BAB II
TINJAUAN PUSTAKA

2.1 Pengertian Data Mining

Data Mining adalah proses menyaring sejumlah besar data yang disimpan dalam repositori, memanfaatkan teknologi statistik, matematika, dan pengenalan pola untuk mengidentifikasi koneksi, pola, dan tren baru yang relevan. *Clustering* adalah salah satu teknik yang digunakan dalam data mining. Salah satu teknik analisis data yang dapat digunakan untuk mengatasi permasalahan dalam suatu pengelompokan data adalah clustering. Nama umum lainnya untuk data mining adalah Knowledge Discovery in Databases (KDD) (Rosida & Wijaya, 2023).

KDD melibatkan pengumpulan dan analisis data masa lalu untuk mengidentifikasi pola, korelasi, atau keteraturan dalam kumpulan data yang sangat besar. Proses Knowledge Discovery in Database (KDD) yang menggunakan algoritma untuk memeriksa data, membuat model, dan mengidentifikasi tren masa lalu, berpusat pada penambangan data. Pengenalan pola (patteran recognition) adalah istilah lain untuk proses mengungkap pola-pola tersembunyi dari data olahan yang sebelumnya tidak ditemukan (Rosida & Wijaya, 2023).

2.1.1 Teknik Data Mining Data

Didalam *data mining* terdapat beberapa teknik yang dapat digunakan (Sari et al., 2020) yaitu sebagai berikut:

1. **Deskripsi**

   Deskripsi dapat digunakan untuk menyoroti tren dan pola dalam data yang disimpan.

2. **Estimasi**

   Dengan pengecualian variabel tujuan estimasi yang lebih bersifat numerik daripada kategorikal, estimasi dan klasifikasi hampir sama.

Nilai variabel target diprediksi oleh model dengan menggunakan catatan lengkap.

3. **Prediksi**

   Prediksi menerka nilai yang tidak diketahui serta perkiraan nilai di masa depan.

4. **Klasifikasi**

   Dalam klasifikasi terdapat target variabel kategori misalnya, pendapatan dapat dibagi menjadi tiga kategori: tinggi, sedang, dan rendah.

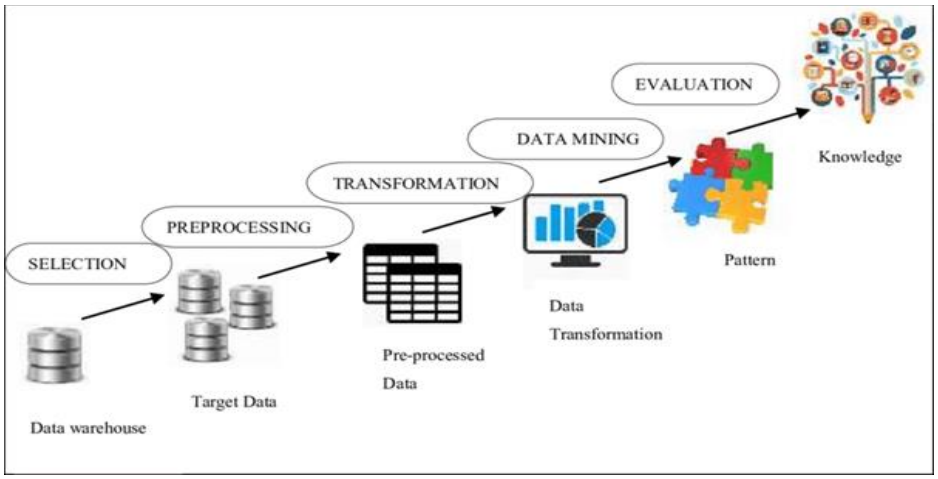
5. **Pengklasteran**

   Merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

6. **Asosiasi**

   Asosiasi bertugas menemukan atribut yang muncul secara bersamaan, seperti yang sering terjadi di ranah retail khususnya.

2.1.2 Tahapan Data Mining



Gambar 2.1 Tahapan KDD

Knowledge Discovery in Database (KDD) KDD adalah proses non-sepele yang menemukan potensi, kegunaan, dan validitas dalam data dan pada akhirnya menghasilkan pola yang dapat dipahami dalam data. KDD berkaitan dengan penemuan ilmiah, interpretasi, dan visualisasi pola dalam

berbagai kumpulan data melalui pendekatan integrasi. Langkah-langkah proses KDD yang digambarkan pada Gambar 2.1 di atas meliputi data mining. (Haris Kurniawan et al., 2020).

