukung pengambilan keputusan. Perusahaan umumnya memiliki data dalam jumlah besar yang berasal dari berbagai aktivitas operasional seperti transaksi penjualan, distribusi barang, data pelanggan, hingga data inventaris. Jika data tersebut hanya disimpan dalam bentuk tabel atau dokumen mentah, maka proses analisis akan menjadi lebih sulit dan membutuhkan waktu yang lama. Oleh karena itu, visualisasi data digunakan untuk mengubah data mentah menjadi informasi yang lebih mudah dipahami sehingga pihak manajemen dapat memperoleh insight bisnis secara lebih cepat.

Visualisasi data juga membantu perusahaan dalam mengidentifikasi pola dan tren yang terdapat dalam data. Sebagai contoh, perusahaan dapat mengetahui produk yang paling sering terjual, pelanggan dengan tingkat transaksi tertinggi, maupun perubahan tren penjualan dalam periode tertentu melalui grafik atau dashboard visual. Dengan adanya visualisasi tersebut, perusahaan dapat melakukan evaluasi terhadap performa bisnis secara lebih efektif. Selain itu, visualisasi data juga membantu dalam proses monitoring operasional perusahaan karena informasi dapat ditampilkan secara real-time dan lebih mudah dipahami oleh pengguna.

Dalam proses analisis data, visualisasi sering digunakan sebagai bagian dari exploratory data analysis (EDA), yaitu tahap eksplorasi data untuk memahami karakteristik data sebelum dilakukan proses analisis lebih lanjut.

Melalui visualisasi, pengguna dapat melihat distribusi data, mendeteksi outlier, memahami hubungan antar variabel, hingga mengetahui pola tertentu dalam data. Oleh karena itu, visualisasi data tidak hanya berfungsi sebagai alat presentasi informasi, tetapi juga sebagai alat bantu dalam proses analisis data itu sendiri.

Perkembangan teknologi juga menyebabkan visualisasi data menjadi semakin interaktif dan dinamis. Saat ini, banyak perusahaan menggunakan dashboard interaktif yang memungkinkan pengguna untuk melakukan filter data, memilih kategori tertentu, hingga melihat perubahan data secara langsung. Dashboard interaktif memberikan kemudahan bagi pengguna dalam memahami data tanpa harus membaca laporan secara manual. Selain itu, visualisasi yang interaktif juga membantu perusahaan dalam menyampaikan informasi kepada berbagai pihak, baik manajemen maupun karyawan operasional, secara lebih efektif.

Dalam konteks penelitian ini, visualisasi data digunakan untuk menyajikan hasil *clustering* terhadap data *invoice* perusahaan agar informasi yang dihasilkan lebih mudah dipahami oleh pengguna. Hasil segmentasi pelanggan dan pengelompokan data transaksi akan divisualisasikan dalam bentuk grafik dan dashboard interaktif sehingga perusahaan dapat melihat pola transaksi, kategori pelanggan, serta karakteristik data secara lebih jelas. Dengan adanya visualisasi data, proses analisis tidak hanya menghasilkan informasi dalam bentuk angka, tetapi juga menghasilkan tampilan visual yang dapat membantu perusahaan dalam memahami data dan mendukung pengambilan keputusan bisnis secara lebih efektif.

2.2.7 Dashboard Interaktif

Dashboard interaktif merupakan media visualisasi data yang digunakan untuk menampilkan informasi penting dalam bentuk tampilan visual yang dapat dipantau dan dianalisis secara langsung oleh pengguna[32]. Dashboard biasanya menyajikan data dalam bentuk grafik, diagram, tabel, indikator performa, maupun visualisasi lainnya yang dirancang agar informasi dapat

dipahami dengan lebih cepat dan mudah. Berbeda dengan laporan statis konvensional, dashboard interaktif memungkinkan pengguna untuk berinteraksi langsung dengan data melalui fitur seperti filter, pencarian, pemilihan kategori, hingga eksplorasi data secara dinamis.

Dalam lingkungan bisnis modern, dashboard interaktif menjadi salah satu alat penting dalam mendukung proses monitoring dan pengambilan keputusan. Perusahaan yang memiliki data dalam jumlah besar membutuhkan sistem yang mampu menyajikan informasi secara ringkas namun tetap informatif. Dashboard interaktif membantu perusahaan dalam mengubah data mentah menjadi informasi visual yang mudah dipahami sehingga pengguna dapat melihat kondisi bisnis secara real-time. Dengan adanya dashboard, pihak manajemen dapat memantau performa perusahaan, aktivitas transaksi, maupun kondisi operasional tanpa harus membaca laporan yang panjang dan kompleks.

Salah satu keunggulan utama dashboard interaktif adalah kemampuannya dalam menampilkan informasi secara dinamis dan fleksibel. Pengguna dapat memilih data tertentu yang ingin dianalisis, melakukan filter berdasarkan kategori tertentu, maupun melihat perubahan data secara langsung. Sebagai contoh, perusahaan dapat menampilkan data pelanggan berdasarkan wilayah, kategori produk, atau periode transaksi tertentu. Dengan fitur interaktif tersebut, proses analisis menjadi lebih cepat dan efisien karena pengguna dapat memperoleh informasi sesuai kebutuhan tanpa harus melakukan pengolahan data secara manual.

Dashboard interaktif juga memiliki peran penting dalam meningkatkan efektivitas komunikasi informasi di dalam perusahaan. Informasi yang disajikan dalam bentuk visual cenderung lebih mudah dipahami dibandingkan data mentah dalam bentuk tabel atau dokumen. Hal ini membantu pengguna dari berbagai latar belakang, termasuk manajemen maupun karyawan operasional, dalam memahami kondisi bisnis secara lebih cepat. Selain itu, dashboard juga dapat membantu perusahaan dalam mengidentifikasi pola, tren,

maupun permasalahan operasional yang terjadi berdasarkan data yang ditampilkan.

Dalam proses analisis data dan Business Intelligence, dashboard interaktif sering digunakan sebagai media untuk menyajikan hasil analisis dan visualisasi data. Hasil pengolahan data seperti *clustering*, segmentasi pelanggan, maupun analisis transaksi dapat ditampilkan dalam bentuk dashboard sehingga pengguna dapat memahami hasil analisis secara lebih jelas. Dashboard juga memungkinkan perusahaan untuk melakukan monitoring terhadap hasil analisis secara berkelanjutan dan membantu dalam proses evaluasi strategi bisnis yang diterapkan.

Perkembangan teknologi web dan data *analytics* menyebabkan dashboard interaktif semakin banyak digunakan dalam berbagai bidang industri, termasuk distribusi dan logistik. Dalam perusahaan distribusi, dashboard dapat digunakan untuk memantau aktivitas transaksi, performa pelanggan, distribusi produk, hingga pola penjualan perusahaan. Dengan adanya dashboard interaktif, perusahaan dapat mengakses informasi penting secara lebih cepat dan terstruktur sehingga mendukung proses pengambilan keputusan berbasis data.

Dalam penelitian ini, dashboard interaktif digunakan sebagai media visualisasi untuk menampilkan hasil *clustering* terhadap data *invoice* perusahaan. Dashboard akan menampilkan informasi terkait segmentasi pelanggan, pola transaksi, serta kategori data tertentu dalam bentuk visual yang mudah dipahami. Dengan adanya dashboard interaktif, perusahaan diharapkan dapat memahami hasil analisis data secara lebih efektif, membantu karyawan dalam memahami informasi bisnis, serta mendukung pengambilan keputusan perusahaan secara lebih cepat dan akurat.

2.2.8 Visual Studio Code

Visual Studio Code (VSCode) merupakan salah satu software source-code editor yang dikembangkan oleh Microsoft dan banyak digunakan dalam

proses pengembangan aplikasi maupun pengolahan data[33]. VSCode mendukung berbagai bahasa pemrograman seperti Python, JavaScript, Java, dan C++, serta menyediakan berbagai fitur pendukung seperti syntax highlighting, debugging, extension support, dan integrated terminal yang membantu proses pengembangan program menjadi lebih efisien dan terstruktur.

Dalam bidang data science dan machine learning, VSCode sering digunakan sebagai lingkungan pengembangan (development environment) untuk melakukan proses coding, *preprocessing data*, *modeling*, visualisasi data, hingga *deployment* aplikasi berbasis Python. Dukungan extension Python pada VSCode memungkinkan pengguna menjalankan script Python secara langsung, mengelola *library*, serta melakukan integrasi dengan Jupyter Notebook dan virtual environment dalam satu aplikasi.

