BAB II
LANDASAN TEORI

2.1 *Data Mining*

Data mining telah menarik perhatian yang signifikan dalam beberapa tahun terakhir di sektor informasi dan masyarakat pada umumnya. Hal ini disebabkan oleh banyaknya data yang tersedia dalam skala besar dan kebutuhan mendesak untuk mengubah data tersebut menjadi informasi serta pengetahuan yang bermanfaat. Informasi dan pengetahuan yang diperoleh dari proses tersebut dapat diterapkan dalam berbagai konteks, mulai dari analisis pasar, pencegahan penipuan, hingga retensi pelanggan, pengendalian produksi, dan ilmu pengetahuan secara umum. Dalam perspektif evolusi teknologi informasi, data mining dapat dianggap sebagai hasil alami dari perkembangan tersebut (Marisa, 2005) .

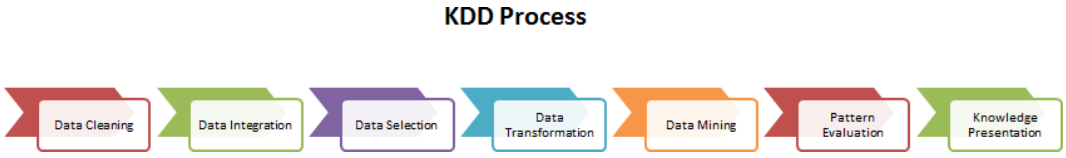
Data mining merupakan proses yang memanfaatkan berbagai teknik statistik, matematika, kecerdasan buatan, dan machine learning guna mengekstraksi serta mengidentifikasi informasi yang bernilai dan pengetahuan terkait dari kumpulan data besar yang ada di berbagai database (Gustientiedina et al., 2019).

Menurut (Campbell, 2021) *Data mining* yakni proses mengidentifikasi dan mengekstrak pola, variabel, dan tren tersembunyi dalam kumpulan data apa pun yang dikumpulkan untuk di analisis. Dengan kata sederhana, proses melihat data untuk mengidentifikasi pola tersembunyi dan tren informasi yang dapat digunakan untuk mengkategorikan data menjadi analisis yang berguna disebut *data mining* atau knowledge discovery of data (KDD). Anda dapat mempergunakan penambangan data guna mengubah data atau informasi mentah menjadi data, yang dapat digunakan oleh bisnis.

2.2 Operasi *Data Mining*

Menurut (Navlani et al., 2021) *Data mining* disebut sebagai proses penemuan pengetahuan dari pola-pola yang menarik. Tujuan utama dari Knowledge

Discoveryin Databases (KDD) adalah untuk mengekstrak atau menemukan pola menarik yang tersembunyi dari database besar, data warehouse, dan web serta penyimpanan informasi lainnya.



Gambar 2.1 Proses KDD

Tahapan proses pada penggunaan *data mining* yang merupakan proses knowledge discovery in database (KDD) misalnya yang tampak pada Gambar 2.2 mampu diuraikan sebagaimana berikut :

1. Pembersihan Data: Pada tahap pertama ini, data diproses terlebih dahulu.Di sini, noise dihilangkan, nilai yang hilang ditangani, dan outlier dideteksi.
2. Integrasi Data: Pada fase ini, data dari berbagai sumber digabungkan dan terintegrasi bersama menggunakan migrasi data dan alat ETL.
3. Pemilihan Data: Pada fase ini, data yang relevan untuk tugas analisis dikumpulkan kembali.
4. Transformasi Data: Pada fase ini, data direkayasa sesuai kebutuhan bentuk yang tepat untuk analisis.
5. *Data mining*: Pada fase ini, teknik *data mining* dipergunakan gunamenemukan hal-hal yang berguna dan pola yang tidak diketahui.
6. Evaluasi Pola: Pada fase ini, pola yang diekstraksi dievaluasi.
7. Presentasi Pengetahuan: Setelah evaluasi pola, pengetahuan yang diekstraksi perlu divisualisasikan dan disajikan kepada pebisnis untuk pengambilan keputusan tujuan.

2.3 *Clustering*

Menurut (Madyatmadja et al., 2021) Clustering adalah suatu teknik yang digunakan untuk memisahkan objek-objek ke dalam kelompok-kelompok berdasarkan kesamaan atribut atau karakteristik dengan data-data lain. Clustering merupakan suatu pendekatan dalam data mining di mana algoritma yang digunakan bersifat

tanpa supervisi, yang berarti bahwa metode ini tidak memerlukan pelatihan khusus dan tidak memerlukan panduan, bahkan keluaran tidak perlu.