1. **Data Selection**

   Data *selection* (pemilihan data) adalah proses memilih subset data target, kualitas, atau indikator dari sejumlah data yang besar.

2. **Pre-pocessing / Cleaning Data**

   Pembersihan adalah proses mengidentifikasi dan memperbaiki ketidakkonsistenan data, seperti masalah tipografi atau pencetakan, dan menghilangkan data yang berlebihan.

3. **Transformation Data**

   Transformasi data adalah proses pengkodean atau pengubahan data. Transformasi data dilakukan untuk mempermudah pengolahan data dan menyederhanakan pemrosesan data.

4. **Data Mining**

   Data mining merupakan proses pencarian pola atau informasi yang menarik/bermanfaat dalam data yang besar dengan menggunakan teknik atau metode tertentu.

5. **Interpretation/ Evaluation**

   Proses penyajian pola, data, atau pengetahuan yang telah diekstraksi dari proses data mining sedemikian rupa sehingga mudah dipahami oleh pihak yang berkepentingan disebut interpretasi dan evaluasi.

6. **Knowledge Presentation**

   Knowledge presentation merupakan tahapan akhir dalam proses data mining. Bagaimana pengetahuan yang telah ditemukan akan disajikan

kepada pengguna. Tidak semua user memahami data mining, oleh karena itu penting untuk menyusun dengan baik penyajian hasil data mining dalam bentuk yang dapat dipahami oleh user. Dalam hal ini, visualisasi juga dapat digunakan untuk membantu menyampaikan hasil *data mining* (Andini et al., 2022).

2.2 Clustering

Clustering adalah proses pengorganisasian data, melakukan observasi, memperhatikan, dan mengelompokkan objek-objek pembanding ke dalam kelompok-kelompok. *Cluster* adalah sekelompok record yang berbeda dengan record di cluster lain namun serupa atau mirip satu sama lain,dimana kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal (Nabila et al., 2021).

Clustering dalam *data mining* berguna untuk menemukan pola distribusi di dalam sebuah data set yang berguna untuk proses analisa data. Kesamaan objek biasanya diperoleh dari kedekatan nilai-nilai atribut yang menjelaskan objek-objek data, objek data biasanya direpresentasikan sebagai titik dalam ruang multidimensi.(Nabila et al., 2021).