Salah satu keunggulan utama VSCode adalah kemampuannya dalam mengintegrasikan berbagai file dan komponen proyek dalam satu workspace yang terstruktur. Pengguna dapat mengelola dataset, file program, *library*, maupun konfigurasi aplikasi secara lebih mudah dalam satu lingkungan pengembangan. Selain itu, VSCode juga memiliki performa yang ringan dan mendukung berbagai extension tambahan yang dapat disesuaikan dengan kebutuhan pengembangan aplikasi.

Dalam penelitian ini, Visual Studio Code (VSCode) digunakan sebagai software utama dalam proses pengembangan dashboard *customer* segmentation berbasis *Streamlit*. VSCode digunakan untuk menyatukan seluruh file yang diperlukan dalam penelitian, mulai dari script *preprocessing data*, proses *clustering*, visualisasi data, hingga implementasi dashboard *deployment* menggunakan Python dan *Streamlit*. Dengan menggunakan VSCode, proses pengembangan aplikasi menjadi lebih terstruktur dan memudahkan integrasi antara hasil analisis data dengan dashboard visualisasi yang dibangun.

2.2.9 *Streamlit*

Streamlit merupakan salah satu *framework* berbasis Python yang digunakan untuk membangun aplikasi web dan dashboard interaktif secara cepat dan sederhana, khususnya dalam bidang data science, machine learning, dan visualisasi data[34]. *Streamlit* dirancang untuk memudahkan proses pengembangan aplikasi analisis data tanpa memerlukan pengetahuan mendalam mengenai pengembangan web seperti HTML, CSS, maupun JavaScript. Dengan menggunakan *Streamlit*, pengembang dapat mengubah script Python menjadi aplikasi web interaktif hanya dengan beberapa baris kode sehingga proses pengembangan dashboard menjadi lebih efisien.

Dalam bidang analisis data, *Streamlit* banyak digunakan untuk menampilkan hasil pengolahan data dalam bentuk visualisasi yang interaktif dan mudah dipahami. *Framework* ini mendukung integrasi dengan berbagai *library* Python seperti Pandas, NumPy, Matplotlib, Plotly, dan Scikit-learn, sehingga sangat cocok digunakan untuk kebutuhan *data mining* dan machine learning. Selain itu, *Streamlit* juga memungkinkan pengguna untuk menampilkan tabel data, grafik, model machine learning, hingga hasil *clustering* dalam satu dashboard yang terintegrasi.

Salah satu keunggulan utama *Streamlit* adalah kemampuannya dalam membangun dashboard interaktif dengan proses implementasi yang relatif sederhana. Pengguna dapat menambahkan berbagai komponen interaktif seperti tombol, filter, dropdown, slider, maupun input data tanpa harus membuat antarmuka web secara manual. Dengan adanya fitur interaktif tersebut, pengguna dapat melakukan eksplorasi data secara lebih fleksibel dan dinamis. Selain itu, *Streamlit* juga mendukung proses pembaruan data secara langsung sehingga visualisasi yang ditampilkan dapat berubah sesuai dengan input atau filter yang dipilih pengguna.

Dalam lingkungan bisnis dan Business Intelligence, *Streamlit* sering digunakan untuk membangun dashboard analitik yang membantu perusahaan dalam memahami data dan mendukung pengambilan keputusan. Dashboard berbasis *Streamlit* dapat digunakan untuk menampilkan informasi seperti

performa penjualan, aktivitas pelanggan, hasil segmentasi pelanggan, hingga hasil analisis data transaksi perusahaan. Dengan tampilan yang sederhana namun interaktif, *Streamlit* membantu pengguna memahami informasi bisnis secara lebih cepat dan efektif.

Selain mudah digunakan, *Streamlit* juga memiliki fleksibilitas yang tinggi dalam proses pengembangan aplikasi. *Framework* ini dapat diintegrasikan dengan berbagai metode analisis data dan machine learning sehingga cocok digunakan dalam penelitian yang melibatkan proses *clustering* maupun visualisasi hasil analisis. Dalam proses *clustering*, hasil pengelompokan data dapat divisualisasikan secara langsung dalam bentuk grafik, tabel, maupun dashboard interaktif sehingga pengguna dapat memahami hasil analisis dengan lebih mudah.

Penggunaan *Streamlit* juga mendukung konsep data-driven decision making karena informasi yang dihasilkan dapat diakses secara lebih cepat dan terstruktur. Dengan adanya dashboard interaktif berbasis *Streamlit*, perusahaan dapat memantau hasil analisis data secara real-time dan melakukan evaluasi terhadap kondisi bisnis berdasarkan data yang tersedia. Hal ini membantu perusahaan dalam meningkatkan efisiensi operasional dan kualitas penambilan keputusan.

Dalam penelitian ini, *Streamlit* digunakan sebagai *framework* utama untuk membangun dashboard interaktif yang menampilkan hasil *clustering* terhadap data *invoice* perusahaan. Dashboard yang dibangun menggunakan *Streamlit* akan menyajikan visualisasi data transaksi, segmentasi pelanggan, serta hasil pengelompokan data dalam bentuk yang mudah dipahami oleh pengguna. Dengan adanya dashboard tersebut, perusahaan diharapkan dapat memahami pola transaksi dan karakteristik pelanggan secara lebih efektif serta mendukung proses pengambilan keputusan berbasis data.

2.2.10 Data Preprocessing

Data preprocessing merupakan tahap awal dalam proses analisis data yang bertujuan untuk mempersiapkan data agar dapat digunakan dalam proses pengolahan dan analisis secara lebih efektif. Data yang diperoleh dari perusahaan umumnya masih berupa data mentah yang belum terstruktur dengan baik dan sering kali memiliki berbagai permasalahan seperti data kosong, data duplikat, inkonsistensi format, maupun kesalahan pencatatan. Oleh karena itu, proses data preprocessing menjadi langkah yang sangat penting untuk meningkatkan kualitas data sebelum dilakukan proses analisis lebih lanjut seperti *clustering* atau machine learning.

Dalam proses analisis data, kualitas data sangat mempengaruhi hasil analisis yang dihasilkan. Data yang tidak bersih atau tidak konsisten dapat menyebabkan hasil *clustering* menjadi kurang akurat dan sulit diinterpretasikan. Sebagai contoh, data transaksi yang memiliki nilai kosong atau duplikat dapat mempengaruhi proses pengelompokan data sehingga menghasilkan *cluster* yang tidak sesuai dengan kondisi sebenarnya. Oleh karena itu, proses preprocessing dilakukan untuk memastikan bahwa data yang digunakan memiliki kualitas yang baik dan siap untuk dianalisis.

Salah satu tahapan penting dalam data preprocessing adalah data cleaning atau pembersihan data. Proses ini dilakukan untuk mengatasi berbagai permasalahan dalam data seperti missing value, duplicate data, maupun data yang tidak valid. Missing value merupakan kondisi ketika terdapat atribut data yang kosong atau tidak memiliki nilai, sedangkan duplicate data terjadi ketika terdapat data yang tercatat lebih dari satu kali. Jika tidak ditangani dengan baik, kedua permasalahan tersebut dapat mempengaruhi hasil analisis data. Oleh karena itu, proses cleaning dilakukan untuk menghapus atau memperbaiki data yang bermasalah agar data menjadi lebih konsisten.

Selain data cleaning, proses preprocessing juga mencakup transformasi data dan normalisasi data. Transformasi data dilakukan untuk mengubah format data agar lebih sesuai dengan kebutuhan analisis. Sebagai contoh, data tanggal dapat diubah menjadi format tertentu atau data kategorikal dapat dikonversi

menjadi bentuk numerik agar dapat diproses oleh algoritma tertentu. Sementara itu, normalisasi data dilakukan untuk menyamakan rentang nilai antar atribut sehingga tidak terjadi dominasi atribut tertentu dalam proses analisis. Hal ini sangat penting terutama pada algoritma *clustering* yang sensitif terhadap perbedaan skala data.

Tahap preprocessing juga mencakup proses seleksi atribut atau feature selection, yaitu proses memilih atribut data yang dianggap paling relevan terhadap tujuan analisis. Dalam data transaksi perusahaan, tidak semua atribut memiliki pengaruh yang signifikan terhadap proses *clustering*. Oleh karena itu, pemilihan atribut yang tepat dapat membantu meningkatkan kualitas hasil *clustering* dan mengurangi kompleksitas proses analisis data. Selain itu, preprocessing juga membantu dalam mengurangi noise atau data yang tidak relevan sehingga hasil analisis menjadi lebih optimal.

