BAB II
TINJAUAN PUSTAKA

2.1. DATA MINING

2.1.1. Pengertian Data Mining

Data mining merupakan proses eksplorasi dan analisis sejumlah besar data untuk menemukan pola, hubungan, atau pengetahuan yang berguna. Tujuan utama dari data mining adalah untuk mengekstrak informasi yang sebelumnya tidak dikenal dan mengubahnya menjadi pengetahuan yang dapat dimanfaatkan dalam pengambilan keputusan. Data mining, atau pengetahuan dalam database Knowledge Discovery in Databases (KDD), adalah proses untuk menemukan pola dan hubungan dalam kumpulan data besar (Yani et al., 2023). Data historis dugunakan untuk memprediksi masa depan dan membantu pengambilan keputusan (Ananda Mustari dkk., 2024). Salah satu teknik data mining yang populer adalah clustering. Clustering merupakan pengelompokan data yang mirip kedalam cluster (kelompok). Data dalam cluster yang sama memiliki kesamaan dan berbeda dengan data di cluster lain (Tiara Alifa dkk., 2024).

Berikut ini merupakan serangkaian proses KDD pada data mining:

1. *Data Selection* (pemilihan data), merupakan tahapan pemilihan data yang akan di eksekusi sesuai dengan kebutuhan analisis.
2. Pembersihan data dan proses data (*cleaning and processing*), proses ini dilakukan untuk membersihkan data yang mencakup pembuangan data yang duplikat, noise, dan menangani data yang tidak konsisten dan relevan.
3. Tranformasi data (*transformation*), pada tahap ini data disesuaikan dengan format ekstensi yang sesuai untuk proses data mining. Hal ini dilakukan karena beberapa metode data mining memerlukan format tertentu sebelum diproses.
4. Data mining, pada tahap ini dilakukan untuk mencari pola atau informasi menarik dalam data yang telah dipilih dengan menggunakan fungsi-fungsi tertentu.

5. Evaluasi pola dan presentasi pengetahuan (*knowledge extraction*), tahap ini merupakan pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya (Afiasari dkk., 2023).

2.1.2. Metode Data Mining

Secara umum, metode data mining dapat dibagi menjadi dua yaitu, deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif adalah data mining digunakan untuk membentuk sebuah model pengetahuan yang akan digunakan untuk melakukan prediksi (Setiyani dkk., 2020).

Metode yang ada dalam data mining adalah sebagai berikut:

1. *Clasification*

   Klasifikasi adalah proses untuk mencari model atau fungsi yang menggambarkan dan membedakan kelas-kelas atau konsep data.
2. *Clustering*

   Pengelompokan data yang tidak diketahui kelasnya kedalam sejumlah kelompok tertentu sesuai dengan tingkat kemiripanya. Metode inilah yang digunakan dalam tugas akhir ini.
3. *Association*

   Tujuan dari metode ini yaitu untuk menghasilkan sejumlah rule yang menjelaskan sejumlah data yang terhubung kuat dengan yang lainya.
4. *Regression*

   Bertujuan untuk mencari prediksi dari suatu pola yang ada
5. *Forecasting*

   Berfungsi untuk meramalkan waktu yang akan datang berdasarkan trend yang telah terjadi di waktu sebelumnya.
6. *Sequence Analysis*

   Tujuan dari metode ini adalah untuk mengenali pola dari data diskrit sebagai contoh adalah menemukan kelompok gen dengan tingkat ekspresi yang mirip.

7. *Deviation Analysis*

   Deviation Analysis adalah suatu metode analisis yang digunakan utnuk mengidentifikasi dan memahami perbedaan antara nilai aktual dan nilai yang diharapkan atau direncanakan dalam suatu proses atau sistem.

2.2 PROGRAM KELUARGA HARAPAN (PKH)

Program Keluarga Harapan (PKH) adalah salah satu program bantuan sosial bersyarat yang dikelola oleh pemerintah Indonesia, khususnya oleh Kementerian Sosial. Program ini bertujuan untuk memberikan bantuan finansial kepada keluarga miskin dan rentan yang memiliki komponen tertentu, seperti ibu hamil, anak usia dini, anak sekolah, penyandang disabilitas berat, dan lanjut usia. Bantuan ini diharapkan dapat mengurangi kemiskinan serta meningkatkan kualitas sumber daya manusia melalui peningkatan akses pendidikan, kesehatan, dan kesejahteraan sosial (Siti Paridah & Martanto, 2024).

