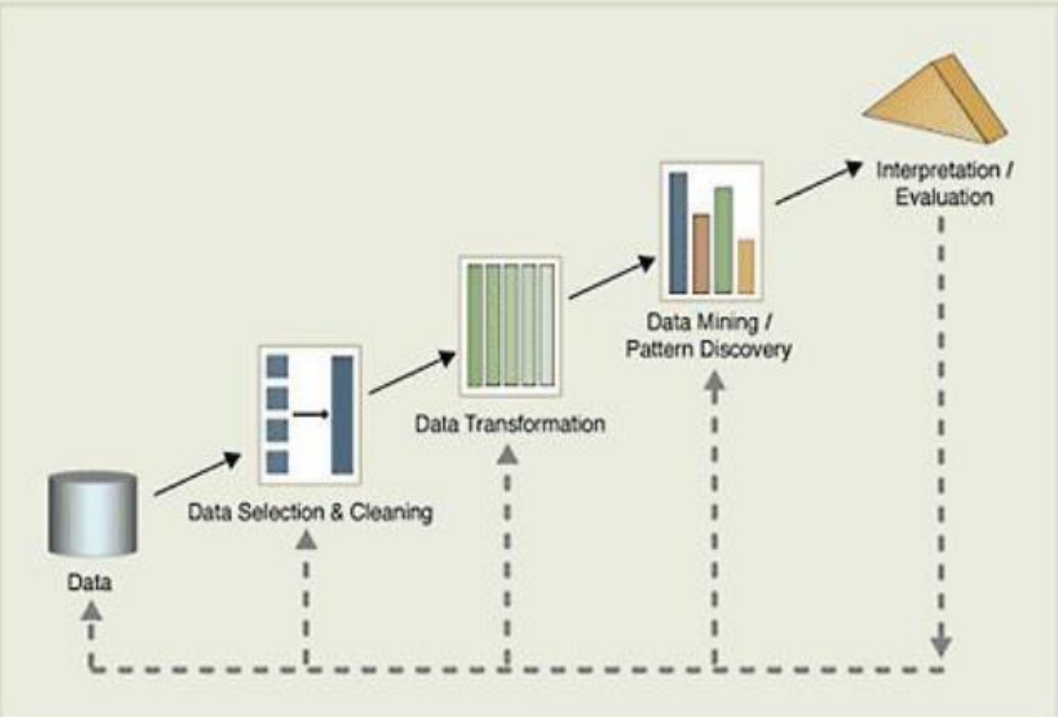
BAB II
LANDASAN TEORI

2.1 *Knowledge Discovery in Database*

KDD adalah suatu proses sistematis dalam mengeksplorasi data, yang terdiri dari lima tahapan utama: *Selection*, *Preprocessing*, *Transformation*, *Data Mining*, dan *Interpretation* (Oktavian et al., 2025). Istilah data mining dan *knowledge discovery in database* (KDD) sering kali digunakan secar bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, teteapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD secara garis besar dapat dijelaskan sebagi berikut:



Gambar 2. 1 *Knowledge Discovery in Database* (KDD)

Sumber: (Prasetya et al., 2021)

Proses dalam KDD adalah proses yang terdiri dari rangkaian proses iteratif sebagai berikut:

a. *Data selection*

Pemilihan (seleksi) data dari sekumpulan data oprasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk porses data mining, disimpan dalam suatu berkas, terpisah dari basis data oprasional.

b. *Pre-processing / Cleaning*

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjaadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*). Juga dilakukan proses enrichmen, yaitu proses "memperkaya" data yang sudah ada denagan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

c. *Transformation*

Coding adalah proses transformasi pada data yang taleh dipilih, sehingga data tersebut sesuai proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantunng pada jenis atau pola informasi yang akan dicari dalam basis data.

d. *Data Mining / Pattern Discovery*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknk ataumetode tertentu. Teknik, metode, atau

algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

e. *Interpretation / Evaliation*

Pola informasi yang dihasilkan dari proses dat mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan proses dari KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.2 Data Mining

Data Mining adalah proses ekstraksi informasi dari kumpulan data melalui penggunaan algoritma dan teknik yang melibatkan bidang ilmu statistik, mesin pembelajaran, dan sistem manajemen database (Saifudin et al., 2024). Data Mining digunakan untuk ekstraksi informasi penting yang tersembunyi dari dataset yang besar. Data mining dibagai menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

a. Deskripsi

Terkdang penelitian dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data.

b. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variable target estimasi lebih kearah numerik dari pada kearah kategori. Model dibangun menggunakan

record lengkap yang menyediakan nilai dari varibel target sebagai nilai prediksi.

c. Prediksi

Predeksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa motode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaaan yang tepat) untuk prediksi.

d. Klasifikasi

Dalma klasifikasi, terdapat target varibel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

e. Pengklusteran

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidamiripan denagn *record* dalam kluster lain. Dari algoritma yang telah dijaleaskan algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan *record* dalam suatu kelompok lain akan bernilai minimal.

f. Asosiasi

Tugas asosiasi dalam data mining adalah menemukan atribut uang muncul dalam suatu waktu.

2.3 Algoritma *K-Means*

Algoritma *K-Means* kelompokkan data ke dalam kluster berbeda sehingga data yang serupa berakhir di kelompok yang sama dan data yang berbeda berada di kelompok terpisah (Irsandi et al., 2025). Lebih spesifiknya, algoritma *k-means* bekerja seperti ini:

a. Pilih *K* sebagai jumlah grup yang dibuat.

b. Pilih posisi acak untuk titik awal setiap grup (disebut sentroid).

c. Hitung jarak terdekat setiap titik data dan setiap sentroid dengan rumus jarak *Euclidean*. Rumus untuk jarak *Euclidean* adalah:

$$d\ (x_i, \mu j) = \sqrt{\sum (x_i - \ \mu j)\ ^2}$$

Keterangan:

xi: kriteria

μj: sentroid klaster ke-j

d. Klasifikasikan setiap titik data berdasarkan jaraknya ke pusat massa, pilih titik dengan jarak terpendek. Kemudian, perbarui nilai pusat massa. Pusat massa yang baru dihitung dengan mengambil rata-rata semua titik data dalam klaster tersebut, menggunakan rumus yang diberikan:

$$C_k = \frac{\sum d_i}{n_k}$$

Keterangan:

nk : Jumlah data klaster *k*

di : Jumlah dari nilai jarak yang masuk dalam klaster.

e. Ulangi langkah 2 - 5 hingga tidak ada lagi perubahan pada setiap kelompok.

f. Setelah Anda menyelesaikan langkah terakhir, gunakan nilai pusat kelompok (μj) dari putaran terakhir untuk membantu mengklasifikasikan data.

Memahami Pengelompokan *K-Means*, K adalah angka yang menunjukkan berapa banyak kelompok yang ingin kita bentuk. Istilah "Means" mengacu pada rata-rata sekelompok titik data, dan setiap kelompok disebut sebagai klaster. Pengelompokan adalah teknik yang digunakan untuk mengelompokkan data dan merupakan bagian dari analisis data atau penambangan data. Metode ini tidak memerlukan data berlabel, sehingga dikenal sebagai pembelajaran tanpa pengawasan. Algoritma K-Means digunakan untuk membagi data menjadi beberapa kelompok, di mana setiap kelompok memiliki karakteristiknya sendiri yang berbeda tetapi juga memiliki beberapa kesamaan dengan kelompok lain:

a. Cari nilai *k*, yang menunjukkan berapa banyak grup atau klaster yang perlu kita buat.
b. Mulailah dengan memilih k titik acak sebagai pusat awal klaster.
c. Ukur jarak setiap titik data dari setiap pusat menggunakan rumus Jarak *Euclidean*:

$$\mathrm{d}\,(\mathrm{P},\mathrm{Q}) \;=\; \sqrt{\sum_{j=1}^{p} \left(x_j(P) - x_j\,(Q)\right)^2}$$

d. Masukkan setiap titik data ke dalam kelompok yang paling dekat dengan pusat klasternya.
e. Temukan lokasi baru pusat klaster (k).
f. Kembali ke langkah 3 jika pusat klaster yang baru berbeda dari yang lama.

Pengelompokan *K-Means* adalah teknik non-hierarkis untuk pengelompokan data. Algoritma ini membagi data menjadi satu atau beberapa kelompok, menempatkan data yang serupa dalam kelompok yang sama dan data yang berbeda dalam kelompok terpisah. Metode ini populer karena mudah digunakan. Namun, salah satu masalah dengan metode ini adalah cara pembentukan klaster sangat bergantung pada bagaimana titik awal dipilih.