Dalam konteks data transaksi perusahaan, proses preprocessing memiliki peran yang sangat penting karena data *invoice* biasanya terdiri dari berbagai jenis atribut dengan format yang beragam. Data transaksi yang besar dan kompleks memerlukan proses preprocessing agar data dapat dianalisis secara lebih terstruktur dan efisien. Tanpa preprocessing yang baik, proses *clustering* dapat menghasilkan pengelompokan data yang kurang akurat dan sulit digunakan sebagai dasar pengambilan keputusan.

Dalam penelitian ini, data preprocessing dilakukan terhadap data *invoice* perusahaan sebelum proses *clustering* dilakukan. Tahapan preprocessing meliputi proses pembersihan data, penghapusan data duplikat, penanganan missing value, transformasi data, serta normalisasi data agar data siap digunakan dalam proses *clustering*. Dengan adanya proses preprocessing, kualitas data diharapkan menjadi lebih baik sehingga hasil segmentasi dan *clustering* yang dihasilkan dapat lebih akurat, relevan, dan mudah dipahami oleh perusahaan.

2.3 *Framework* dan Algoritma Penelitian

2.3.1 CRISP-DM

CRISP-DM atau Cross Industry Standard Process for *Data mining* merupakan salah satu *framework* atau metodologi yang paling banyak digunakan dalam proses *data mining* dan analisis data[35]. CRISP-DM dikembangkan untuk memberikan tahapan kerja yang terstruktur dalam proses pengolahan data, mulai dari memahami permasalahan bisnis hingga implementasi hasil analisis. *Framework* ini bersifat fleksibel dan dapat diterapkan pada berbagai bidang industri, termasuk distribusi, logistik, *retail*, keuangan, hingga *e-commerce*. Oleh karena itu, CRISP-DM menjadi salah satu metode yang paling umum digunakan dalam penelitian berbasis *data mining* dan machine learning.

CRISP-DM terdiri dari enam tahapan utama yang saling berhubungan, yaitu *Business understanding*, *Data understanding*, *Data preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Keenam tahapan tersebut dirancang untuk membantu proses analisis data agar berjalan secara sistematis dan sesuai dengan tujuan bisnis yang ingin dicapai. Selain itu, CRISP-DM juga memungkinkan peneliti untuk kembali ke tahap sebelumnya apabila ditemukan permasalahan atau kebutuhan perbaikan dalam proses analisis data. Dengan demikian, *framework* ini memiliki sifat iteratif dan fleksibel dalam proses implementasinya.

A. *Business understanding*

Tahap pertama dalam CRISP-DM adalah *Business understanding*. Pada tahap ini, peneliti memahami permasalahan bisnis yang terjadi pada perusahaan serta menentukan tujuan penelitian yang ingin dicapai. Dalam penelitian ini, permasalahan utama perusahaan adalah jumlah data *invoice* yang sangat besar dan belum terorganisir dengan baik sehingga menyulitkan perusahaan dalam memahami pola transaksi pelanggan maupun produk unggulan. Oleh karena itu, penelitian dilakukan untuk membantu perusahaan mengelompokkan

data transaksi menggunakan metode *clustering* agar informasi bisnis menjadi lebih terstruktur dan mudah dipahami.

B. *Data understanding*

Tahap kedua adalah *Data understanding*, yaitu proses memahami data yang digunakan dalam penelitian. Pada tahap ini dilakukan pengumpulan data, identifikasi atribut data, serta analisis awal terhadap karakteristik data yang dimiliki perusahaan. Data yang digunakan dalam penelitian ini berupa data *invoice* perusahaan yang memuat informasi transaksi pelanggan. Tahap ini juga mencakup proses eksplorasi data awal untuk mengetahui kondisi data seperti jumlah data, jenis atribut, missing value, duplicate data, maupun distribusi data yang akan dianalisis.

C. *Data preparation*

Tahap ketiga adalah *Data preparation*, yaitu proses persiapan data sebelum dilakukan proses *modeling* atau analisis data. Tahap ini merupakan salah satu tahap yang paling penting karena kualitas data sangat mempengaruhi hasil analisis yang dihasilkan. Dalam tahap ini dilakukan proses data preprocessing seperti pembersihan data, penghapusan duplicate data, penanganan missing value, transformasi data, normalisasi data, serta pemilihan atribut yang relevan terhadap penelitian. Tujuan dari tahap ini adalah menghasilkan dataset yang bersih, konsisten, dan siap digunakan dalam proses *clustering*.

D. *Modeling*

Tahap keempat adalah *Modeling*, yaitu proses penerapan algoritma atau metode analisis data terhadap dataset yang telah dipersiapkan sebelumnya. Dalam penelitian ini, proses *modeling* dilakukan menggunakan beberapa algoritma *clustering* seperti *K-means*,*DBSCAN*, *Agglomerative Clustering*, dan untuk mengelompokkan data *invoice* berdasarkan karakteristik tertentu. Pada

tahap ini juga dilakukan penyesuaian parameter algoritma agar proses *clustering* dapat menghasilkan *cluster* yang optimal dan sesuai dengan tujuan penelitian.

E. Evalution

Tahap kelima adalah *Evaluation*, yaitu proses evaluasi terhadap hasil model atau hasil *clustering* yang telah dihasilkan sebelumnya. Tahap ini bertujuan untuk memastikan bahwa hasil *clustering* yang diperoleh telah sesuai dengan tujuan bisnis dan memiliki kualitas yang baik. Dalam penelitian ini, evaluasi *clustering* dilakukan menggunakan beberapa metode evaluasi seperti *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index* untuk menentukan algoritma *clustering* yang memberikan hasil terbaik. Selain itu, hasil *clustering* juga dianalisis untuk memahami pola pelanggan dan karakteristik data transaksi perusahaan.

F. *Deployment*

Tahap terakhir dalam CRISP-DM adalah *Deployment*, yaitu proses implementasi atau penyajian hasil analisis kepada pengguna. Dalam penelitian ini, hasil *clustering* dan analisis data divisualisasikan menggunakan dashboard interaktif berbasis *Streamlit* sehingga informasi yang dihasilkan dapat dipahami dengan lebih mudah oleh perusahaan. Dashboard tersebut menampilkan hasil segmentasi pelanggan, pola transaksi, serta visualisasi data lainnya yang dapat membantu perusahaan dalam mendukung pengambilan keputusan berbasis data. Dengan adanya tahap *deployment*, hasil penelitian tidak hanya berhenti pada proses analisis, tetapi juga dapat dimanfaatkan secara langsung oleh perusahaan dalam aktivitas bisnis sehari-hari.

Salah satu alasan utama pemilihan CRISP-DM dalam penelitian ini adalah karena *framework* tersebut memiliki tahapan yang terstruktur, sistematis, serta fleksibel dalam proses analisis data. CRISP-DM tidak hanya berfokus

pada proses teknis pengolahan data, tetapi juga memperhatikan pemahaman terhadap permasalahan bisnis yang menjadi dasar penelitian. Dalam penelitian ini, permasalahan utama berasal dari banyaknya data *invoice* perusahaan yang belum terorganisir dengan baik sehingga menyulitkan perusahaan dalam memahami pola transaksi, pelanggan unggulan, serta produk yang paling dominan. Oleh karena itu, CRISP-DM dipilih karena mampu menghubungkan kebutuhan bisnis perusahaan dengan proses analisis data secara menyeluruh mulai dari pemahaman masalah, pengolahan data, proses *clustering*, hingga implementasi hasil analisis dalam bentuk dashboard interaktif.

Selain itu, CRISP-DM juga dipilih karena *framework* ini sangat sesuai digunakan dalam penelitian berbasis *data mining* dan machine learning, khususnya pada penelitian *clustering*. Tahapan yang dimiliki CRISP-DM memungkinkan proses penelitian dilakukan secara bertahap dan lebih terarah sehingga mempermudah peneliti dalam melakukan analisis data. *Framework* ini juga bersifat iteratif, yang berarti setiap tahapan dapat diperbaiki atau diulang kembali apabila ditemukan permasalahan dalam proses penelitian. Hal tersebut menjadi salah satu keunggulan CRISP-DM dibandingkan metode lain karena proses analisis data sering kali membutuhkan penyesuaian berulang agar menghasilkan model yang optimal.