Menurut (Milenković et al., 2020) Analisis klaster yakni tugas mengkalsifikasikan sekumpulan objek sedemikian rupa sehingga objek-objek pada klaster tersebut serupa satu sama lain dan dengan demikian berbeda dari objek-objek dalam klaster lainnya. Tidak seperti klasifikasi, kita tidak memiliki solusi "tepat" di sini:

- Evaluasi kinerja algoritma jauh lebih sulit daripada klasifikasi.
- Kesesuaian solusi bergantung pada domain dan kasus aplikasi.
- Satu solusi yang sama dapat dievaluasi secara berbeda dalam kasus aplikasi yang berbeda.
- Memerlukan keterlibatan pakar domain untuk mengevaluasi solusi.

Proses analisis klaster terdiri dari dua langkah dasar:

1) pemilihan ukuran jarak (kesamaan) yang tepat,
2) pemilihan algoritma pengelompokan, yaitu serangkaian prosedur untuk mengelompokkan elemen sehingga ada perbedaan kecil dalam klaster dan perbedaan besar antar klaster. Ada berbagai algoritma untuk memecahkan masalah pengelompokan. Namun, tidak ada algoritma terbaik secara objektif untuk pengelompokan, karena algoritma tertentu dapat menghasilkan hasil yang baik pada satu set data dan buruk pada yang lain karena pengelompokan bergantung pada dimensionalitas, struktur, dan jenis data. Terdapat metode hierarkis dan non-hierarkis, termasuk metode k-mean, yang menjadi subjek penelitian kami. Metode sub-kluster non-hierarkis, yang lebih andal daripada metode hierarkis, mengasumsikan bahwa jumlah kluster diketahui sebelumnya, atau seperti beberapa metode, bervariasi selama proses pengelompokan.

2.4 Algoritma *K-means*

Menurut penelitian sebelumnya (Gustientiedina et al., 2019) disimpulkan bahwasanya Algoritma K-means Clustering yakni suatu metode pengelompokkan iteratif yang membagi suatu himpunan data menjadi sejumlah K cluster sesuai dengan yang telah ditetapkan sebelumnya. Algoritma K-means Clustering terbukti mudah dalam implementasinya, beroperasi secara efisien, dapat dengan cepat beradaptasi, dan umumnya sering digunakan dalam berbagai aplikasi. Kelompok data yang sama akan ditempatkan dalam satu cluster, sedangkan data dengan karakteristik yang berbeda akan dielompokkan ke dalam cluster yang berbeda pula. Sehingga, variansi data dalam satu cluster akan cenderung kecil. Dalam pengelompokkan, kedekatan antara dua objek diukur berdasarkan jarak antara keduanya. Begitu pula, kedekatan antara suatu data dengan pusat cluster ditentukan oleh jarak dari data tersebut ke pusat cluster. Data akan termasuk ke dalam cluster yang tepat berdasarkan jarak terdekat antara data tersebut dengan pusat cluster yang bersangkutan.

Pengukuran jarak antara setiap data dengan masing-masing pusat klaster dihitung dengan memanfaatkan formula jarak euclidea yang dirumuskan sebagaimana berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{r} (x_i - y_i)^2} \text{..(1)}$$

Dimana :

$d(x,y)$ = Jarak data $x$ke y
x_i = Nilai fitur ke-$i$dari x
y_i = Nilai fitur ke-i dari y
r = Jumlah fitur dalam vektor

Sebuah rekaman bakal dimasukkan ke dalam sebuah cluster tertentu apabila jarak antara rekaman tersebut dengan pusat cluster tersebut adalah yang terendah dibandingkan dengan jarak ke pusat cluster lainnya. Kemudian, data-data yang termasuk anggota dalam setiap cluster dikelompokkan. Pusat cluster yang baru akan diperoleh dengan menghitung rata-rata nilai dari setiap fitur dari seluruh data yang ada dalam masing-masing cluster. Pusat dari semua data dalam cluster diwakili oleh centroid, disebutkan dengan:

$$c_j = \frac{1}{Nk} \sum_{l=1}^{Nk} x_{jl}$$

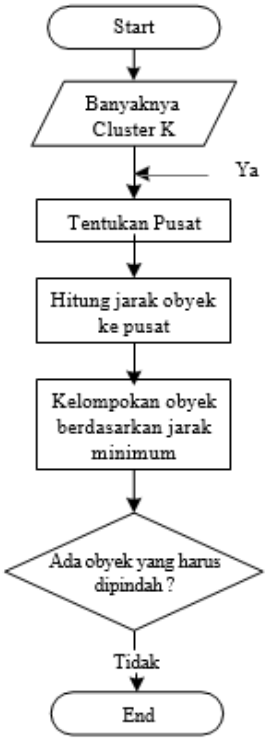
Dimana :

c_j = Centroid baru
x_{jl} = Anggota cluster *l*pada atribut *ke-j*
Nk = Jumlah data dalam cluster

X_{jl} merupakan nilai dari anggota cluster l pada atribut ke-j setelah diukur menggunakan metode perhitungan jarak euclidean. Sementara itu, Nk menunjukkan jumlah data yang termasuk dalam suatu cluster setelah dihitung menggunakan teori jarak euclidean. Langkah saat melangsungkan Algoritma *K- means Clustering* yakni sebagaimana berikut :

1. Tentukan nilai k sebagai jumlah cluster yang harus dibuat.
2. Pilih k dari dataset x sebagai titik tengah (centroid) awal.
3. Alokasikan setiap data ke centroid terdekat menggunakan perhitungan jarak berdasarkan metode euclidean.
4. Hitung ulang posisi centroid c berdasarkan data yang tergabung dalam masing-masing cluster dengan mengaplikasikan rumus yang menentukan centroid baru.
5. Lakukan iterasi langkah 3 dan 4 secara berulang hingga mencapai kondisi konvergensi, di mana (a) perubahan dalam fungsi objektif telah turun di bawah ambang batas yang telah ditentukan, ataupun (b) tidak ada data yang beralih cluster lagi, atau perubahan lokasi centroid telah turun di bawah ambang batas yang telah ditetapkan.

Alur Algortima *K-means Clustering* diperlihatkan pada gambar dibawah ini :



Gambar 2.3 *Flowchart* Algoritma *K-means*

Menurut penelitian sebelumnya (Alfina & Santosa, 2012), mengungkapkan bahwasanya k-means adalah salah satu metode clustering yang paling sederhana dan umum digunakan. Keunggulan utama dari metode k-means adalah kemampuannya dalam mengelompokkan data dengan efisien dan cepat, bahkan untuk jumlah data yang besar. Meskipun demikian, metode k-means memiliki kelemahan yang disebabkan oleh proses penentuan pusat awal cluster. Ketepatan hasil clustering yang dihasilkan oleh k-means sangat tergantung pada nilai awal pusat cluster yang dipilih. Karena itu, kemungkinan hasil akhir cluster yang terbentuk merupakan solusi lokal optimal.

2.5 *Python*

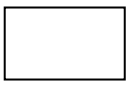
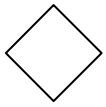
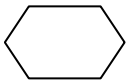
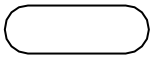
Python sudah berkembang menjadi bahasa yang umum digunakan dalam berbagai aplikasi ilmu data. Ini menggabungkan kemampuan bahasa pemrograman umum dengan keunggulan bahasa skrip khusus di bidang seperti MATLAB atau R. Python dilengkapi dengan berbagai perpustakaan yang mencakup pemrosesan data, visualisasi, statistik, pemrosesan bahasa alami, dan pemrosesan gambar. Dengan

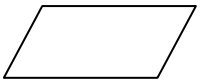
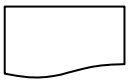
berbagai alat yang luas ini, para ilmuwan data memiliki kemungkinan untuk memilih dari sekumpulan fungsi umum dan tujuan khusus. Salah satu keunggulan utama menggunakan Python adalah kemampuannya untuk berinteraksi langsung dengan kode, yang dapat dilakukan melalui terminal atau alat seperti Jupyter Notebook. Dalam pembelajaran mesin dan analisis data yang melibatkan proses berulang secara mental, interaksi yang cepat dan mudah melalui alat yang mendukung iterasi merupakan hal penting. Python juga mendukung pembuatan antarmuka pengguna grafis (GUI) dan pengembangan layanan web yang kompleks, serta integrasi ke dalam sistem yang sudah ada (N a v l a n i e t a l . , 2 0 2 1) .