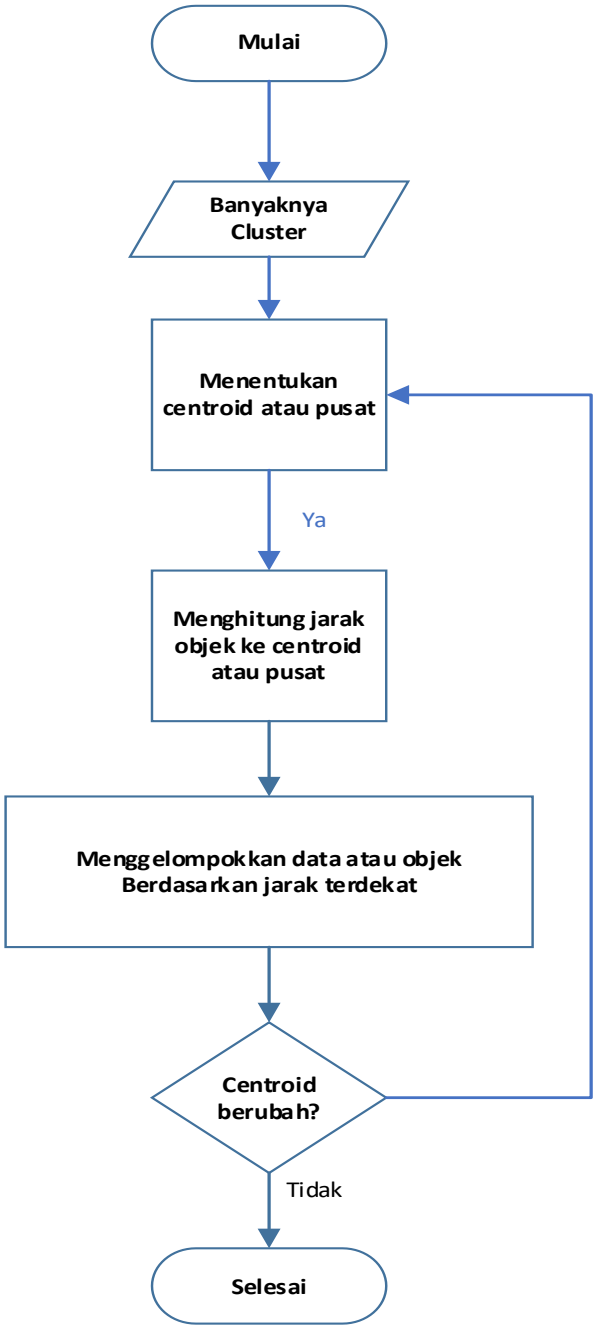
2.3 K-Means Clustering

K-Means Clustering merupakan metode yang digunakan dalam data mining yang cara kerjanya mencari dan mengelompokkan data yang mempunyai kemiripan karakteristik antara data satu dengan yang lain (Fajri & Purnamasari, 2022). *K-Means* dapat digambarkan sebagai algoritma pengelompokan berulang yang membagi kumpulan data tertentu menjadi k cluster yang telah ditentukan. Saat memproses objek dalam jumlah besar, algoritme *K-Means* relatif lebih dapat diukur dan efisien karena presisinya yang relatif tinggi mengenai ukuran objek. Menentukan *centroid*, jumlah *cluster*, dan jarak centroid merupakan salah satu langkah penting dalam penerapan *K-Means Cluster*. (Rosida & Wijaya, 2023).

Istilah-istilah didalam algoritma *K-Means Clustering* (Muliono & Sembiring, 2019):

1. Cluster: *Cluster* adalah kelompok atau grup.
2. Cendroid: *Cendroid* dadalah titik pusat untuk menentukan *auclidian distance*.
3. Iterasi: Iterasi adalah pengulangan proses, berhenti ketika hasil iterasi telah konvergen (Muliono & Sembiring, 2019).

Langkah-langkah pada algoritma *K-Means* dapat dilihat pada gambar 2 di bawah:



Gambar 2.2 Flowchart proses algoritma K-Means

Adapun proses dan langkah-langkah pada algoritma *K-Means Clustering* dapat dibaca sebagai berikut:

1. Menentukan jumlah k sebagai *cluster* yang akan dibentuk. Penentuan banyaknya jumlah *cluster* k biasanya dilakukan dengan

beberapa faktor pertimbangan baik teoritis dan konseptual yang kemudian diusulkan untuk menentukan jumlah cluster.

2. Buat k *centroid* awal yang dipilih secara acak, atau titik pusat klaster. *Centroid* pertama dipilih secara acak dari sejumlah objek, hingga maksimum k *cluster*, dan *centroid cluster* berikutnya kemudian dihitung.
3. Menghitung jarak dari setiap objek ke masing-masing *centroid* dari masing-masing data *cluster* dan menghitung jarak antara objek dengan *centroid* yang bisa di lakukan dengan menggunakan rumus matematik *euclidian distance*:

   $$d(x, y) = |x - y| = \sum_{i=1}^{n}(x1 - y1)^2 \quad ; i = 1,2,3, \ldots, n$$

   Keterangan:

   x= pusat cluster

   y= data

   n= banyaknya objek
4. Tetapkan setiap item ke pusat centroid yang paling dekat. Secara umum, *k-means* dapat digunakan untuk menetapkan item ke setiap *cluster* selama iterasi, setiap objek dinyatakan sebagai anggota *cluster* dengan mengukur jarak kedekatan sifatnya terhadap titik pusat cluster tersebut.
5. Dengan menggunakan persamaan sebelumnya, hitung rata-rata setiap cluster untuk menemukan pusat cluster atau centroid baru.

   $$v = \frac{\sum_{i}^{n} = 1^{Xi}}{N}; \quad i = 1,2,3, \ldots n$$

   Keterangan:

   v = Centroid pada cluster

   xi = Objek ke-i

   n = Banyaknya jumlah objek yang menjadi anggota cluster

6. Jika posisi pusat centroid yang baru berbeda, ulangi langkah pada nomor 3. Perulangan atau iterasi akan berakhir jika rasio tidak bertambah melebihi nilai awal hingga semua data hasil perhitungan konvergen (Muliono & Sembiring, 2019).

