BAB II
LANDASAN TEORI

2.1 Data Mining

Menurut (Wanto et al., 2020) Data mining merupakan metode untuk memperoleh informasi bermanfaat dari gudang data besar, yang dapat diartikan sebagai proses ekstraksi informasi baru dari kumpulan data besar yang membantu pengambilan keputusan. Data mining dapat mengungkapkan tren dan pola tersembunyi yang tidak terlihat dalam analisis query sederhana sehingga memiliki peran penting dalam penemuan pengetahuan dan pengambilan keputusan.

Menurut (Rivanthio & Ramdhani, 2020) Data mining merupakan upaya analisis data pengamatan set dengan tujuan menemukan keterkaitan yang tidak terduga dan merangkum data dalam format baru yang mudah dipahami dan berguna bagi pemilik data. Tugas utama dari data mining ialah untuk mengelompokkan data (klaster), selain itu juga memiliki fungsi lain seperti deskripsi, estimasi, prediksi, klasifikasi, dan asosiasi.

Menurut (Rerung, 2018) Data mining dibagi menjadi beberapa kelompok berdasarkan tugas/pekerjaan yang dapat dilakukan yaitu:

1. Deskripsi: terkadang para ahli dan pengamat secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.
2. Estimasi: Sama dengan klasifikasi, hanya saja variabel target estimasi lebih cenderung ke arah numerik daripada kategori. Model dibuat dengan menggunakan baris data lengkap yang memberikan nilai variabel target sebagai prediksi. Kemudian, pada peninjauan berikutnya, nilai estimasi variabel target dibuat berdasarkan nilai prediksi variabel.
3. Prediksi: hampir sama dengan klasifikasi dan estimasi, kecuali bahwa pada prediksi, nilai hasil akan diproyeksikan ke masa depan. Beberapa cara dan teknik

yang diterapkan pada pengelompokan dan perkiraan juga dapat digunakan (jika sesuai) pada prediksi.

4. Klasifikasi: terdapat target variabel kategori. Sebagai ilustrasi, pengelompokan pendapatan bisa dibagi menjadi tiga kategori, yakni pendapatan tinggi, pendapatan menengah, dan pendapatan rendah.
5. Pengklasteran: Pengklasteran adalah suatu teknik pengelompokan data dengan cara mengamati dan memperhatikan kemiripan antara obyek-obyek yang ada, sehingga membentuk kelas-kelas yang serupa. Kelompok atau klaster adalah gabungan dari beberapa data yang memiliki kemiripan satu sama lainnya dan berbeda dari klaster lainnya. Bedanya dengan klasifikasi, pengklasteran tidak memperhatikan variabel target. Tujuan dari pengklasteran bukan untuk melakukan klasifikasi, estimasi, atau prediksi nilai variabel target, melainkan mencoba untuk membagi data menjadi kelompok-kelompok yang homogen, di mana kesamaan antara data dalam satu kelompok akan maksimal dan kesamaan dengan data dalam kelompok lain akan minimal.
6. Asosiasi: Tugas asosiasi dalam data mining adalah mencari atribut yang muncul secara bersamaan. Salah satu contoh implementasi dari asosiasi adalah analisis keranjang belanja di pasar, yang akan dijelaskan dalam penelitian ini.

2.2 Clustering

Menurut (Madhulatha, 2012) *Clustering* adalah proses mengelompokkan objek yang serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah set data ke dalam subset, sehingga setiap subset memiliki makna yang berguna. Setiap kelompok terdiri dari benda-benda yang serupa satu sama lain dan berbeda dari benda dalam kelompok lain. Algoritma pengelompokan terdiri dari dua bagian, yaitu hirarkis dan partisional. Algoritma hirarkis menemukan kelompok secara berurutan di mana kelompok ditetapkan sebelumnya, sedangkan algoritma partisional menentukan semua kelompok pada waktu tertentu.

Menurut (Bataineh et al., 2011) Pengelompokan data, atau *clustering*, memiliki peran vital dalam kehidupan sehari-hari, karena sangat terkait dengan sejumlah besar

data yang dapat memberikan informasi penting untuk memenuhi kebutuhan hidup. Salah satu alat yang sangat penting dalam hal ini adalah klasterisasi atau pengelompokan data ke dalam sejumlah kategori atau cluster. *Clustering* dapat diterapkan pada berbagai aplikasi di berbagai bidang, seperti analisis data statistik, pembelajaran mesin, penambangan data, pengenalan pola, analisis citra, dan bioinformatika.

