

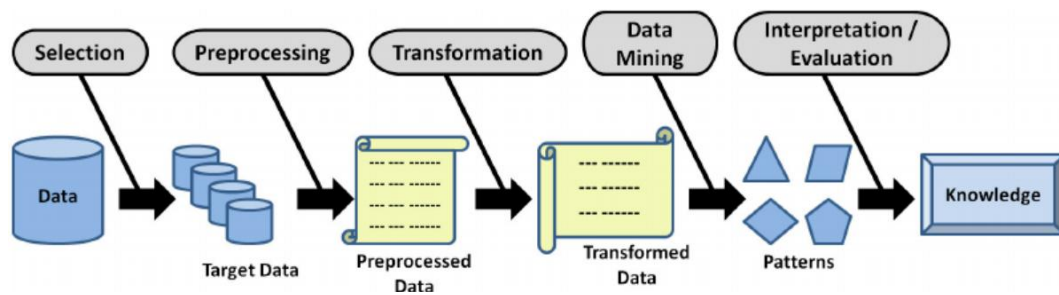
BAB II

LANDASAN TEORI

2.1. Knowledge Discovery in Database (KDD)

Knowledge Discovery in Databases (KDD) merupakan proses sistematis yang bertujuan untuk menemukan informasi atau pola baru yang bermanfaat dari data dalam basis data, khususnya data historis yang berukuran besar. (Firdaus et al., 2025). Ini mencakup tahapan pengumpulan, pemilihan, dan pengolahan data secara terstruktur guna mengidentifikasi keteraturan, pola, serta hubungan tertentu yang tersembunyi di dalam data. Melalui analisis yang sistematis, data mentah dapat diubah menjadi pengetahuan yang bernilai, sehingga hasil dari kegiatan data mining tersebut dapat dimanfaatkan sebagai dasar yang lebih akurat dan objektif dalam mendukung serta meningkatkan kualitas pengambilan keputusan di masa mendatang(Yulia et al., 2026).

Proses ini mencakup beberapa tahapan utama, yaitu seleksi data (*selection*), pra-pemrosesan atau pengolahan (*processing*), transformasi data (*transformation*), penambangan data (*data mining*), serta evaluasi atau interpretasi hasil (*interpretation/evaluation*). Setiap tahap dirancang secara sistematis untuk memastikan bahwa data yang dianalisis mampu menghasilkan informasi yang relevan dan sesuai dengan tujuan penelitian(Ermillian & Nugroho, 2024). sebagaimana ditampilkan pada Gambar 2.1



Gambar 2.1 Tahapan Dalam *Knowledge Discovery in Database (KDD)*

Tahapan Dalam Proses KDD (Ubaid et al., 2025) :

1. *Data Selection* (Seleksi Data)

Tahap awal untuk memilih data yang relevan dari database atau dataset besar sesuai dengan tujuan analisis, sehingga diperoleh subset data yang mendukung proses penemuan pola.

2. *Data Preprocessing / Cleaning* (Pra-pemrosesan / Pembersihan Data)

Proses pembersihan data dengan menghapus duplikasi, memperbaiki inkonsistensi, dan menangani nilai yang hilang agar kualitas data layak untuk dianalisis.

3. *Data Transformation* (Transformasi Data)

Tahap mengubah data ke format yang sesuai agar dapat diproses oleh algoritma data mining, seperti melalui normalisasi, reduksi dimensi, atau pengkodean variabel.

4. *Data Mining* (Penambangan Data)

Tahap inti KDD yang menerapkan algoritma seperti klasifikasi, klusterisasi, atau regresi untuk menemukan pola atau informasi penting dari data.

5. *Evaluation / Interpretation* (Evaluasi / Interpretasi)

Proses menilai dan menafsirkan pola yang dihasilkan untuk memastikan hasil analisis bermakna dan dapat digunakan dalam pengambilan keputusan

2.2. Data Mining

Data mining adalah proses untuk memperoleh informasi penting dari data dalam jumlah besar melalui analisis tertentu. Proses ini bertujuan mengekstrak pengetahuan baru agar dapat membantu pengambilan keputusan secara lebih akurat. Selain itu, data mining juga mampu menemukan pola dan tren tersembunyi yang tidak terlihat dari analisis sederhana (Sulistio et al., 2026).