2.4 Algoritma Apriori

Algoritma apriori merupakan algoritma klasik yang digunakan agar komputer mempelajari aturan asosisasi dan mecari pola hubungan antar item dataset (Prasetya et al., 2021). Algoritma Apriori adalah algoritma yang digunakan dalam data mining untuk menemukan aturan asosiasi antara item dalam kumpulan data transaksi (Suseno, 2024).

Berdasarkan pendapat tersebut, dapat disimpulkan bahwa algoritma Apriori berperan penting dalam proses penggalian data (data mining), khususnya dalam mengidentifikasi pola keterkaitan dan hubungan antar item yang sering muncul secara bersamaan dalam suatu dataset transaksi. Algoritma ini membantu peneliti dalam menghasilkan aturan asosiasi yang bersifat informatif, sehingga dapat digunakan sebagai dasar pengambilan keputusan dan analisis perilaku data secara sistematis dan terstruktur.

Aturan asosiasi ini mengidentifikasi hubungan antara item-item yang sering muncul bersama dalam transaksi. Algoritma ini memanfaatkan konsep "prinsip apriori", yang menyatakan bahwa jika sebuah itemset jarang terjadi, maka subsetnya juga jarang terjadi.

Algoritma apriori memiliki 2 syarat dalam penggunaannya yaitu harus menggunakan nilai minimum support dan nilai minimum confidence. nilai minimum support menunjukkan persentase atau jumlah minimum transaksi di mana sebuah aturan harus terjadi agar dianggap relevan. Sedangkan nilai minimum confidence menentukan persentase minimum transaksi di mana item yang terkait harus terjadi bersama-sama dalam transaksi agar aturan dianggap relevan atau layak untuk dianalisis lebih lanjut.

Nilai confidence dihitung sebagai persentase transaksi yang mengandung item A dan B (contoh: produk A dan B dibeli bersamaan) dibagi dengan persentase transaksi yang mengandung item A (contoh: produk A dibeli). Parameter Pada penelitian ini nilai minimum support yang digunakan yaitu 30% dan nilai minimum *confidence* yang digunakan sebesar 60%.

Tahap ini bertujuan untuk menemukan kombinasi item yang memenuhi syarat minimum nilai support dalam basis data. Nilai support suatu item dihitung menggunakan rumus berikut.

$$Support\ (A) = \frac{\text{Jumlah transaksi mengandung A}}{\text{Total Transaksi}}$$

Selanjutnya untuk mencari nilai suuport dari gabungan 2 item dengan rumus berikut.

$$Support\ (A,B) = \frac{\text{Jumlah transaksi mengandung A dan B}}{\text{Total Transaksi}}$$

Setelah semua pola frekuensi tinggi ditemukan, langkah berikutnya adalah mencari aturan asosiasi yang memenuhi syarat minimum untuk confidence dengan menghitung confidence aturan asosiatif A → B. Nilai *confidence* dari aturan A → B dihitung menggunakan rumus berikut (Syaifullo et al., 2024).

$$Confidence\ P(A/B) = \frac{\text{Jumlah transaksi mengandung A dan B}}{\text{Total Transaksi}}$$

Algoritma ini mampu mengidentifikasi kombinasi item yang sering muncul bersamaan dan membentuk aturan asosiasiyang relevan, sehingga penerapannya luas pada market basket analysis, khususnya dalam mendukung strategi pemasaran dan pengelolaan produk. Beberapa istilah penting dalam algoritma Apriori antara lain:

a. *Itemset*

   Itemset adalah sekumpulan produk dalam sebuah transaksi, yang bisa mencakup satu ataupun beberapa item untuk dianalisis.

b. *Support*

   Frekuensi kemunculan suatu itemset dalam seluruh transaksi, digunakan untuk menilai kelayakan analisislebih lanjut.

c. *Minimum Support*

   Ambang batas minimum supportyang harus dipenuhi agar itemset dapat dikategorikan sebagai frequent itemset.

d. *Kandidat Itemset (K-Item)*

   Itemset awal hasil proses join yang akan diuji terhadap nilai minimum support.

e. *Frequent Itemset*

   Itemset yang memenuhi syarat support minimum dan digunakan sebagai acuan dalam membentuk aturan asosiasi.

f. *Pruning*

   Proses penyaringan kandidat itemset yang tidak memenuhi minimum supportuntuk meningkatkan efisiensi analisis.

g. *Join*

Penggabungan dua atau lebih itemset berbeda untuk membentuk kandidat itemset dengan panjang lebih besar.

h. *Confidence*

Tingkat kepercayaan atau probabilitas bahwa item B akan dibeli jika item A dibeli, digunakan untuk mengukur kekuatan aturan asosiasi.