2.3 K-Means

Algoritma K-Means adalah algoritma klasterisasi yang mengelompokkan data berdasarkan titik pusat klaster (centroid) terdekat dengan data. K-Means adalah salah satu teknik pengelompokan non-hirarki yang berupaya mempartisi data menjadi satu atau lebih klaster. Teknik ini mempartisi data menjadi klaster sehingga data dengan karakteristik yang sama dan data dengan karakteristik yang berbeda dikelompokkan ke dalam klaster yang berbeda (Arofah & Marisa, 2018).

Algoritma K-means melibatkan pengulangan proses untuk memperoleh basis data cluster. Input yang dibutuhkan adalah jumlah cluster awal yang diinginkan, dan outputnya adalah titik centroid akhir. Metode K-means akan memilih pola k secara acak sebagai titik awal centroid. Jumlah iterasi yang diperlukan untuk mencapai centroid cluster dipengaruhi oleh pemilihan awal centroid secara acak. Oleh karena itu, algoritma dapat ditingkatkan dengan menentukan centroid cluster berdasarkan kepadatan data awal yang tinggi, sehingga meningkatkan kinerja (F. Eltibi & M. Ashour, 2011; Hung et al., 2005). Adapun langkah-langkah dalam algoritma K-Means (Ningsih, 2022) sebagai berikut:

1. Tentukan nilai *k* sebagai jumlah klaster yang ingin dibentuk. Inisialisasi *k* pusat klaster ini bisa dilakukan dengan berbagai cara, namun yang paling seringdilakukan adalah dengan cara random yang di ambil dari data yang ada.
2. Menghitung jarak setiap data *input* terhadap masing – masing *centroid* menggunakan rumus jarak *Euclidean* (*Euclidean Distance*) hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Persamaan *Euclidean Distance* ditampilkan pada **persamaan 2.2**.

3. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
4. Memperbaharui nilai *centroid*. Nilai *centroid* baru di peroleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan persamaan 2.1 berikut:

   $$\mu j(t+1) = \frac{1}{Nsj}\sum_{j \in sj} xj \qquad \text{.....(2.1)}$$

   Dimana:

   μj(t+1) : centroid baru pada iterasi ke (t+1)

   Nsj : banyak data pada cluster sj
5. Melakukan perulangan dari langkah 2 hingga 4, sampai anggota tiap *cluster* tidak ada yang berubah. Jika langkah 5 telah terpenuhi, maka nilai pusat *cluster* (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

2.4 Jarak Euclidean

Menurut (Eviana et al., 2022) Jarak Euclidean merupakan perhitungan jarak antara dua buah titik. Konsep Euclidean terkait dengan Teorema Phytagoras, dimana perhitungan akar kuadrat digunakan. Euclidean space diperkenalkan oleh Euclid, seorang matematikawan dari Yunani pada sekitar tahun 300 B.C.E. untuk mempelajari hubungan antara sudut dan jarak. Teorema Phytagoras sering diterapkan pada dimensi yang lebih tinggi dalam Euclidean. Fungsi heuristik Euclidean dapat diperoleh dari perhitungan jarak langsung, seperti ketika mencari panjang garis diagonal pada segitiga. Persamaan 2.2 menunjukkan rumus Jarak Euclidean.

$$d(x,y) = \sqrt{\sum(x-y)^2} \qquad \text{.....(2.2)}$$

Keterangan :

d(x,y) = Jarak antara data pada titik x dan titik y

x = Data

y = Pusat Cluster

2.5 Davies-Bouldin Index

Menurut (Bates & Kalita, 2016) Indeks Davies Bouldin (DBI) merupakan salah satu teknik yang dipakai untuk mengukur keabsahan atau jumlah klaster yang optimal dalam suatu teknik pengelompokan di mana kohesi didefinisikan sebagai total kedekatan data dengan titik pusat klaster yang diikuti. Evaluasi dengan menggunakan Indeks Davies Bouldin ini menggunakan skema evaluasi dari klaster internal, di mana kebaikan atau ketidakbaikan hasil klaster dilihat dari kuantitas dan kedekatan antar data hasil klaster.