Data mining tidak hanya digunakan pada bidang bisnis, tetapi juga dapat diterapkan pada sektor sosial, khususnya dalam menganalisis pemasukan/pendapatan . Melalui penerapan metode K-Means, data mining digunakan untuk mengelompokkan masyarakat berdasarkan karakteristik sosial ekonomi, seperti penghasilan, jumlah tanggungan, dan kondisi tempat tinggal. Selanjutnya, algoritma yang digunakan dalam pengolahan data mining untuk pengelompokan data pemasukan pada penelitian ini adalah metode K-Means , yang mampu membagi data ke dalam beberapa kelompok berdasarkan tingkat kemiripan karakteristik yang dimiliki (Marbun et al., 2022).

Data mining memiliki beberapa fungsi utama yang digunakan untuk memperoleh informasi atau pengetahuan baru dari data yang diolah. Fungsi-fungsi tersebut meliputi prediksi (prediction), yaitu memperkirakan nilai di masa mendatang; estimasi, yaitu menentukan nilai berdasarkan data yang tersedia;

klasifikasi (classification), yaitu mengelompokkan data ke dalam kategori tertentu; klastering (clustering), yaitu mengelompokkan data berdasarkan tingkat kemiripan karakteristik; serta asosiasi (association rule), yaitu menemukan hubungan atau keterkaitan antar data dalam suatu kumpulan data (Suwanto & Pradesan, 2026).

2.3. Pengertian Analisis

Analisis merupakan suatu proses mengkaji atau menyelidiki suatu peristiwa, data, maupun objek secara sistematis untuk memperoleh gambaran mengenai kondisi yang sebenarnya. Melalui proses analisis, informasi yang masih bersifat umum dapat diuraikan menjadi bagian-bagian yang lebih spesifik sehingga lebih mudah dipahami. Analisis dilakukan untuk menemukan hubungan antar data, mengidentifikasi permasalahan, serta menghasilkan informasi yang dapat dijadikan dasar dalam pengambilan keputusan. Dengan demikian, analisis menjadi tahapan penting dalam suatu penelitian karena membantu peneliti memperoleh kesimpulan yang akurat berdasarkan fakta dan data yang tersedia (Ismail et al., 2022).

2.3.1. Tujuan Analisis

Tujuan analisis adalah untuk memahami suatu permasalahan atau objek penelitian secara lebih mendalam melalui proses pengumpulan, penguraian, dan pengolahan data sehingga diperoleh informasi yang akurat. Analisis dilakukan untuk mengidentifikasi hubungan antar data, menemukan pola atau kecenderungan tertentu, serta menghasilkan informasi yang dapat dijadikan dasar dalam pengambilan keputusan. Dengan adanya proses analisis, data yang semula

bersifat mentah dapat diubah menjadi informasi yang lebih terstruktur, mudah dipahami, dan bermanfaat dalam menyelesaikan suatu permasalahan maupun mendukung kegiatan penelitian (Ismail et al., 2022).

2.4. Pengertian Pemasukan

pemasukan atau pendapatan adalah arus masuk manfaat ekonomi yang timbul dari aktivitas normal perusahaan selama suatu periode akuntansi yang mengakibatkan bertambahnya ekuitas, selain yang berasal dari kontribusi pemilik. Pemasukan diperoleh melalui kegiatan utama perusahaan, seperti penjualan produk atau pemberian jasa kepada pelanggan (Saputri et al., 2026).

2.4.1. Faktor-Faktor yang Mempengaruhi Pemasukan

Pemasukan suatu usaha dipengaruhi oleh beberapa faktor yang berkaitan dengan aktivitas penjualan dan kondisi pasar. Faktor-faktor tersebut antara lain (Silviana et al., 2022):

1. Volume Penjualan, semakin banyak produk yang terjual maka semakin besar pemasukan yang diperoleh.
2. Harga Jual, harga yang sesuai dengan kondisi pasar dapat meningkatkan nilai pemasukan.
3. Metode Pembayaran, transaksi tunai memberikan pemasukan penuh, sedangkan pembayaran DP hanya menghasilkan sebagian nilai transaksi.
4. Permintaan Konsumen, tingginya permintaan terhadap produk akan meningkatkan jumlah penjualan dan pemasukan.
5. Kondisi Ekonomi, daya beli masyarakat dan tingkat persaingan usaha turut memengaruhi besarnya pemasukan yang diterima perusahaan.

Dalam konteks penelitian ini, pemasukan diartikan sebagai jumlah uang yang diterima Showroom Rafare Motor Desa Lingga Tiga dari setiap transaksi penjualan sepeda motor, baik melalui pembayaran tunai maupun pembayaran uang muka (DP). Nilai pemasukan tersebut menjadi atribut utama yang digunakan sebagai dasar dalam proses pengelompokan menggunakan algoritma K-Means, sehingga dapat diketahui kategori pemasukan rendah, sedang, dan tinggi.