2.6 Diagram *Flowchart*

(Rosa A.S. M. Salahudin, 2019) Flowchart merupakan sebuah diagram yang menggunakan simbol-simbol grafis untuk menggambarkan aliran dari sebuah algoritma atau proses tertentu, dimana tindakan-tindakan simbolis direpresentasikan dalam bentuk kotak dan diurutkan dengan menyambungkan setiap langkah menggunakan panah.

Tabel 2.1 Simbol –Simbol *Flowchart*

	*Process*	Simbol yang memperlihatkan pengolahan yang dilangsungkan Komputer
	*Decision*	Simbol guna kondisi yang bakal menciptakan beberapa kemungkinan jawaban / aksi
	*Predefined Process*	Simbol guna penyimpanan yang bakal dipergunakan sebagaimana tempat pengolahan didalam storage
	Terminal	Simbol guna permulaan ataupujn akhir pada suatu program
	Manual Input	Simbol guna pemasukan data secara

		manual on-line keyboard
	Arus / Flow	Penghubung antara prosedur/ proses
	*Connector*	Simbol keluar / masuk prosedur ataupun proses pada lembar/ halamanyang sama
	*Off-line Connector*	Simbol keluar / masuk prosedur ataupun proses pada lembar / halaman yang lain
	*Input-Output*	Simbol yang mengungkapkan proses input dan *output* tanpa bergantung dengan jenis peralatannya
	*Document*	Simbol yang mengungkapkan input berasal dari dokumen dalam bentuk kertas ataupun *output* di cetak dikertas

2.7 *White Box Testing*

Pengujian white box adalah metode pengujian yang memusatkan perhatian pada internal sistem, khususnya source code program. Fokus utama dari pengujian white box adalah untuk mengevaluasi kompleksitas code program sebagai alat pengujian. Bagi para programmer, pengujian white box sangatlah esensial dalam menilai tingkat kompleksitas suatu code. Selain itu, pengujian white box juga bermanfaat sebagai validasi terhadap beberapa aspek penting, seperti kepatuhan source code terhadap desain yang ada, kesesuaian source code dengan kebutuhan fungsional, dan keberadaan kerentanan pada source code tersebut (C Munaiseche & Rorimpandey, 2021).

2.8 *Flowgraph*

Flowgraph yakni representasi visual dari alur logika program. Flowgraph bisa disusun berdasarkan kode program ataupun flowchart sistem yang ada. Penandaan dalam flowgraph biasanya terdiri dari lingkaran untuk menandakan node dan anak panah untuk menunjukkan edge. Node mewakili pernyataan prosedural, sementara edge menggambarkan urutan jalannya logika program (C Munaiseche & Rorimpandey, 2021).

2.9 Penelitian Terdahulu

Pada bagian ini akan membahas mengenai tabel penelitian terdahulu yang berhubungan dengan karya ilmiah penulis, yang terdapat pada tabel berikut :

Tabel 2.2 Penelitian Terdahulu

No	Nama	Judul	Keterangan
1.	(Prayogo Putra Tjaya, Rino, 2021)	MPLEMENTASI METODE *CLUSTERING K- MEANS* UNTUK REKOMENDASI PENGADAAN STOK LAMPU DI PT GLOBAL LIGHTING INDONESIA	Pada penelitian ini menghasilkan suatu model program yang mampu mengelompokkan produk paling diminati, diminati dan kurang diminati
2.	(Supardi & Kanedi, 2020)	IMPLEMENTASI METODE ALGORITMA *K- MEANS CLUSTERING* PADA TOKO EIDELWEIS	Penelitian ini menghasilkan sistem yang menerapkan metode *k- means clustering*.
3.	(Annur, 2019)	Penerapan *Data mining* Menentukan Strategi Penjualan Variasi Mobil Menggunakan Metode *K*-	Pada penelitian ini dibahas pengimplementasian metode *K-MEANS CLUSTERING* dalam Pengelompokkan data

		means Clustering	penjualan variasi mobil.
4.	(Wulandari, 2020)	*Clustering* Kecamatan di Kota Bandung Berdasarkan Indikator Jumlah Penduduk dengan Menggunakan Algoritma *K-means*	Dari hasil penelitian dan perhitungan *clustering* Kecamatan di Kota Bandung mempergunakan metode *K-means Clustering* dapat diterapkan