2.4 RapidMiner

RapidMiner adalah perangkat lunak sumber terbuka yang tersedia untuk semua orang. Salah satu alat untuk memproses dan menganalisis data disebut *RapidMiner*. *RapidMiner* menggunakan sejumlah metode, termasuk metode prediktif dan deskriptif. *Java* adalah bahasa yang digunakan *RapidMiner*. Dengan bantuan analisis data berbasis komputer dan teknologi algoritma komputasi canggih, perangkat lunak *Rapidminer* dapat memproses *data mining*. Perhitungan algoritma *K-Means Clustering* akan dilakukan dengan menggunakan *Rapidminer*. Hasil penelitian ini diharapkan dapat memberikan manfaat bagi Puskesmas Hanura mengelompokkan penyakit pasien berdasarkan usia pasien yang dapat membantu pihak Puskemas dalam melakukan pencegahan dan pengobatan dengan sosialisasi kepada masyarakat tentang pentingnya menjaga kesehatan (Sari et al., 2020).

2.5 Tableau

Tableau adalah platform yang memfasilitasi pembuatan visualisasi data bergaya *dasbord*, membuat data lebih mudah untuk berinteraksi dan dipahami. *Tableau* adalah *platform* yang membantu membuat representasi data dalam format grafik atau pemetaan grafik yang membuat data lebih mudah untuk berinteraksi dan dipahami. Dalam penelitian ini aplikasi *Tableau* membantu menampilkan *persentase* % pada setiap penyakit (Afikah et al., 2022).

2.6 David Bouldin Index (DBI)

Indeks Davies-Bouldin (DBI), yang dikembangkan oleh David L. Davies dan Donald W. Bouldin, adalah metode untuk mengevaluasi *cluster*. Skema penilaian klaster internal dari Skor *Indeks* Davies-Bouldin menetapkan skor klaster yang baik atau buruk tergantung pada kuantitas skor klaster dan tingkat kemiripan antar skor klaster. Dengan membandingkan jarak *intra-cluster* (jarak rata-rata semua titik data dalam sebuah cluster dari pusat centroid) ke jarak antar-*cluster* (jarak antara dua centroid). Semakin kecil nilai DBI, semakin baik kekompakan atau pemisahannya, sebaliknya untuk nilai DBI yang besar (Hastari et al., 2023). Dalam Pengujian DBI, *cluster* dengan niali DBI terkecil atau mendekati 0, dijadikan sebagai cluster terbaik. (Hasanah et al., 2024).

2.7 Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

No.	Judul/Tahun	Metode	Tujuan	Hasil
1	KLASTERISASI POLA PENYEBARAN PENYAKIT PASIEN BERDASARKAN USIA PASIEN MENGGUNAKAN K MEANS CLUSTERING (2022)			
2	DATA MINING DENGAN ALGORITMA NEURAL NETWORK DAN	Metode data mining klasifikasi yang akan digunakan adalah	Algoritma neural network digunakan untuk memprediksi	Hasil penelitian yang dilakukan dari tahap

	VISUALISASI DATA UNTUK PREDIKSI KELULUSAN MAHASISWA (2020)	algoritma neural network	kelulusan mahasiswa yang sulit dilakukan secara manual, sedangkan visualisasi digunakan untuk menampilkan data rekapitulasi secara visual sehingga lebih menarik dan mudah dipahami.	awal sampai dengan tahap pengujian menggunakan Data Mining dengan algoritma Neural Network untuk kelulusan mahasiswa menghasilkan prediksi dengan Precision 87.80%, Recall 86.90% dan nilai akurasi sebesar 92.83%. Sedangkan dari visualisasi data dari dataset yang berjumlah 2742 record menampilkan beberapa rekapitulasi pelaporan berupa dashboard yang sangat komplit, sehingga dengan prediksi dan visualisasi data tersebut dapat membantu dalam kelulusan

				mahasiswa dan memberikan rekomendasi tindakan yang tepat dan harus dilakukan oleh manajemen atau pihak yang berwenang untuk mengambil keputusan.
2	Penerapan Data Mining dalam Perancangan Sistem Pendukung Keputusan Seleksi Penerimaan Beasiswa Menggunakan Naive Bayes Classifier (Studi Kasus: IIB Darmajaya) 2020	Teknik data mining dalam perancangan sistem pendukung keputusan seleksi penerimaan beasiswa bagi mahasiswa berprestasi menggunakan metode Naïve Bayes	Tujuan dari pelaksanaan penelitian ini adalah membangun suatu sistem yang akan digunakan guna mendukung proses pengambilan keputusan dengan menerapkan Teknik datamingung memanfaatkan algoritma Naïve Bayes Classiffier agar dapat membantu pihak IIB Darmajaya khususnya Unit	Hasil dari implementasi sistem ini ialah memberikan keterangan tentang informasi penerima beasiswa berdasarkan rengking yang dapat digunakan sebagai alat bantu dalam proses pengambilan keputusan. Dengan adanya sistem ini, proses perhitungan untuk menentukan penerima beasiswa dapat dilakukan