Selain CRISP-DM, terdapat beberapa *framework* lain yang juga dapat digunakan dalam penelitian berbasis *data mining* dan analisis data. Salah satunya adalah Knowledge Discovery in Databases (KDD). *Framework* KDD merupakan metode yang berfokus pada proses penemuan pengetahuan dari data melalui beberapa tahapan seperti selection, preprocessing, transformation, *data mining*, dan interpretation. KDD banyak digunakan dalam penelitian *data mining* karena memiliki proses yang cukup sistematis dalam pengolahan data. Namun, KDD lebih berorientasi pada proses teknis *data mining* dan kurang menekankan pada pemahaman kebutuhan bisnis dibandingkan CRISP-DM. Oleh karena itu, *framework* ini dianggap kurang sesuai untuk penelitian yang

memiliki fokus pada penyelesaian permasalahan bisnis perusahaan secara langsung.

Framework lain yang juga dapat digunakan adalah SEMMA (Sample, Explore, Modify, Model, Assess) yang dikembangkan oleh SAS Institute. SEMMA berfokus pada tahapan analisis data dan pengembangan model machine learning melalui proses sampling data, eksplorasi data, modifikasi data, *modeling*, dan evaluasi model. *Framework* ini cukup efektif digunakan untuk penelitian berbasis machine learning karena memiliki fokus yang kuat pada proses analisis data dan pengembangan model. Namun, SEMMA lebih berorientasi pada proses teknis pengolahan data dan tidak memiliki tahapan *business understanding* secara khusus seperti pada CRISP-DM. Oleh karena itu, dalam penelitian ini CRISP-DM dianggap lebih sesuai karena mampu mengintegrasikan kebutuhan bisnis perusahaan dengan proses analisis data secara lebih lengkap.

Penerapan CRISP-DM dalam penelitian ini dimulai dari tahap *Business understanding*, yaitu memahami permasalahan yang terjadi pada perusahaan terkait data *invoice* yang belum terorganisir dengan baik. Permasalahan tersebut menyebabkan perusahaan kesulitan dalam memahami pola transaksi pelanggan serta menentukan pelanggan dan produk unggulan. Setelah itu, tahap *Data understanding* dilakukan dengan mempelajari karakteristik data *invoice* perusahaan, memahami atribut data, serta melakukan eksplorasi awal terhadap data yang dimiliki. Tahap ini bertujuan untuk mengetahui kondisi data sebelum dilakukan proses analisis lebih lanjut.

Selanjutnya, pada tahap *Data preparation* dilakukan proses *preprocessing data* seperti pembersihan data, penghapusan data duplikat, penanganan missing value, transformasi data, serta normalisasi data agar dataset siap digunakan dalam proses *clustering*. Setelah data siap digunakan, penelitian dilanjutkan ke tahap *Modeling* dengan menerapkan beberapa algoritma *clustering* seperti *K-means*, *DBSCAN*, *Agglomerative clustering*, dan GMM untuk mengelompokkan data *invoice* berdasarkan karakteristik tertentu. Pada

tahap ini juga dilakukan penyesuaian parameter untuk memperoleh hasil *clustering* yang optimal.

Tahap berikutnya adalah *Evaluation*, yaitu melakukan evaluasi terhadap hasil *clustering* menggunakan beberapa metode evaluasi seperti *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index* untuk menentukan algoritma yang memberikan performa terbaik. Selain itu, hasil *clustering* juga dianalisis untuk memahami pola transaksi pelanggan dan karakteristik data perusahaan. Tahap terakhir adalah *Deployment*, yaitu menyajikan hasil *clustering* dalam bentuk dashboard interaktif berbasis *Streamlit* agar perusahaan dapat memahami hasil analisis data secara lebih mudah dan mendukung proses pengambilan keputusan berbasis data. Dengan penerapan CRISP-DM, penelitian ini diharapkan mampu menghasilkan proses analisis data yang terstruktur, sistematis, dan relevan terhadap kebutuhan perusahaan.

2.3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan proses analisis awal terhadap dataset yang bertujuan untuk memahami karakteristik data sebelum dilakukan proses *modeling* atau analisis lanjutan[36]. EDA digunakan untuk mengeksplorasi struktur data, mengetahui pola distribusi data, mendeteksi adanya missing value, outlier, duplikasi data, serta memahami hubungan antar variabel dalam dataset. Proses ini menjadi salah satu tahapan penting dalam *data mining* dan machine learning karena kualitas data sangat mempengaruhi hasil analisis dan performa model yang digunakan.

Dalam proses Exploratory Data Analysis, analisis dapat dilakukan menggunakan pendekatan statistik maupun visualisasi data. Analisis statistik digunakan untuk melihat informasi dasar seperti jumlah data, rata-rata, nilai minimum, nilai maksimum, standar deviasi, dan distribusi data. Sementara itu, visualisasi data digunakan untuk membantu memahami pola dan persebaran data secara lebih intuitif melalui grafik seperti histogram, scatter plot, boxplot, heatmap, maupun diagram distribusi lainnya.

Tujuan utama dari EDA adalah memperoleh pemahaman yang lebih mendalam terhadap dataset sehingga peneliti dapat menentukan teknik preprocessing, transformasi data, serta metode analisis yang paling sesuai dengan karakteristik data yang digunakan. Selain itu, EDA juga membantu dalam mengidentifikasi potensi permasalahan pada dataset seperti data tidak konsisten, noise, skewness, maupun ketidakseimbangan distribusi data yang dapat mempengaruhi hasil *modeling*.

Dalam penelitian *clustering*, Exploratory Data Analysis memiliki peran penting karena proses *clustering* sangat bergantung pada pola distribusi dan karakteristik data. Melalui EDA, peneliti dapat memahami persebaran data *customer*, pola transaksi, hubungan antar atribut, serta mendeteksi adanya outlier yang dapat mempengaruhi kualitas *cluster* yang dihasilkan. Hasil EDA juga dapat digunakan sebagai dasar dalam menentukan metode preprocessing seperti standardization, transformation, maupun dimensionality reduction sebelum proses *clustering* dilakukan.

Pada penelitian ini, Exploratory Data Analysis dilakukan terhadap data *invoice* perusahaan untuk memahami karakteristik data transaksi pelanggan sebelum proses *clustering* dilakukan. Analisis dilakukan dengan melihat tipe data, distribusi data transaksi, jumlah data, serta hubungan antar atribut yang digunakan dalam penelitian. Selain itu, EDA juga digunakan untuk membantu proses identifikasi data yang tidak valid dan menentukan atribut yang paling relevan untuk digunakan dalam proses segmentasi pelanggan.

2.3.3 Algoritma dan Metode Evaluasi yang digunakan

2.3.2.1 *K-means*

K-means merupakan salah satu algoritma *clustering* yang paling banyak digunakan dalam proses *data mining* dan machine learning, khususnya pada penelitian segmentasi pelanggan dan analisis data transaksi[37]. *K-means* termasuk ke dalam kategori unsupervised learning karena proses pengelompokan data dilakukan tanpa menggunakan label atau kategori data sebelumnya. Algoritma ini bekerja dengan cara

mengelompokkan data ke dalam beberapa *cluster* berdasarkan tingkat kemiripan antar data sehingga data yang berada dalam satu *cluster* memiliki karakteristik yang lebih mirip dibandingkan dengan data pada *cluster* lainnya.

Tujuan utama dari algoritma *K-means* adalah meminimalkan jarak antar data dalam satu *cluster* dan memaksimalkan jarak antar *cluster* yang berbeda. Dalam proses *clustering*, *K-means* menggunakan titik pusat *cluster* yang disebut centroid sebagai acuan dalam menentukan pengelompokan data. Setiap data akan ditempatkan ke dalam *cluster* yang memiliki centroid terdekat berdasarkan perhitungan jarak tertentu, yang umumnya menggunakan Euclidean Distance. Oleh karena itu, algoritma *K-means* sangat bergantung pada posisi centroid awal dan karakteristik data yang digunakan.