Setelah tahap pemahaman data, dilakukan proses preprocessing untuk menyiapkan data setagar sesuai dengan kebutuhan analisis menggunakan algoritma Apriori. Proses ini mencakup beberapa langkah teknis utama, yaitu:

a. Data *Cleaning*

Data *cleaning* merupakan tahap awal yang bertujuan untuk memperbaiki dan membersihkan data dari berbagai permasalahan yang dapat mengganggu proses analisis. Pada tahap ini dilakukan penghapusan data yang tidak lengkap (*missing value*), data ganda (*duplicate*), serta data yang tidak konsisten atau mengandung kesalahan pencatatan. Selain itu, data yang bersifat tidak relevan dengan tujuan penelitian juga dihilangkan. Proses data *cleaning* sangat penting karena kualitas data yang buruk dapat menghasilkan pola dan aturan yang tidak akurat pada tahap analisis selanjutnya.

b. Data *Reduction*

Data *reduction* adalah proses penyederhanaan data dengan cara mengurangi jumlah data tanpa menghilangkan informasi penting yang terkandung di dalamnya. Tahapan ini dilakukan untuk meningkatkan efisiensi proses pengolahan data dan mengurangi kompleksitas komputasi. Data *reduction*

dapat dilakukan melalui pemilihan atribut yang relevan, pengelompokan data, atau penghapusan atribut yang memiliki pengaruh rendah terhadap tujuan penelitian. Dengan melakukan data *reduction*, dataset menjadi lebih ringkas namun tetap representatif, sehingga proses analisis dapat berjalan lebih cepat dan efektif.

c. *Data Transformation*

Data *transformation* merupakan tahap mengubah data ke dalam format atau struktur yang sesuai dengan kebutuhan algoritma yang digunakan. Transformasi data dapat berupa normalisasi, pengkodean data kategorikal menjadi numerik, pengelompokan nilai, maupun penyesuaian skala data. Pada tahap ini, data juga dapat disesuaikan agar memenuhi syarat input algoritma data mining, sehingga proses pembentukan pola dan aturan asosiasi dapat dilakukan secara optimal. Data *transformation* memastikan bahwa data siap diproses dan mampu menghasilkan informasi yang bermakna.

2.5 Tools Pendukung

Tools pendukung merupakan perangkat dan sarana yang digunakan untuk membantu peneliti dalam melaksanakan seluruh tahapan penelitian, mulai dari pengumpulan data, pengolahan data, analisis, hingga penyajian hasil penelitian. Dalam penelitian ini, tools pendukung berperan penting dalam memastikan proses analisis data transaksi penjualan dapat dilakukan secara sistematis, akurat, dan efisien.

Penggunaan tools pendukung memungkinkan peneliti untuk mengolah data dalam jumlah besar, menerapkan algoritma K-Means dan Apriori secara tepat, serta

memvisualisasikan hasil analisis sehingga lebih mudah dipahami. Tools pendukung yang digunakan dalam penelitian ini meliputi perangkat keras (*hardware*) dan perangkat lunak (*software*) yang saling terintegrasi untuk mendukung keberhasilan penelitian.

2.6.1 Perangkat Keras (*Hardware*)

Perangkat keras personal komputer merupakan seluruh bagian fisik personal komputer, perbedaannya terletak dalam data yg berada pada dalamnya atau berjalan pada atasnya dan tidak selaras menggunakan software yg menaruh instruksi pada perangkat keras buat menuntaskan tugasnya (Simanullang, 2021). Perangkat keras yang digunakan dalam penelitian ini berfungsi sebagai media utama dalam pengolahan data transaksi Cafe Lega serta penerapan algoritma data mining.