			Kemahasiswaan dalam penagambilan keputusan dalam penyeleksian penerimaan beasiswa prestasi.	dengan mudah, cepat dan akurat.
3	PENGELOMPOKKAN PENYAKIT PASIEN MENGGUNAKAN ALGORITMA K-MEANS (2022)	Algoritma yang digunakan dalam penelitian ini adalah algoritma KMeans.	Untuk mempermudah proses pengelolaan data yang banyak, Puskesmas Bahorok memerlukan suatu sistem dalam mengambil keputusan untuk mengetahui pengelompok kan penyakit berdasarkan usia pasien yang sering terkena penyakit pada Puskesmas Bahorok.	Pengelompokan dengan metode KMeans dapat menghasilkan jumlah cluster yang sama dengan jumlah data yang berbeda – beda tanpa harus memiliki data yang sama. Dengan dibangunnya sistem ini untuk mempermudahkan user dalam mengelompokan penyakit pada pasien berdasarkan usia secara efektif dan efisien khususnya untuk Staff Pegawai dan Administrasi.

				Dengan metode KMeans sangatlah mempermudah kan user dalam mengelompok an suatu data hanya dengan memiliki karakteristik yang sama. s
4	Komparasi Metode Apriori dan FP-Growth Data Mining Untuk Mengetahui Pola Penjualan (2023)	Algoritma yang dapat digunakan untuk mengelola data penjualan dalam mengatasi masalah tersebut adalah Apriori dan FP-Growth. Adapun metode penelitian yang digunakan pada penelitian ini adalah proses KDD (Knowledge Discovery in Database)	Tujuan penelitian ini adalah untuk mengetahui pola penjualan produk terlaris dan untuk meningkatka n kuantitas penjualan produk parfum	Penelitian ini menghasilkan pola frekuensi tinggi untuk itemsets dengan nilai minimum support 20% menghasilkan produk yang menjadi The Most Tree Items adalah Jo Malone 82,49%, Zarra 28,25%, dan Zwitsal 20,34%. Sedangkan aturan asosiasi yang terbentuk dari nilai Min. Supp 20% dan Min. Conf 80%, mendapatkan kombinasi 2 itemsets yaitu Jo Malone dan

				Zarra. Sedangkan untuk kombinasi 3 itemsets yaitu Jo Malone, Zarra dan Baccarte dengan status valid dan kuat dibuktikan dengan nilai lift lebih besar dari 1, oleh karena itu aturan asosiasi tersebut sangat tepat untuk dapat digunakan.
5	Komparasi Metode Apriori dan FP-Growth Data Mining Untuk Mengetahui Pola Penjualan (2023)	Algoritma yang dapat digunakan untuk mengelola data penjualan dalam mengatasi masalah tersebut adalah Apriori dan FP-Growth. Adapun metode penelitian yang digunakan pada penelitian ini adalah proses KDD (Knowledge	Tujuan penelitian ini adalah untuk mengetahui pola penjualan produk terlaris dan untuk meningkatkan kuantitas penjualan produk parfum	Penelitian ini menghasilkan pola frekuensi tinggi untuk itemsets dengan nilai minimum support 20% menghasilkan produk yang menjadi The Most Tree Items adalah Jo Malone 82,49%, Zarra 28,25%, dan Zwitsal 20,34%. Sedangkan aturan asosiasi yang terbentuk dari nilai Min. Supp

		Discovery in Database)		20% dan Min. Conf 80%, mendapatkan kombinasi 2 itemsets yaitu Jo Malone dan Zarra. Sedangkan untuk kombinasi 3 itemsets yaitu Jo Malone, Zarra dan Baccarte dengan status valid dan kuat dibuktikan dengan nilai lift lebih besar dari 1, oleh karena itu aturan asosiasi tersebut sangat tepat untuk dapat digunakan.