Menurut (Sitompul, 2018) Tahapan dalam perhitungan Davies Bouldin-Index adalah sebagai berikut:

1. *Sum of Square Within-cluster (SSW)*

Untuk mengetahui kohesi dalam sebuah *cluster* ke-i adalah dengan menghitung nilai dari *Sum of Square Within-cluster* (SSW). Kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari sebuah cluster yang diikuti. Persamaan yang digunakan untuk menghitung nilai Sum of Square Within-cluster ditampilkan pada persamaan 2.3.

$$\text{SSW} = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - c_{pi}\right\|^2 \qquad \text{.....(2.3)}$$

2. *Sum of Square Beetwen-cluster (SSB)*

Perhitungan *Sum of Square Between-cluster* (SSB) bertujuan untuk mengetahui separasi antar *cluster*. Persamaan yang digunakan untuk memperoleh nilai *Sum of Square Between cluster* ditampilkan pada persamaan 2.4.

$$SSB = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,j\neq 1}^{M}\left\|c_i - c_j\right\|^2 \qquad \text{.....(2.4)}$$

3. *Ratio*

Perhitungan *Ratio* bertujuan untuk mengetahui nilai perbandingan antara *cluster* ke-i dan *cluster* ke-j. Persamaan yang digunakan untuk menghitung nilai *Ratio* ditampilkan pada persamaan 2.5(Nawrin et al., 2017).

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \qquad \text{.....(2.5)}$$

4. *Davies Bouldin Index*

Dari persamaan berikut ini, *k* adalah jumlah *cluster*. Semakin kecil nilai Davies- Bouldin Index (DBI) yang diperoleh (non-negatif >= 0), maka semakin baik *cluster* yang diperoleh dari pengelompokan menggunakan algoritma *clustering* (Bates & Kalita, 2016). Persamaan DBI ditampilkan pada persamaan 2.6.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \underset{i \neq j}{max}\left(R_{i,j}\right) \quad \text{.....(2.6)}$$

2.6 Penelitian Terdahulu

Sebagai bahan referensi penulis melakukan analisis terhadap beberapa penelitian sebelumnya yang berkaitan dengan topik penelitian yang dilakukan. Berikut merupakan hasil dari analisis penelitian sebelumnya:

a. Amir Ali (2019) dengan judul penelitian "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo" mendapatkan kesimpulan, bahwa peneliti dapat mengidentifikasi data rekam medis dari rumah sakit anwar medika sebanyak 534 data pasien dengan waktu penyelesaian sebanyak 0.06 detik oleh sistem (Ali, 2019).
b. Darmansah dan Ni Wayan Wardani (2021) dengan judul penelitian "Analisis Pesebaran Penularan Virus Corona di Provinsi Jawa Tengah Menggunakan Metode K-Means Clustering" mendapatkan kesimpulan, dengan menggunakan metode tersebut dapat mengelompokkan persebaran penularan virus corona di jawa tengah (Darmansah & Wardani, 2021).
c. Dezty Adhe Chajannah Rachman, Rito Goejantoro, dan Fidia Deny Tisna Amijaya (2020) dengan judul penelitian "Implementasi *Text Mining* Pengelompokkan Dokumen Skripsi Menggunakan Metode *K-Means Clustering*" mendapatkan kesimpulan, Banyaknya kelompok optimal yang terbentuk dari dokumen skripsi menggunakan metode *KMeans Clustering* adalah 2 cluster dengan nilai *silhouette coefficient* 0,12 yang berarti *no structure* (Adhe et al., 2020).