PKH adalah program yang berbasis pada pemberian bantuan tunai dengan syarat bahwa penerima harus memenuhi kewajiban tertentu, seperti memastikan anak-anak mereka bersekolah atau mengikuti layanan kesehatan dasar seperti imunisasi dan pemeriksaan kesehatan rutin. PKH telah menjadi salah satu instrumen penting dalam upaya pemerintah untuk mengurangi angka kemiskinan, PKH sendiri telah dilaksanakan sejak tahun 2007 dan merupakan salah satu langkah untuk memutus angka kemiskinan dan mengangkat kesejahteraan masyarakat (Moruk dkk., 2024).

2.3. ALGORITMA *K-MEANS*

K-Means merupakan bagian dari metode pengelompokan data non-hirarki (sekatan) yang memiliki kemampuan mempartisi data kedalam bentuk dua kelompok ataupun lebih. Metoda tersebut akan mempartisi data ke dalam cluster-cluster sehingga data yang memiliki kemiripan berada pada cluster yang sama dan data yang memiliki ketidaksamaan berada pada cluster yang lain. Tujuan dari pengelompokan yaitu untuk meminimalkan

dari fungsi objektif yang diset dalam proses pengelompokan, pada umumnya akan berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok (Nugraha dkk., 2022).

Langkah-langkah dalam mengcluster menggunakan metode K-Means adalah sebagai berikut:

1. Menentukan k sebagai jumlah cluster yang akan dibentuk.
2. Membangkitkan nilai random untuk pusat *cluster* awal (centroid) sebanyak k.
3. Menghitung jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak *Euclidean* (*Euclidean Distance*) hingga ditemukan jarak terdekat dari setiap data dengan *centroid.* Berikut adalah persamaan *Euclidean Distance* (2.1)

$$d(xi, \mu j) = \sqrt{\sum(xi - \mu j)^2} \qquad (2.1)$$

Keterangan :

xi : data kriteria.

μj : *centroid* pada *cluster* ke-*j*

4. Mengklasifikasikan setiap data berdasarkan kedekatanya dengan *centroid* (jarak terkecil).
5. Memperbaharui nilai *centroid.* Nilai *centroid* baru di peroleh dari rata-rata cluster yang bersangkutan dengan menggunakan rumus persamaan (2.2)

$$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \epsilon Sj} xj \qquad (2.2)$$

Keterangan:

$\mu j(t+1)$: *centroid* baru pada iterasi ke (t + 1)

Nsj : banyak data pada *cluster Sj.*

6. Melakukan perulangan dari langkah 2 sampai 5, hingga anggota tiap *cluster* tidak ada yang berubah.

2.4. CLUSTERING

Clustering merupakan suatu kumpulan data yang dibagi menjadi banyak kelompok (cluster) dengan menggunakan pendekatan

pengelompokan, yang didasarkan pada kesamaan yang telah ditentukan sebelumnya. Objek yang ada didalam cluster memiliki kemiripan karakteristik antar satu sama yang lain dan berbeda dengan cluster yang lain. Berkat pengelompokan ini kita dapat mengetahui hubungan yang menarik antara data atribut, menentukan kelompok prioritas, dan mengungkap pola distribusi umum (Wahyu dkk., 2024).

2.5. *EUCLIDEAN DISTANCE*

Euclidean Distance merupakan salah satu metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam euclidean space (meliputi bidang Euclidean dua dimensi, tiga dimensi, atau bahkan lebih). Perhitungan menggunakan geometri Euclidean memberikan hasil yang positif dengan memilih nilai terendah, kesamaan yang diukur menggunakan jarak Euclidean ditentukan (Wahyu Pribadi dkk., 2022). Berikut adalah rumus persamaan *Euclidean Distance* (2.3)

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \quad (2.3)$$

Keterangan:

$d(x,y)$: jarak antara cluster x dengan titik centroid y pada data ke-i.

x : data pada pusat cluster ke-i.

y : bobot data ke-i pada titik centroid.

2.6. *SILHOUETTE COEFFICIENT*

Silhouette Coefficient adalah salah satu teknik yang dilakukan dilakukan untuk mengukur tingkat keoptimalan atau keabsahan sebuah *cluster* yang telah diperoleh dari *clustering*. Pengujian *Silhouette Score* dilakukan setelah mencapai konvergensi 0 dimana hasil pengelompokan terakhir sama dengan pengelompokan sebelumnya. Dengan kata lain, tidak ada data yang berpindah klaster (Zurfani et al., 2024).

Berikut tahapan perhitungan *Silhouette Coefficient* :

1. Menghitung rata-rata jarak suatu data dengan data lain dalam suatu *cluster* sama menggunakan persamaan (2.4)

$$a(i) = \frac{1}{[A]-1} + \sum \mathrm{j}\epsilon_{A,\mathrm{j}}, {}_{i} \;\; d(i,\mathrm{j}) \qquad (2.4)$$

Keterangan :

$a(i)$: Perbedaan rata-rata pada data (i) ke semua data lain di cluster A.