Adapun spesifikasi perangkat keras yang digunakan dalam penelitian ini antara lain:

a. Laptop atau komputer dengan prosesor minimal Intel Core i3 atau setara
b. Memori (RAM) minimal 4 GB
c. Media penyimpanan berupa hard disk atau SSD untuk menyimpan data transaksi dan hasil analisis
d. Perangkat input seperti keyboard dan mouse untuk mendukung proses pengolahan data
e. Perangkat output seperti monitor untuk menampilkan hasil analisis dan visualisasi data

Spesifikasi perangkat keras tersebut dinilai telah memadai untuk menjalankan proses pengolahan data dan analisis menggunakan algoritma K-Means dan Apriori tanpa kendala berarti.

2.6.2 Perangkat Lunak (*Software*)

Perangkat lunak merupakan sarana penting untuk memastikan keamanan perangkat lunak (Siregar, 2020). Maka dapat diartikan perangkat lunak (*software*) merupakan aplikasi atau program yang digunakan untuk mengolah data, melakukan analisis, serta menyajikan hasil penelitian. Dalam penelitian ini, perangkat lunak memiliki peran penting dalam mendukung implementasi metode data mining.

Perangkat lunak yang digunakan dalam penelitian ini meliputi:

a. Sistem Operasi Windows, sebagai sistem operasi utama untuk menjalankan seluruh aplikasi pendukung penelitian

Gambar 2. 2 Tampilan awal windows 10

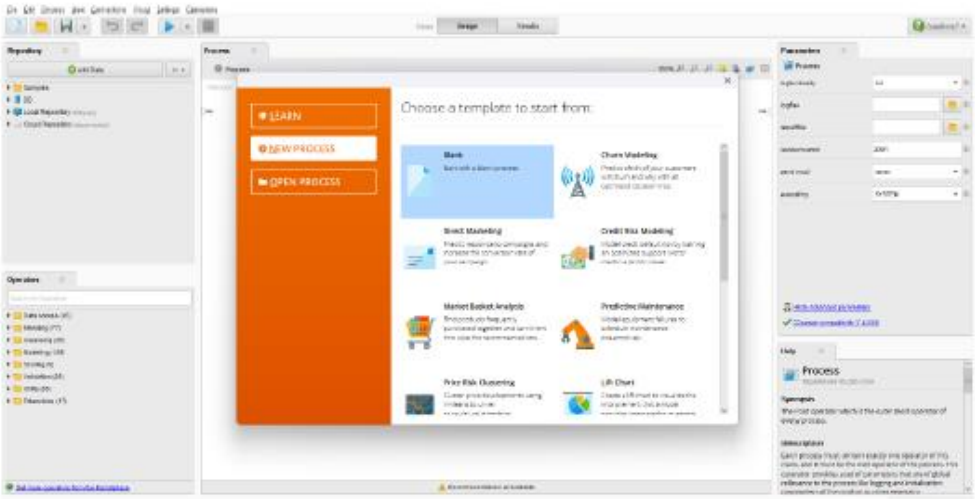
Sumber: (Dalimunthe et al., 2020)

b. Microsoft Excel, digunakan untuk proses input data, pembersihan data (data cleaning), serta pengolahan data awal

Gambar 2. 3 Lembar kerja MS.Excel

Sumber: (Suryati et al., 2020)

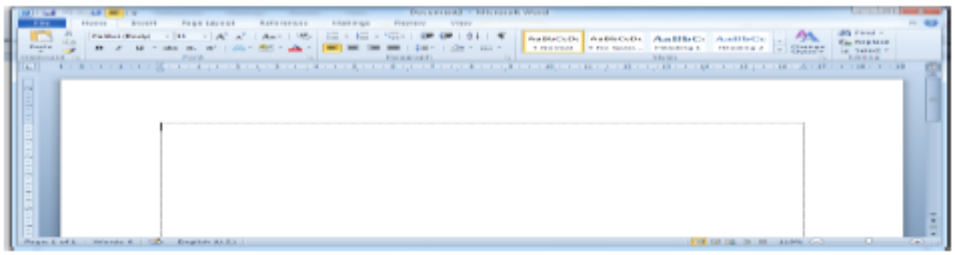
c. RapidMiner, digunakan untuk menerapkan algoritma K-Means dan Apriori dalam proses analisis data mining



Gambar 2. 4 Tampilan awal Rapidminer

Sumber: (Hartono & Albertus, 2024)

d. Microsoft Word, digunakan untuk penyusunan laporan hasil penelitian dan dokumentasi



Gambar 2. 5 Tampilan Microsoft Word 2010

Sumber: (Nasution & Putri, 2021)

Penggunaan perangkat lunak tersebut memungkinkan proses analisis data dilakukan secara terstruktur, akurat, dan sesuai dengan kebutuhan penelitian.