Proses kerja algoritma *K-means* dimulai dengan menentukan jumlah *cluster* atau nilai K yang akan digunakan dalam proses *clustering*. Setelah jumlah *cluster* ditentukan, algoritma akan memilih centroid awal secara acak. Selanjutnya, setiap data akan dihitung jaraknya terhadap seluruh centroid menggunakan metode perhitungan jarak seperti Euclidean Distance. Data kemudian akan dimasukkan ke dalam *cluster* yang memiliki jarak terdekat terhadap centroid. Setelah seluruh data dikelompokkan, centroid baru akan dihitung berdasarkan rata-rata nilai seluruh data dalam masing-masing *cluster*. Proses tersebut dilakukan secara berulang hingga posisi centroid tidak mengalami perubahan signifikan atau hingga mencapai jumlah iterasi tertentu.

Perhitungan jarak pada algoritma *K-means* umumnya menggunakan Euclidean Distance karena metode ini efektif dalam mengukur jarak antar data numerik. Rumus Euclidean Distance dapat dituliskan sebagai berikut:

Algoritma *K-means* memiliki beberapa kelebihan yang menyebabkan metode ini banyak digunakan dalam penelitian *clustering*.

Salah satu kelebihannya adalah proses komputasi yang relatif cepat dan efisien, terutama pada dataset dengan jumlah data yang besar. Selain itu, *K-means* juga mudah diimplementasikan dan hasil *clustering* yang dihasilkan cukup mudah dipahami serta divisualisasikan. Oleh karena itu, algoritma ini sangat cocok digunakan dalam segmentasi pelanggan dan analisis data transaksi perusahaan.

Meskipun demikian, *K-means* juga memiliki beberapa kelemahan. Salah satu kelemahan utama algoritma ini adalah sensitivitas terhadap pemilihan centroid awal, yang dapat mempengaruhi hasil *clustering* yang dihasilkan. Selain itu, *K-means* juga kurang optimal dalam menangani data dengan distribusi yang tidak merata, data yang memiliki noise atau outlier, serta data non-numerik. Oleh karena itu, proses preprocessing dan pemilihan fitur yang tepat menjadi faktor penting dalam meningkatkan kualitas hasil *clustering* menggunakan *K-means*.

Dalam penelitian ini, algoritma *K-means* digunakan untuk mengelompokkan data *invoice* perusahaan berdasarkan karakteristik transaksi pelanggan. Data yang digunakan telah melalui proses preprocessing dan transformasi agar dapat diproses menggunakan algoritma *K-means*. Hasil *clustering* dari algoritma ini diharapkan dapat membantu perusahaan dalam mengidentifikasi kelompok pelanggan berdasarkan tingkat prioritas, pola transaksi, serta kontribusi pelanggan terhadap perusahaan. Selain itu, hasil *clustering* juga akan dibandingkan dengan algoritma lain menggunakan beberapa metode evaluasi untuk menentukan algoritma yang memberikan performa terbaik dalam penelitian ini.

2.3.2.2 *Agglomerative Clustering*

Agglomerative Clustering merupakan salah satu metode *clustering* yang digunakan untuk mengelompokkan data berdasarkan tingkat kemiripan antar data secara bertahap hingga membentuk struktur hierarki *cluster*[38]. Berbeda dengan algoritma *K-means* yang mengharuskan jumlah *cluster* ditentukan di awal, *Agglomerative Clustering* membentuk

cluster secara bertingkat sehingga hubungan antar data dan antar *cluster* dapat divisualisasikan dalam bentuk struktur pohon yang disebut dendrogram. Metode ini termasuk ke dalam kategori unsupervised learning karena proses pengelompokan dilakukan tanpa menggunakan label data sebelumnya.

Agglomerative Clustering merupakan bagian dari metode hierarchical *clustering* dengan pendekatan bottom-up, di mana setiap data pada awalnya dianggap sebagai satu *cluster* tersendiri, kemudian *cluster-cluster* tersebut akan digabungkan secara bertahap berdasarkan tingkat kemiripan hingga membentuk *cluster* yang lebih besar. Proses penggabungan tersebut dilakukan terus menerus hingga seluruh data tergabung dalam satu struktur hierarki *cluster*. Pendekatan ini menjadi salah satu metode hierarchical *clustering* yang paling umum digunakan karena implementasinya yang relatif sederhana dan mampu menghasilkan struktur *cluster* yang cukup baik pada berbagai jenis data.

Proses kerja *Agglomerative Clustering* dimulai dengan menghitung jarak atau tingkat kemiripan antar data menggunakan metode distance measurement tertentu seperti Euclidean Distance. Setelah itu, algoritma akan mencari dua data atau *cluster* yang memiliki jarak paling dekat untuk digabungkan menjadi satu *cluster* baru. Proses penggabungan dilakukan secara berulang hingga jumlah *cluster* yang diinginkan tercapai atau seluruh data tergabung dalam satu *cluster* besar. Hasil pengelompokan kemudian dapat divisualisasikan dalam bentuk dendrogram yang menunjukkan hubungan hierarki antar *cluster* berdasarkan tingkat kemiripan data.

Dalam *Agglomerative Clustering*, terdapat beberapa metode linkage yang digunakan untuk menentukan jarak antar *cluster*. Metode linkage yang umum digunakan antara lain single linkage, complete linkage, average linkage, dan ward linkage. Single linkage menggunakan jarak terdekat antar anggota *cluster*, sedangkan complete linkage menggunakan jarak terjauh antar anggota *cluster*. Average linkage menggunakan rata-rata jarak antar

anggota *cluster*, sementara ward linkage berfokus pada minimisasi variasi dalam *cluster*. Pemilihan metode linkage dapat mempengaruhi bentuk dan kualitas *cluster* yang dihasilkan sehingga perlu disesuaikan dengan karakteristik data yang digunakan.

Salah satu keunggulan utama *Agglomerative Clustering* adalah kemampuannya dalam menunjukkan hubungan antar data dan antar *cluster* secara lebih jelas melalui dendrogram. Dengan adanya dendrogram, pengguna dapat memahami struktur *cluster* dan menentukan jumlah *cluster* yang optimal berdasarkan visualisasi hubungan antar data. Selain itu, *Agglomerative Clustering* juga lebih fleksibel dalam menangani bentuk *cluster* yang kompleks dibandingkan beberapa algoritma *clustering* lainnya. Metode ini juga tidak terlalu bergantung pada pemilihan centroid awal seperti pada *K-means*.

Meskipun memiliki beberapa keunggulan, *Agglomerative Clustering* juga memiliki kelemahan, terutama dalam hal kompleksitas komputasi. Proses perhitungan jarak antar seluruh data menyebabkan algoritma ini membutuhkan waktu komputasi yang lebih besar dibandingkan *K-means*, khususnya pada dataset dengan jumlah data yang sangat besar. Selain itu, hasil *clustering* pada *Agglomerative Clustering* juga sensitif terhadap noise dan outlier sehingga proses *preprocessing data* menjadi sangat penting sebelum algoritma diterapkan.

Dalam penelitian ini, *Agglomerative Clustering* digunakan sebagai salah satu algoritma pembanding dalam proses segmentasi pelanggan berdasarkan data *invoice* perusahaan. Algoritma ini digunakan untuk melihat struktur hubungan antar data pelanggan berdasarkan karakteristik transaksi yang dimiliki. Selain itu, penggunaan *Agglomerative Clustering* juga bertujuan untuk membandingkan performa dan kualitas *cluster* yang dihasilkan dengan algoritma *clustering* lainnya seperti *K-means*, *DBSCAN*, dan *Gaussian Mixture Model* (GMM). Hasil *clustering* yang diperoleh kemudian akan dievaluasi menggunakan beberapa metode evaluasi

clustering untuk menentukan algoritma yang paling optimal dalam mengelompokkan data pelanggan berdasarkan tingkat prioritas *customer*.

2.3.2.3 *DBSCAN*

DBSCAN atau *Density-based* Spatial *Clustering* of Applications with Noise merupakan salah satu algoritma *clustering* berbasis density atau kepadatan data yang digunakan untuk mengelompokkan data berdasarkan tingkat kepadatan tertentu. Berbeda dengan algoritma[39] *K-means* dan *Agglomerative clustering* yang berfokus pada jarak antar data atau centroid, *DBSCAN* bekerja dengan cara mengidentifikasi area data yang memiliki kepadatan tinggi dan memisahkannya dari area dengan kepadatan rendah. Algoritma ini termasuk ke dalam kategori unsupervised learning karena proses *clustering* dilakukan tanpa menggunakan label data sebelumnya.