$d(i,\mathrm{j})$: Jarak antara data i dan data j.

2. Menghitung rata-rata jarak data tersebut dengan semua data di *cluster* lain menggunakan persamaan (2.5)

$$d(i,C) = \frac{1}{|C|} \sum_{J \in C} d(i,\mathrm{j}) \qquad (2.5)$$

Keterangan :

$d(i,C)$: Perbedaan rata-rata pada data (i) ke seluruh data C

C : Cluster lain selain A

3. Memilih nilai jarak dengan nilai paling kecil atau minimum menggunakan persamaan (2.6)

$$b(i) = \min_{c,A} d(i,\mathrm{j}) \qquad (2.6)$$

Keterangan :

$b(i)$: Nilai minimum jarak rata-rata data ke-*i* dengan semua data di *cluster* berbeda.

4. Hitung nilai *Silhouette Coefficient* dengan persamaan (2.7)

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (2.7)$$

Keterangan :

$a(i)$: Perbedaan rata-rata pada data (*i*) ke semua data lain di *cluster* A

$b(i)$: Perbedaan rata-rata pada data (*i*) ke semua data pada *cluster* berbeda.

$s(i)$: Nilai *Silhouette Coefficient.*

Range nilai *s(i)* yakni antara -1 dan 1, interprestasi nilai tersebut yaitu:

$s(i)$: -1: data ke- i digolongkan lemah (dekat pada *cluster* B daripada A)

$s(i)$: 0: data ke- i berada di tengah dua kluster (A dan B)

$s(i)$: 1: data ke- i digolongkan baik

Interpretasi nilai *Silhouette Coefficient* ditunjukan pada tabel 2.1

Tabel 2. 1 Interpretasi Nilai *Silhouette Coefficient*

Silhouette Coefficient	Interpretasi
<= 0.25	Tidak terstruktur
0.26-0.50	Hasil struktur lemah
0.51-0.70	Hasil struktur baik
0.71-1.00	Hasil struktur kuat

2.7. PENELITIAN TERKAIT

Sebagai penguat topik penelitian, dilakukan beberapa analisis dari beberapa penelitian yang ada sebelumnya yang berkaitan dengan topik penelitian.

Berikut tabel 2.2 merupakan beberapa hasil dari penelitian terkait:

Tabel 2. 2 Penelitian terkait

No	Nama penulis/tahun	Judul	Hasil penelitian
1.	Siti amaliyah, 2023	Penerapan data mining untuk menentukan kelompok prioritas penerima bantuan PKH menggunakan *metode clustering K-Means* pada desa kuala dendang	penerapan data mining untuk menentukan prioritas penerima bantuan Program Keluarga Harapan (PKH) dilakukan menggunakan metode Clustering K-Means. Penelitian ini membagi data penduduk Desa Kuala Dendang tahun 2020 menjadi tiga cluster dengan kategori prioritas: cluster 1 (prioritas 1) 114 kepala keluarga (11%), cluster 2 (prioritas 2) 690 kepala keluarga

			(68%), cluster 3 (prioritas 3) 218 kepala keluarga (13%).
2.	Yunan fauzi wijaya, 2024	Implementasi data mining untuk penerima bantuan PKH pemerintah dengan menerapkan algoritma clustering k-medoids	Hasil penelitian dari artikel tersebut menunjukkan bahwa algoritma K-Medoids dapat digunakan untuk mengelompokkan data penerima bantuan Program Keluarga Harapan (PKH). Dalam penelitian ini, ditemukan bahwa: Terdapat 2 cluster yang terbentuk dari pengelompokan data penerima bantuan PKH. Pada cluster 1 terdapat 7 keluarga, sedangkan pada cluster 2 terdapat 8 keluarga.
3.	Siti paridah & martanto, 2024	Klasterisasi penerima dana bantuan program keluarga harapan menggunakan metode K-means pada desa gereba	Penerapan algoritma K-Means clustering untuk menganalisis data program keluarga harapan (PKH) didesa gereba berhasil mengelompokan penerima bantuan menjadi 2 cluster yaitu: cluster 0 terdapat 163 item, cluster 1 terdapat 167 item.
4.	Ibrahim, 2024	Penggunaan algoritma K-Means untuk klasifikasi penerima bantuan PKH sekota banjarmasin	penggunaan algoritma K-Means dalam mengklasifikasikan penerima bantuan Program Keluarga Harapan (PKH) di Kota Banjarmasin berhasil