2.6.3 Peran Tools Pendukung dalam Penelitian

Tools pendukung memiliki peran yang sangat penting dalam menunjang keberhasilan penelitian ini. Tanpa adanya tools pendukung yang memadai, proses pengolahan dan analisis data transaksi penjualan akan sulit dilakukan secara optimal.

Peran tools pendukung dalam penelitian ini antara lain:

a. Membantu pengolahan data transaksi agar siap dianalisis melalui proses data *cleaning*, data *reduction*, dan data *transformation*
b. Mempermudah penerapan algoritma K-Means dan Apriori secara sistematis dan terukur
c. Meningkatkan akurasi dan efisiensi analisis, terutama dalam menangani data transaksi dalam jumlah besar
d. Menyajikan hasil analisis dalam bentuk yang mudah dipahami, baik berupa tabel, grafik, maupun aturan asosiasi

e. Mendukung dokumentasi dan pelaporan hasil penelitian secara rapi dan terstruktur

Dengan dukungan perangkat keras dan perangkat lunak yang sesuai, penelitian ini diharapkan mampu menghasilkan analisis pola pembelian pelanggan yang akurat dan dapat dijadikan dasar pengambilan keputusan strategis bagi pihak Cafe Lega.

2.7 Penelitian Terdahulu

Penelitian terdahulu digunakan sebagai landasan dan pembanding dalam penelitian ini, guna menunjukkan posisi penelitian yang dilakukan serta memperkuat relevansi penggunaan metode yang dipilih. Beberapa penelitian sebelumnya telah menerapkan teknik data mining, khususnya algoritma K-Means dan Apriori, dalam menganalisis pola pembelian dan perilaku konsumen pada berbagai bidang usaha. Berikut ringkasan penelitian terdahulu yang relevan dengan penelitian ini.

Tabel 2. 1 Penelitian Terdahulu

No	Penulis & Tahun	Objek Penelitian	Metode / Algoritma	Teknik Analisis Data	Hasil Penelitian
1	Prasetya dkk. (2021)	Data transaksi penjualan ritel	Algoritma Apriori	Association Rule Mining	Penelitian berhasil menemukan pola keterkaitan antar produk yang sering dibeli bersamaan, sehingga dapat digunakan untuk mendukung strategi penataan

					produk dan promosi penjualan.
2	Suseno (2024)	Data penjualan produk	K-Means dan Apriori	Clustering dan Association Rules	Hasil penelitian menunjukkan bahwa kombinasi K-Means dan Apriori mampu mengelompokkan transaksi serta menemukan hubungan antar produk yang signifikan untuk mendukung pengambilan keputusan bisnis.
3	Oktavian dkk. (2025)	Data penjualan produk ritel	K-Means dan Apriori	Clustering dan Market Basket Analysis	Penelitian menghasilkan segmentasi pelanggan berdasarkan pola pembelian serta rekomendasi strategi product bundling yang lebih efektif.
4	Strategi Pemasaran dkk. (2025)	Platform Edulink.id	K-Means dan Apriori	Analisis klaster dan aturan asosiasi	Hasil penelitian menunjukkan bahwa penerapan K-Means mampu mengelompokkan pelanggan secara optimal, sedangkan Apriori mengidentifikasi kombinasi produk/layanan yang sering digunakan bersama.

5	Arif Saifudin dkk. (2024)	Marketplace Shopee	K-Means dan FP-Growth	Clustering dan Association Mining	Penelitian membuktikan bahwa pengelompokan konsumen dan pencarian pola pembelian dapat meningkatkan pemahaman perilaku pelanggan pada marketplace.
6	Faris Syaifulloh dkk. (2024)	Toko bangunan	Apriori, FP-Growth, Squeezer	Analisis pola pembelian	Hasil penelitian menunjukkan bahwa algoritma Apriori mampu menghasilkan aturan asosiasi yang kuat dan informatif terkait pola pembelian pelanggan.

Sumber: Penelitian 2025

.