Salah satu karakteristik utama *DBSCAN* adalah kemampuannya dalam mendeteksi noise dan outlier pada dataset. Dalam proses *clustering*, data yang berada pada area dengan kepadatan rendah dan tidak memenuhi syarat pembentukan *cluster* akan dianggap sebagai noise atau data outlier. Hal ini menjadi salah satu keunggulan *DBSCAN* dibandingkan beberapa algoritma *clustering* lainnya karena dalam data transaksi perusahaan sering terdapat data yang memiliki pola transaksi tidak normal atau berbeda secara signifikan dibandingkan data lainnya.

DBSCAN menggunakan dua parameter utama dalam proses *clustering*, yaitu epsilon (ε) dan minimum points (MinPts). Parameter epsilon digunakan untuk menentukan radius atau jarak maksimum antar data agar dapat dianggap berada dalam satu area kepadatan yang sama. Sementara itu, parameter MinPts digunakan untuk menentukan jumlah minimum data yang harus berada dalam radius epsilon agar suatu area dapat dianggap sebagai *cluster*. Pemilihan nilai epsilon dan MinPts sangat mempengaruhi hasil *clustering* yang dihasilkan sehingga perlu dilakukan penyesuaian berdasarkan karakteristik dataset yang digunakan.

Dalam *DBSCAN*, data dibagi menjadi tiga kategori utama, yaitu core point, border point, dan noise point. Core point merupakan data yang memiliki jumlah tetangga sesuai atau melebihi nilai MinPts dalam radius epsilon tertentu. Border point merupakan data yang berada di sekitar core point tetapi tidak memiliki jumlah tetangga yang cukup untuk menjadi core point. Sedangkan noise point adalah data yang tidak termasuk ke dalam *cluster* manapun karena berada pada area dengan kepadatan rendah. Dengan adanya pembagian tersebut, *DBSCAN* mampu membentuk *cluster* dengan bentuk yang lebih fleksibel dibandingkan algoritma *clustering* berbasis centroid seperti *K-means*.

Proses kerja *DBSCAN* dimulai dengan memilih satu data secara acak, kemudian algoritma akan mencari seluruh data yang berada dalam radius epsilon dari data tersebut. Jika jumlah data di sekitar titik tersebut memenuhi nilai minimum points, maka data tersebut akan menjadi core point dan membentuk *cluster* baru. Selanjutnya, algoritma akan terus memperluas *cluster* dengan mencari data lain yang terhubung dengan core point tersebut hingga seluruh area dengan kepadatan yang sama tergabung dalam satu *cluster*. Proses ini dilakukan hingga seluruh data telah diperiksa dan dikelompokkan.

Salah satu kelebihan utama *DBSCAN* adalah kemampuannya dalam membentuk *cluster* dengan bentuk yang kompleks dan tidak beraturan. Selain itu, algoritma ini juga tidak memerlukan penentuan jumlah *cluster* di awal seperti pada *K-means*, sehingga lebih fleksibel dalam proses *clustering*. *DBSCAN* juga sangat efektif dalam mendeteksi noise dan outlier pada data sehingga cocok digunakan pada dataset yang memiliki distribusi data tidak merata. Oleh karena itu, algoritma ini banyak digunakan dalam penelitian yang melibatkan data dengan pola distribusi yang kompleks.

Meskipun memiliki berbagai keunggulan, *DBSCAN* juga memiliki beberapa kelemahan. Salah satu kelemahan utama algoritma ini adalah sensitivitas terhadap pemilihan parameter epsilon dan MinPts. Jika nilai

parameter yang digunakan tidak sesuai, maka hasil *clustering* dapat menjadi kurang optimal. Selain itu, *DBSCAN* juga kurang efektif pada dataset dengan tingkat kepadatan yang sangat bervariasi karena algoritma ini menggunakan parameter kepadatan yang sama untuk seluruh data. Pada dataset dengan jumlah data yang sangat besar, proses perhitungan tetangga antar data juga dapat meningkatkan kompleksitas komputasi.

Dalam penelitian ini, *DBSCAN* digunakan sebagai salah satu algoritma *clustering* untuk mengelompokkan data *invoice* perusahaan berdasarkan pola transaksi pelanggan. Penggunaan *DBSCAN* bertujuan untuk melihat kemampuan algoritma dalam mendeteksi pola transaksi pelanggan berdasarkan kepadatan data serta mengidentifikasi kemungkinan adanya outlier atau pelanggan dengan karakteristik transaksi yang berbeda secara signifikan. Hasil *clustering* dari *DBSCAN* kemudian akan dibandingkan dengan algoritma *K-means* dan *Agglomerative Clustering* menggunakan beberapa metode evaluasi *clustering* untuk menentukan algoritma yang memberikan performa terbaik dalam segmentasi prioritas pelanggan perusahaan.

2.3.2.4 *Gaussian Mixture Model* (GMM)

Gaussian Mixture Model (GMM) merupakan salah satu metode *clustering* berbasis probabilistik yang digunakan untuk mengelompokkan data berdasarkan distribusi probabilitas Gaussian atau distribusi normal[40] .Berbeda dengan algoritma *K-means* yang mengelompokkan data berdasarkan jarak terhadap centroid terdekat, GMM menentukan probabilitas suatu data termasuk ke dalam suatu *cluster* tertentu. Metode ini termasuk ke dalam kategori unsupervised learning karena proses pengelompokan dilakukan tanpa menggunakan label data sebelumnya.

Pada *Gaussian Mixture Model*, setiap *cluster* diasumsikan memiliki distribusi Gaussian dengan parameter tertentu seperti mean dan covariance. Setiap data tidak secara mutlak dimasukkan ke dalam satu *cluster* tertentu, melainkan memiliki nilai probabilitas terhadap seluruh *cluster* yang

tersedia. Data kemudian akan dimasukkan ke dalam *cluster* dengan probabilitas tertinggi. Pendekatan probabilistik tersebut membuat GMM lebih fleksibel dalam menangani bentuk distribusi data yang kompleks dibandingkan algoritma *clustering* berbasis centroid seperti *K-means*.

Proses kerja *Gaussian Mixture Model* umumnya menggunakan algoritma Expectation-Maximization (EM). Pada tahap expectation, algoritma menghitung probabilitas setiap data terhadap masing-masing *cluster* berdasarkan parameter distribusi Gaussian yang dimiliki. Selanjutnya pada tahap maximization, parameter distribusi seperti mean, covariance, dan probabilitas *cluster* akan diperbarui berdasarkan hasil probabilitas sebelumnya. Proses tersebut dilakukan secara berulang hingga model mencapai kondisi konvergen atau stabil.

Salah satu keunggulan utama *Gaussian Mixture Model* adalah kemampuannya dalam menangani *cluster* dengan bentuk distribusi yang tidak simetris dan memiliki overlap antar *cluster*. GMM juga lebih fleksibel dalam memodelkan data yang memiliki variasi bentuk dan ukuran *cluster* yang berbeda-beda. Selain itu, pendekatan probabilistik yang digunakan memungkinkan algoritma ini memberikan interpretasi yang lebih baik terhadap kemungkinan suatu data termasuk ke dalam *cluster* tertentu.

Meskipun memiliki beberapa keunggulan, *Gaussian Mixture Model* juga memiliki kelemahan, terutama dalam hal kompleksitas komputasi dan sensitivitas terhadap pemilihan jumlah *cluster*. Proses perhitungan probabilitas dan covariance matrix menyebabkan GMM membutuhkan waktu komputasi yang lebih besar dibandingkan beberapa algoritma *clustering* lainnya. Selain itu, jika distribusi data tidak sesuai dengan asumsi distribusi Gaussian, maka kualitas *cluster* yang dihasilkan dapat menjadi kurang optimal.

Dalam penelitian ini, *Gaussian Mixture Model* digunakan sebagai salah satu algoritma pembanding dalam proses segmentasi pelanggan

berdasarkan data *invoice* perusahaan. Algoritma ini digunakan untuk melihat kemampuan pendekatan probabilistik dalam mengelompokkan pelanggan berdasarkan karakteristik transaksi yang dimiliki. Selain itu, penggunaan GMM juga bertujuan untuk membandingkan performa dan kualitas *cluster* yang dihasilkan dengan algoritma *clustering* lainnya seperti *K-means*, *DBSCAN*, dan *Agglomerative Clustering*. Hasil *clustering* yang diperoleh kemudian akan dievaluasi menggunakan beberapa metode evaluasi *clustering* untuk menentukan algoritma yang paling optimal dalam mengelompokkan data pelanggan berdasarkan tingkat prioritas *customer*.