			mengelompokkan data masyarakat miskin menjadi tiga cluster. Berikut adalah rincian dari masing-masing cluster: cluster 0 terdapat 1529 sebagai prioritas pertama, cluster 1 terdapat 69 orang sebagai prioritas menengah, cluster 2 terdapat 191 orang sebagai prioritas rendah.
5.	Alvendo wahyu aranski, 2024	Pengaplikasian data mining dalam mengelompokan data penerima bantuan subsidi rumah menggunakan metode K-Means Clustering	Hasil dari penelitian tersebut menunjukkan bahwa dari total 170 data yang dianalisis, terdapat 91 orang yang layak untuk menerima bantuan subsidi rumah (dikelompokkan dalam cluster 0) dan 79 orang yang tidak layak (dikelompokkan dalam cluster 1). Penelitian ini menggunakan Algoritma K-Means untuk mengelompokkan data berdasarkan beberapa metrik, seperti jumlah anggota keluarga, pekerjaan, kondisi rumah, dan pendapatan.
6.	Zunaida sitorus, 2024	Penerapan data mining untuk clustering penduduk miskin dikota tanjungbalai menggunakan metode algoritma K-Means	penerapan metode data mining, khususnya algoritma K-Means, berhasil dalam mengklasifikasikan penduduk miskin di Kota Tanjungbalai. Melalui pengolahan dan pengujian data, penelitian ini

			menghasilkan tingkat persentase keakuratan yang tinggi dalam penentuan status kemiskinan, yang diharapkan dapat berdampak positif pada peningkatan taraf hidup masyarakat.
7.	Nirwana hendrastuty, 2024	Penerapan data mining menggunakan algoritma k-means clustering dalam evaluasi hasil pembelajaran siswa	hasil dari pembahasan artikel mengenai penerapan algoritma K-Means Clustering dalam evaluasi hasil pembelajaran siswa: C0 (rajin): 63 siswa, C1 (sangat rajin): 91 siswa. Metrik evaluasi yang digunakan adalah Silhouette Score dan Within-Cluster Sum of Squares (WCSS). Silhouette Score tertinggi yang diperoleh adalah 0,9168, menunjukkan hasil clustering yang baik dan optimal.
8.	Talitha tiara alifa dkk, 2024	Implementasi data mining menggunakan *K-Means Clustering* dalam analisis penjualan produk	pembahasan artikel mengenai penerapan algoritma K-Means Clustering dalam analisis penjualan produk fashion pria di Department Store X. K-Means Clustering terbukti efektif dalam mengidentifikasi kinerja produk. Penelitian ini juga menyarankan untuk mengeksplorasi metode clustering lainnya untuk analisis

			yang lebih baik, mengingat tantangan dalam memanfaatkan data penjualan yang terus meningkat dan pentingnya memahami pola penjualan serta preferensi konsumen.
9.	Agie sidik permana, 2023	Penerapan data minig dalam pengelompokan bahan sembako laris menggunakan K-Means clustering (studi kasus toko gunung bumi)	penerapan algoritma K-Means Clustering berhasil mengelompokkan barang di Toko Gunung Bumi menjadi dua kategori, yaitu barang laris dan tidak laris. Dalam penelitian ini, diperoleh dua iterasi dengan nilai centroid yang berbeda untuk masing-masing kategori. Centroid untuk barang tidak laris adalah 397.950 dan 349.400, sedangkan untuk barang laris adalah 1.248.300 dan 1.272.700. Hasil evaluasi cluster menunjukkan nilai Davies-Bouldin Index sebesar 0.604, yang mengindikasikan bahwa pengelompokan yang dilakukan cukup baik.
10.	Mohammad Ferdiansyah, 2024	Implementasi Algoritma *K-Means* ++ untuk *Clustering* Penjualan Bahan Bangunan	implementasi algoritme K-Means++ untuk clustering penjualan bahan bangunan memberikan hasil yang signifikan. Dalam evaluasi cluster, digunakan dua metode, yaitu Davies Bouldin

			Index (DBI) dan Silhouette Coefficient. Hasil perhitungan DBI menunjukkan bahwa penggunaan K-Means++ menghasilkan nilai 0.5890, yang lebih baik dibandingkan dengan K-Means biasa yang memiliki nilai 0.6795. Ini menunjukkan bahwa K-Means++ lebih efektif dalam mengelompokkan data. Selain itu, analisis menunjukkan bahwa 2 cluster dipilih sebagai centroid terbaik berdasarkan nilai Silhouette Coefficient yang mencapai 0.665, yang menunjukkan struktur cluster yang baik. Hal ini mengindikasikan bahwa dengan menggunakan K-Means++.