d. Zulfa Nabila, Auliya Rahman Isnain, Permata, dan Zaenal Abidin (2021) dengan judul penelitian "Analisis Data Mining Untuk *Clustering* Kasus Covid-19 di Provinsi Lampung Dengan *Algoritma K-Means*" mendapatkan kesimpulan, hasil pengelompokkan yang berbeda dikarenakan jumlah pada atribut Suspek, Probable, Konfirmasi Positif, Selesai Isolasi, dan Kematian pada setiap Kabupaten/Kota yang tidak sama (Nabila et al., 2021).
e. Sabrina Aulia Rahmah (2020) dengan judul penelitian "Klasterisasi Pola Penjualan Pestisida Menggunakan Metode *K-Means Clustering* (Studi Kasus di Toko Juanda Tani Kecamatan Hutabayu Raja)" mendapatkan kesimpulan, Metode *K-Means clustering* dapat diterapkan pada penjualan Pestisida di Toko Jaunda Tani, metode ini sangat membantu dalam mengelompokan pola penjualan selama satu musim (Aulia, 2021).
f. Lidia Gayatri dan Hendry (2021) dengan judul penelitian "Pemetaan Penyebaran Covid-19 Pada Tingkat Kabupaten/Kota di Pulau Jawa Menggunakan Algoritma *Kmeans Clustering*" mendapatkan kesimpulan, Hasil implementasi persebaran kasus Covid-19 di Pulau Jawa menggunakan algoritma *K-Means clustering* mendapatkan 3 cluster dari hasil pengujian menggunakan DBI dengan nilai 0,609 (Gayatri & Hendry, 2021).
g. Juniar Hutagalung, Yopi Hendro Syahputra, dan Zohana Pertiwi Tanjung (2022) dengan judul penelitian "Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma *K-Means Clustering*" mendapatkan kesimpulan, Penerapan algoritma k-means, mampu mempercepat dalam menentukan pengelompokan siswa kelas unggulan (Hutagalung, 2022).
h. Lilis Suriani (2020) dengan judul penelitian "Pengelompokan Data Kriminal Pada Poldasu Menentukan Pola Daerah Rawan Tindak Kriminal Menggunakan Data Mining Algoritma *K-Means Clustering*" mendapatkan kesimpulan, Pembagian pengelompokan daerah rawan Poldasu belum efektif dan efesien,dikerenakan belum adanya metode khusus untuk mendukung keberhasilan pembagian pengelompokkan daerah Rawan, dan masih dilakukan secara manual (Suriani, 2020).

i. Suhandio Handoko, Fauziah, dan Endah Tri Esti Handayani (2020) dengan judul penelitian "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode *K-Means Clustering*" mendapatkan kesimpulan, hasil dari metode Algoritma *K-Means Clustering* data mining didapatkan daerah penjualan produk yang tinggi, sedang, dan rendah. Daerah dengan penjualan produk yang rendah akan dilakukan promosi penjualan produk sedangkan untuk daerah penjualan yang tinggi tidak diadakan promosi (Handoko et al., 2020).

j. Wildan Maulidi Molyono, Sentot Achmadi, dan Yosep Agus Pranoto (2021) dengan judul penelitian "Pemetaan Tambak Garam Serta Produksi Garam Pada Kabupaten Pamekasan Menggunakan *K-Means Clustering*" mendapatkan kesimpulan, Berdasarkan pengujian dari data tambak garam Dinas Perikanan Kabupaten Pamekasan, terdapat 5,26% daerah yang tidak perlu meningkatkan produksi garam. 26,32% daerah yang kurang perlu meningkatkan produksi, dan 68,42% daerah yang perlu meningkatkan hasil produksi garamnya (Maulidi Molyono et al., 2021).

k. Fintri Indriyani dan Eni Irfiani (2019) dengan judul penelitian "Clustering Data Penjualan pada Toko Perlengkapan Outdoor Menggunakan Metode K-Means" mendapatkan kesimpulan, Penerapan metode K-Means dalam pengelompokan data penjualan pada Toko Genta Corp dapat menghasilkan rekomendasi barang yang laris, Kurang laris dan cukup laris (Indriyani & Irfiani, 2019).

l. Rizki Muliono dan Zulfikar Sembiring (2019) dengan judul penelitian "Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen" mendapatkan kesimpulan, bahwa jumlah ketepatan prediksi yang dilakukan oleh algortima K-means terhadap 15 data mengalami perbedaan sebanyak 53.33 % keakuratan prediksi (Muliono & Sembiring, 2019).

m. Hendro Priyatman, Fahmi Sajid, dan Dannis Haldivany (2019) dengan judul penelitian "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa" mendapatkan kesimpulan, bahwa pada kasus ini implementasi algoritma k-means dalam data mining sudah berhasil,

dan bisa menampilkan informasi prediksi kelulusan mahasiswa (Priyatman et al., 2019).