2.5. K-Means

K-Means merupakan metode pengelompokan data yang populer dalam bidang data mining dan machine learning. Metode ini termasuk dalam kategori non-hierarki karena tidak membentuk struktur bertingkat. K-Means bertujuan membagi data ke dalam sejumlah cluster yang telah ditentukan sebelumnya (k), sehingga objek dalam satu kelompok memiliki kemiripan yang tinggi dibandingkan dengan objek di kelompok lain.

Proses K-Means dimulai dengan penentuan centroid secara acak, kemudian setiap data dihitung jaraknya untuk menentukan kedekatan dengan centroid terdekat. Data akan dikelompokkan ke dalam cluster tersebut, lalu centroid diperbarui berdasarkan rata-rata anggota cluster. Proses ini dilakukan secara berulang hingga tidak terjadi perubahan signifikan. Karena sederhana dan efisien, K-Means banyak digunakan dalam berbagai analisis data seperti segmentasi dan pengelompokan informasi(Ikhsan et al., 2025).

2.5.1. Langkah-Langkah K-means

Berikut adalah tahapan dalam proses algoritma *K-Means* :

1. Menentukan jumlah cluster (K) yang akan dibentuk sesuai dengan kebutuhan analisis data.
2. Menetapkan centroid awal (pusat cluster) secara acak sebagai titik awal pengelompokan.
3. Menghitung jarak setiap data terhadap masing-masing centroid menggunakan metode *Euclidean Distance* sebagai berikut:

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \dots + (x_n - c_n)^2}$$

Keterangan:

d = jarak antara data dan centroid

x_i = nilai data pada atribut ke-i

c_i = nilai centroid pada atribut ke-i

n = jumlah atribut / variabel

4. Mengelompokkan setiap data ke dalam cluster yang memiliki jarak paling dekat dengan centroid.
5. Menghitung centroid baru dengan cara mencari nilai rata-rata dari seluruh data dalam masing-masing cluster.

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i$$

Keterangan:

μ_k = centroid cluster ke-k

n_k = jumlah data pada cluster ke-k

x_i = data ke-i dalam cluster

6. Mengulangi proses perhitungan jarak dan pembaruan centroid hingga tidak terjadi perubahan anggota cluster atau nilai centroid (*konvergen*).

2.5.2. Pengertian Pengelompokan (*Clustering*)

Clustering merupakan teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok (*cluster*) berdasarkan tingkat kemiripan karakteristik yang dimiliki. Data yang memiliki karakteristik yang sama akan berada dalam satu kelompok, sedangkan data yang berbeda akan berada pada kelompok yang berbeda. Clustering termasuk dalam metode unsupervised learning karena tidak menggunakan label kelas sebelumnya (Permadi, 2023). Analisis Cluster merupakan metode untuk mengelompokkan objek berdasarkan tingkat kemiripan yang diukur menggunakan jarak. Salah satu ukuran yang umum digunakan adalah jarak Euclidean, yaitu akar dari jumlah kuadrat selisih nilai antar variabel. Dalam prosesnya, jarak digunakan untuk menentukan kedekatan objek terhadap pusat cluster (*centroid*), sehingga objek akan dikelompokkan ke dalam cluster dengan jarak terdekat. Centroid diperoleh dari rata-rata anggota dalam setiap cluster dan diperbarui hingga hasil pengelompokan optimal (Firda et al., 2025).

2.6. Research Gap Penelitian

Penelitian terdahulu digunakan untuk memperkuat dasar ilmiah dan menunjukkan posisi penelitian. Kajian yang dipilih berhubungan dengan *data mining*, *K-Means*, pengelompokan data penjualan, persediaan, dan data sepeda

motor. Hasil tinjauan menunjukkan bahwa metode *K-Means* sering digunakan untuk membaca pola data usaha, tetapi objek pemasukan sepeda motor pada *showroom* lokal masih memerlukan kajian tersendiri.