2.3.2.5 Metode Evaluasi

Metode evaluasi *clustering* merupakan proses yang digunakan untuk mengukur kualitas hasil *clustering* yang dihasilkan oleh suatu algoritma[41]. Evaluasi *clustering* dilakukan untuk mengetahui seberapa baik data telah dikelompokkan ke dalam *cluster* yang memiliki tingkat kemiripan tinggi di dalam *cluster* yang sama dan perbedaan yang jelas antar *cluster* yang berbeda. Dalam penelitian *clustering*, proses evaluasi menjadi sangat penting karena algoritma *clustering* termasuk ke dalam metode unsupervised learning yang tidak memiliki label data sebagai acuan utama dalam pengukuran performa model.

Tujuan utama dari evaluasi *clustering* adalah menentukan apakah hasil *cluster* yang terbentuk sudah optimal dan sesuai dengan tujuan analisis yang dilakukan. Hasil *clustering* yang baik umumnya memiliki karakteristik berupa jarak antar data dalam satu *cluster* yang kecil serta jarak antar *cluster* yang besar. Dengan kata lain, data yang berada dalam *cluster* yang sama harus memiliki tingkat kemiripan yang tinggi, sedangkan data pada *cluster* yang berbeda harus memiliki karakteristik yang cukup berbeda. Oleh karena itu, metode evaluasi digunakan untuk mengukur kualitas *cluster* secara objektif berdasarkan perhitungan tertentu.

Dalam penelitian ini, metode evaluasi *clustering* digunakan untuk membandingkan performa beberapa algoritma *clustering* seperti *K-means*,

Agglomerative, *DBSCAN*, dam GMM dalam mengelompokkan data *invoice* perusahaan. Hasil evaluasi akan digunakan untuk menentukan algoritma yang memberikan hasil *clustering* paling optimal dalam segmentasi prioritas pelanggan. Selain itu, evaluasi *clustering* juga membantu dalam menentukan jumlah *cluster* yang paling sesuai berdasarkan karakteristik data yang digunakan.

Secara umum, metode evaluasi *clustering* dibagi menjadi beberapa kategori, salah satunya adalah internal *evaluation*. Internal *evaluation* merupakan metode evaluasi yang menggunakan informasi dari data *clustering* itu sendiri tanpa memerlukan label data eksternal. Metode ini mengevaluasi kualitas *cluster* berdasarkan tingkat kohesi dalam *cluster* dan separasi antar *cluster*. Dalam penelitian *clustering* berbasis data transaksi perusahaan, internal *evaluation* menjadi metode yang paling umum digunakan karena data yang digunakan tidak memiliki label *cluster* sebelumnya.

Beberapa metode evaluasi *clustering* yang paling sering digunakan dalam penelitian *clustering* antara lain *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*. Ketiga metode tersebut digunakan karena mampu memberikan pengukuran kualitas *clustering* berdasarkan karakteristik *cluster* yang terbentuk. *Silhouette Score* digunakan untuk mengukur tingkat kemiripan data dalam *cluster* dibandingkan dengan *cluster* lainnya. *Davies-Bouldin Index* digunakan untuk mengukur rasio antara jarak dalam *cluster* dan jarak antar *cluster*, sedangkan *Calinski-Harabasz Index* digunakan untuk mengukur perbandingan antara variasi antar *cluster* dan variasi dalam *cluster*.

Penggunaan lebih dari satu metode evaluasi *clustering* sangat penting karena setiap metode memiliki pendekatan pengukuran yang berbeda. Dalam beberapa kasus, suatu algoritma dapat menghasilkan nilai evaluasi yang baik pada satu metode tetapi kurang optimal pada metode lainnya. Oleh karena itu, kombinasi beberapa metode evaluasi dapat

memberikan hasil analisis yang lebih objektif dan komprehensif dalam menentukan kualitas *clustering* yang dihasilkan. Selain itu, penggunaan multi-*evaluation* juga membantu meningkatkan validitas hasil penelitian.

Dalam penelitian ini, metode evaluasi *clustering* digunakan untuk menentukan algoritma yang paling sesuai dalam melakukan segmentasi prioritas pelanggan berdasarkan data *invoice* perusahaan. Hasil evaluasi dari masing-masing algoritma akan dibandingkan untuk melihat kualitas *cluster* yang dihasilkan. Algoritma dengan hasil evaluasi terbaik akan dipertimbangkan sebagai metode yang paling optimal dalam membantu perusahaan memahami karakteristik pelanggan dan mendukung pengambilan keputusan berbasis data.

2.4 Tools dan software yang digunakan

2.4.1 Python

Python merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan dalam bidang data science, machine learning, artificial intelligence, dan *data mining*[42]. Python memiliki sintaks yang sederhana dan mudah dipahami sehingga menjadi salah satu bahasa pemrograman yang paling populer dalam proses analisis data. Selain itu, Python juga didukung oleh berbagai *library* dan *framework* yang lengkap untuk kebutuhan pengolahan data, visualisasi data, hingga implementasi machine learning. Oleh karena itu, Python banyak digunakan dalam penelitian berbasis analisis data karena mampu mendukung proses pengembangan sistem secara lebih efisien dan fleksibel.

Dalam bidang *data mining* dan machine learning, Python memiliki berbagai keunggulan seperti kemudahan integrasi dengan *library* analisis data, kemampuan pengolahan data dalam jumlah besar, serta dukungan komunitas yang sangat luas. Python juga mendukung berbagai metode analisis data dan algoritma machine learning yang dapat digunakan untuk melakukan proses *clustering*, klasifikasi, prediksi, maupun visualisasi data. Hal tersebut menyebabkan Python menjadi salah satu software utama yang paling sering digunakan dalam penelitian data science dan Business Intelligence.

Dalam penelitian ini, Python digunakan sebagai bahasa pemrograman utama dalam seluruh proses analisis data dan implementasi sistem. Python digunakan untuk melakukan proses *preprocessing data*, pengolahan data transaksi perusahaan, implementasi algoritma *clustering*, evaluasi hasil *clustering*, hingga proses visualisasi data. Selain itu, Python juga digunakan untuk mengintegrasikan berbagai *library* dan *framework* yang dibutuhkan dalam penelitian seperti Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, dan *Streamlit*.

Penggunaan Python dalam penelitian ini juga membantu proses pengolahan data *invoice* perusahaan yang memiliki jumlah data cukup besar dan kompleks. Dengan menggunakan Python, proses analisis data dapat dilakukan secara lebih otomatis, terstruktur, dan efisien dibandingkan proses manual. Selain itu, Python juga memungkinkan implementasi dashboard interaktif berbasis web sehingga hasil analisis data dapat ditampilkan dalam bentuk visualisasi yang lebih mudah dipahami oleh perusahaan.

Dalam implementasinya, Python digunakan melalui environment pengembangan seperti Jupyter Notebook dan Visual Studio Code untuk membantu proses coding, analisis data, serta pengembangan dashboard interaktif. Dengan berbagai keunggulan tersebut, Python menjadi tools utama yang mendukung seluruh proses penelitian mulai dari pengolahan data hingga implementasi hasil analisis.

2.4.2 Microsoft Excel

Microsoft Excel merupakan salah satu software pengolah data berbasis spreadsheet yang banyak digunakan dalam berbagai bidang, termasuk administrasi bisnis, pengelolaan data, serta analisis data sederhana[43]. Excel menyediakan berbagai fitur yang memungkinkan pengguna untuk mengelola data dalam bentuk tabel, melakukan perhitungan, membuat grafik, hingga melakukan filtering dan sorting data. Karena kemudahan penggunaannya, Microsoft Excel menjadi salah satu software yang umum digunakan dalam

proses pengelolaan data perusahaan sebelum dilakukan analisis lebih lanjut menggunakan tools data science.

Dalam perusahaan distribusi dan logistik, Microsoft Excel sering digunakan untuk menyimpan dan mengelola data transaksi seperti *invoice*, data pelanggan, data produk, maupun laporan penjualan. Data *invoice* perusahaan umumnya disimpan dalam format spreadsheet sehingga memudahkan proses pencatatan dan pengelolaan data transaksi. Selain itu, Excel juga membantu pengguna dalam melakukan pemeriksaan data secara manual sebelum data diproses menggunakan software analisis data lainnya.