Tabel 2. 1 Penelitian terdahulu yang relevan

No.	Peneliti	Objek/Fokus	Hasil Utama	Relevansi dengan Penelitian
1	Fithri & Wardhana (2021)	Transaksi penjualan Toko Usaha Mandiri	Data transaksi dapat dikelompokkan untuk membantu membaca pola penjualan.	Memberi dasar pemanfaatan K-Means pada data transaksi usaha.
2	Mauliadi (2022)	Tingkat potongan harga sepeda motor Honda	Data kendaraan dibagi menjadi tiga <i>cluster</i> untuk menentukan potongan harga.	Relevan karena memakai data sepeda motor, tetapi fokusnya diskon, bukan pemasukan unit.
3	Prastiwi dkk. (2022)	Persediaan stok barang mini market	Pengelompokan membantu menentukan persediaan stok barang.	Relevan dengan kebutuhan membaca data barang yang masuk atau tersedia.
4	Yani dkk. (2023)	Data penjualan CV Panji Medical Selaras	Sistem menghasilkan informasi produk paling laris, laris, dan kurang laris.	Mendukung penggunaan K-Means sebagai informasi keputusan usaha.
5	Syafrinal & Febrianti (2023)	Pola penjualan Zahra Mart	Aplikasi <i>data mining</i> membantu mengetahui barang laris dan tidak laris.	Relevan dengan pemanfaatan data transaksi untuk menghindari penumpukan barang.
6	Mutaqin (2023)	Penjualan saldo transportasi online	<i>Dataset</i> dibersihkan sebelum proses <i>cluster</i> dan hasilnya digunakan sebagai masukan strategi penjualan.	Menekankan pentingnya pembersihan data sebelum perhitungan.
7	Syahrani dkk. (2024)	Produk bahan bangunan	Produk dikelompokkan menjadi beberapa	Menunjukkan K-Means dapat membantu

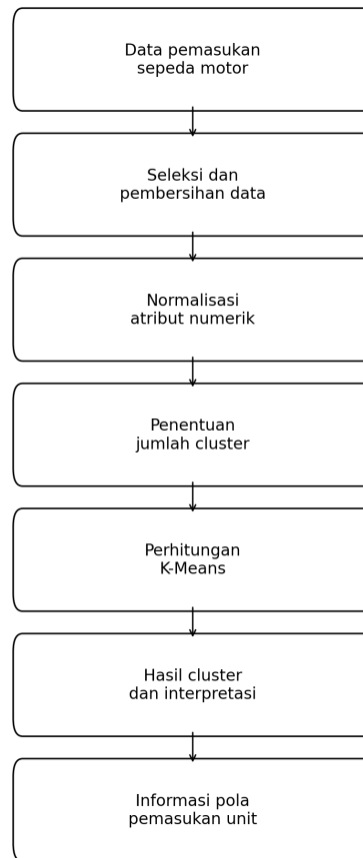
No.	Peneliti	Objek/Fokus	Hasil Utama	Relevansi dengan Penelitian
			<i>cluster</i> berdasarkan transaksi penjualan.	mencegah penumpukan produk.
8	Bui & Bahtiar (2024)	Transaksi penjualan Toko Arino	Pengelompokan transaksi digunakan untuk meningkatkan efisiensi manajemen toko ritel.	Relevan dengan pengelompokan data usaha berskala toko.
9	Wardana (2025)	Penjualan sepeda motor Honda	Data penjualan dikelompokkan menjadi tinggi, sedang, dan rendah.	Sangat relevan pada objek sepeda motor, tetapi penelitian ini berfokus pada pemasukan unit.
10	Miranda & Sriani (2025)	Data penjualan dan stok Zura Mart	Hasil <i>cluster</i> membantu optimasi persediaan dan rekomendasi promosi.	Relevan untuk memahami hubungan pengelompokan data dengan keputusan persediaan.

Tinjauan penelitian terdahulu menunjukkan kesamaan pada metode, yaitu penggunaan *K-Means* untuk mengelompokkan data usaha. Perbedaan utama penelitian ini terletak pada objek dan fokus analisis. Penelitian ini tidak mengelompokkan penjualan, promosi, atau diskon. Kajian ini mengelompokkan pemasukan sepeda motor pada *Showroom* ReFare Motor Desa Lingga Tiga berdasarkan atribut data yang tersedia.

2.7. Kerangka Pemikiran

Kerangka pemikiran penelitian ini disusun berdasarkan hubungan antara data pemasukan sepeda motor, proses praproses data, penerapan algoritma *K-Means*, dan hasil analisis *cluster*. Data pemasukan menjadi bahan utama.

Algoritma menjadi alat analisis. Hasil *cluster* menjadi informasi yang menjelaskan pola pemasukan sepeda motor.



Gambar 2.2 Alur Penerapan K-Means

Proses dimulai dari data pemasukan sepeda motor yang diperoleh dari *showroom*. Data kemudian dipilih, dibersihkan, dan disesuaikan skalanya. Setelah jumlah *cluster* ditentukan, algoritma *K-Means* menghitung jarak data terhadap *centroid* sampai kelompok data stabil. Hasil akhir berupa informasi pola pemasukan unit yang dapat dibaca oleh pihak *showroom*.