Dalam penelitian ini, Microsoft Excel digunakan pada tahap awal pengelolaan data untuk membantu proses pemeriksaan struktur dataset, pengecekan atribut data, serta validasi data transaksi perusahaan. Excel digunakan untuk melihat kondisi awal data *invoice* seperti jumlah data, format data, missing value, duplicate data, maupun kesesuaian atribut yang digunakan dalam penelitian. Dengan adanya proses pemeriksaan awal menggunakan Excel, peneliti dapat memahami karakteristik dataset sebelum dilakukan proses preprocessing dan *clustering* menggunakan Python.

Selain itu, Microsoft Excel juga membantu dalam proses dokumentasi dan penyusunan data sebelum diintegrasikan ke dalam proses analisis data menggunakan Jupyter Notebook dan *Streamlit*. Penggunaan Excel dalam penelitian ini mendukung proses pengelolaan data secara lebih terstruktur dan membantu memastikan bahwa data yang digunakan telah sesuai dengan kebutuhan penelitian.

2.4.3 Jupyter Notebook

Jupyter Notebook merupakan salah satu tools berbasis web yang digunakan untuk proses pengembangan dan analisis data menggunakan bahasa pemrograman seperti Python[44]. Jupyter Notebook banyak digunakan dalam bidang data science, machine learning, dan *data mining* karena menyediakan lingkungan kerja yang interaktif sehingga memudahkan proses eksplorasi data,

pengujian model, serta visualisasi hasil analisis. Selain itu, Jupyter Notebook memungkinkan pengguna untuk menjalankan kode program secara bertahap per cell sehingga proses pengembangan dan debugging menjadi lebih mudah dilakukan.

Salah satu keunggulan utama Jupyter Notebook adalah kemampuannya dalam menggabungkan kode program, visualisasi data, serta dokumentasi penjelasan dalam satu workspace yang sama. Hal tersebut membuat Jupyter Notebook sangat cocok digunakan dalam proses penelitian berbasis analisis data karena peneliti dapat langsung melihat hasil output dari setiap proses yang dijalankan. Selain itu, visualisasi seperti grafik, tabel, maupun hasil *clustering* dapat ditampilkan secara langsung sehingga mempermudah proses exploratory data analysis (EDA) dan evaluasi hasil analisis.

Dalam proses *data mining* dan machine learning, Jupyter Notebook sering digunakan untuk melakukan *preprocessing data*, eksplorasi data, implementasi algoritma, hingga pengujian model. Lingkungan kerja yang interaktif memungkinkan peneliti untuk mencoba berbagai parameter dan metode analisis secara lebih fleksibel. Oleh karena itu, Jupyter Notebook menjadi salah satu tools yang paling umum digunakan dalam penelitian *clustering* dan analisis data transaksi perusahaan.

Dalam penelitian ini, Jupyter Notebook digunakan untuk melakukan proses analisis data awal terhadap data *invoice* perusahaan. Proses seperti data preprocessing, exploratory data analysis, implementasi algoritma *clustering*, serta evaluasi *clustering* dilakukan menggunakan Jupyter Notebook. Selain itu, Jupyter Notebook juga digunakan untuk membantu proses visualisasi data dan pengujian hasil *clustering* sebelum diimplementasikan ke dalam dashboard interaktif menggunakan *Streamlit*.

Penggunaan Jupyter Notebook dalam penelitian ini membantu proses analisis data menjadi lebih terstruktur, interaktif, dan mudah dipahami. Dengan kemampuan untuk menjalankan kode secara bertahap dan menampilkan hasil

analisis secara langsung, Jupyter Notebook mendukung proses eksperimen dan pengembangan model *clustering* secara lebih efektif sehingga mempermudah proses penelitian secara keseluruhan.

2.4.4 Visual Studio Code

Visual Studio Code atau VS Code merupakan salah satu software code editor yang dikembangkan oleh Microsoft dan banyak digunakan dalam proses pengembangan perangkat lunak, analisis data, serta pengembangan aplikasi berbasis web[45]. VS Code memiliki tampilan yang ringan, fleksibel, dan mendukung berbagai bahasa pemrograman termasuk Python. Selain itu, VS Code juga menyediakan berbagai fitur seperti debugging, syntax highlighting, extension support, serta integrasi terminal yang membantu proses pengembangan aplikasi menjadi lebih efisien.

Dalam bidang data science dan machine learning, VS Code sering digunakan sebagai environment pengembangan untuk mengelola project yang memiliki struktur kode yang lebih kompleks. Berbeda dengan Jupyter Notebook yang lebih fokus pada eksplorasi data dan eksperimen, VS Code lebih cocok digunakan untuk pengembangan aplikasi atau sistem secara terstruktur. Oleh karena itu, VS Code banyak digunakan dalam proses integrasi kode program, pengelolaan file project, serta implementasi dashboard berbasis web seperti *Streamlit*.

Salah satu keunggulan VS Code adalah kemampuannya dalam mengintegrasikan berbagai tools dan *framework* dalam satu environment pengembangan. VS Code mendukung penggunaan Python serta berbagai *library* data science sehingga memudahkan proses pengembangan aplikasi analisis data. Selain itu, VS Code juga memiliki fitur terminal bawaan yang memungkinkan pengguna menjalankan program, mengelola environment Python, serta menjalankan dashboard *Streamlit* secara langsung dalam satu software.

Dalam penelitian ini, VS Code digunakan sebagai software utama dalam proses pengembangan dashboard interaktif berbasis *Streamlit*. Setelah proses *preprocessing data*, *clustering*, dan evaluasi model selesai dilakukan menggunakan Jupyter Notebook, kode program kemudian diintegrasikan dan dikembangkan lebih lanjut menggunakan VS Code. Software ini digunakan untuk menyusun struktur project, menghubungkan proses analisis data dengan dashboard interaktif, serta mengelola file program yang digunakan dalam penelitian.

Selain digunakan untuk pengembangan dashboard, VS Code juga membantu proses debugging dan pengelolaan source code penelitian secara lebih rapi dan terstruktur. Dengan fitur extension dan integrasi Python yang dimiliki, VS Code mendukung proses pengembangan aplikasi analisis data secara lebih efisien dan fleksibel. Oleh karena itu, penggunaan VS Code dalam penelitian ini membantu proses implementasi sistem menjadi lebih optimal dan mendukung penyajian hasil *clustering* dalam bentuk dashboard interaktif yang mudah digunakan oleh perusahaan.

2.4.5 *Streamlit*

Streamlit merupakan *framework* open-source berbasis Python yang digunakan untuk membangun aplikasi web interaktif khususnya dalam bidang data science, machine learning, dan data *analytics*[46]. *Streamlit* dirancang untuk mempermudah proses *deployment* dan visualisasi hasil analisis data tanpa memerlukan pengetahuan mendalam mengenai pengembangan web seperti HTML, CSS, maupun JavaScript. Dengan menggunakan *Streamlit*, pengguna dapat mengubah script Python menjadi dashboard web interaktif secara cepat dan sederhana.

Salah satu keunggulan utama *Streamlit* adalah kemampuannya dalam menampilkan visualisasi data, tabel interaktif, grafik, serta hasil machine learning secara real-time melalui web browser. *Streamlit* juga mendukung integrasi dengan berbagai *library* Python seperti Pandas, Matplotlib, Plotly,

Scikit-learn, dan NumPy sehingga sangat cocok digunakan dalam pengembangan aplikasi berbasis data *analytics* dan machine learning.

Dalam proses pengembangan aplikasi, *Streamlit* menyediakan berbagai komponen interaktif seperti sidebar, button, selectbox, filter data, upload file, serta fitur visualisasi yang memudahkan pengguna dalam melakukan eksplorasi data secara langsung. Selain itu, proses *deployment* aplikasi menggunakan *Streamlit* relatif sederhana karena aplikasi dapat dijalankan hanya dengan menggunakan perintah Python melalui terminal atau command prompt.

Dalam penelitian ini, *Streamlit* digunakan sebagai *framework* utama untuk membangun dashboard *customer* segmentation berbasis web. Dashboard tersebut digunakan untuk menampilkan hasil *clustering* pelanggan, visualisasi persebaran *cluster*, distribusi *customer*, *cluster* profiling, serta business insight dari masing-masing algoritma *clustering* yang digunakan. Dengan adanya dashboard berbasis *Streamlit*, hasil penelitian dapat divisualisasikan secara lebih interaktif sehingga memudahkan perusahaan dalam memahami hasil segmentasi pelanggan dan mendukung proses pengambilan keputusan berbasis data.

.