BAB II
TINJAUAN PUSTAKA

1.1 Landasan Teori

1.1.1 *Data mining*

Menurut (Han et al., n.d.) *Data mining* umumnya digunakan sebagai proses ekstraksi pola atau pengetahuan yang bermanfaat dari sebuah dataset besar atau kompleks. *Data mining* merupakan kumpulan proses yang menggabungkan Teknik-teknik yang mempunyai tujuan untuk menemukan informasi yang belum diketahui pada data yang sudah dikumpulkan (Nabila et al., 2021)

1. Pengelompokkan Data mining

Menurut (Yuli Mardi, n.d.) *data mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dilakukan sebagai berikut:

1. *Data Selection*

   Pemilihan atau penentuan data dari beberapa sampel data yang dilakukan sebelum memulai fase mengekstraksi data dalam *Knowledge Discovery in Database.*

2. *Pre-processing/Cleaning*

   Cara paling umum untuk membersihkan data sebelum proses *data mining* adalah yang menjadi fokus dalam *Knowledge Discovery in Database.*

3. *Transformation*

   Cara yang paling umum untuk menyatukan data yang telah dipilih, yang sangat tergantung pada jenis atau contoh data yang akan dicari dalam basis data.

4. *Data mining*

Data mining adalah cara untuk menemukan pola atau informasi data yang menarik dari dalam data terpilih dengan menggunakan metode dan algoritma tertentu.

5. *Interpretation/Evaluation*

   Tahap ini mencakup pemeriksaan apakah informasi data yang ditemukan bertentangan dengan kenyataan atau spekulasi sebelumnya.

6. *Knowledge Presentation*

   Tahap ini merupakan memberikan dan menyajikan informasi mengenai metode yang digunakan untuk memberikan informasi kepada pengguna.

1.1.2 *Clustering*

Analisis *cluster* adalah Teknik multivariat yang mempunyai tujuan utama untuk mengelompokkan objek penelitian berdasarkan karakteristik yang dimiliki (A.S Awalluddin & Taufik, 2017). Analisis *Cluster* mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam *cluster* yang sama (Mara & Intisari, 2013)

1. Penentuan Jumlah κ Optimal

Terdapat metode yang selalu digunakan untuk memastikan jumlah κ yang paling sering digunakan yaitu *Elbow Method*, *Silhouette Analysis*, dan *Gap Statistic*.

1) *Elbow Method*

   Elbow Method adalah metode yang digunakan untuk menghasilkan informasi guna menentukan jumlah kelompok yang membentuk siku pada titik tertentu. (Putu et al., 2015) Menurut (Maori, 2023) metode

Elbow digunakan untuk menentukan jumlah optimal dari *cluster* dalam analisis *clustering*. Proses ini melibatkan penentuan jumlah *cluster* κ dengan *Sum of Squared Errors* (*SSE*) untuk setiap nilai κ.

Berikut formula dari *Elbow Method*:

$$\text{SSE} = \sum_{i=1}^{\kappa} \sum_{xi \epsilon C_i} D(x_i, C_i)^2 \quad (2.1)$$

Dimana:

a) κ adalah jumlah *cluster*.

b) $x_i \in C_i$ adalah nilai keanggotaan titik data x_i ke pusat kelompok C_i.

c) C_i adalah pusat *cluster* ke- i.

d) $D\ (x_i, C_i)$ adalah jarak dari titik x_i ke kelompok C_i yang diikuti.

2) Metode *Silhouette Analysis*

Menurut (Elfan, 2023) metode *Silhouette* adalah salah satu alat yang digunakan untuk mengukur seberapa mirip sebuah objek dengan *cluster*nya sendiri dibandingkan dengan *cluster* lainnya. Nilai *silhouette* berkisar antara -1 dan 1. Nilai mendekati 1 menunjukkan bahwa data berada di *cluster* yang tepat, sedangkan nilai mendekati -1 menunjukkan bahwa data mungkin berada di *cluster* salah.

Berikut formula dari *Silhouette Analysis:*

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2.2)$$

Dimana:

a) $s(i)$ adalah nilai *Silhouette* untuk data point i.

b) $a(i)$ adalah jarak rata-rata antara i dan semua poin dalam *cluster*nya sendiri.

c) $b(i)$ adalah jarak rata-rata antara i dan semua poin dalam *cluster* terdekat yang i tidak menjadi bagiannya.

3) *Gap Statistic*

Metode *Gap Statistic* merupakan Teknik yang digunakan untuk menentukan jumlah *cluster* dalam data *clustering*. Metode ini membandingkan kinerja *clustering* pada data asli dengan kinerja pada data acak yang dihasilkan dari distribusi yang sama. Tujuan dari *Gap Statistic* untuk menemukan jumlah *cluster* yang memaksimalkan perbedaan (GAP) antara data asli dan data acak, yang menunjukkan bahwa struktur *clustering* pada data asli lebih signifikan dari pada yang dihasilkan secara acak (Tibshirani et al., 2001). Menurut (Tibshirani et al., 2001) *Gap statistic* digunakan untuk membandingkan total *within-cluster variation* untuk berbagai angka *cluster* dengan ekspektasi *null reference distribution* dari data. Berikut formula *gap statistic*:

$$\text{Gap } (\kappa) = \frac{1}{B} \sum_{b=1}^{B} log\ (W_k) \quad (2.4)$$

Dimana:

a) κ adalah jumlah kluster

b) B adalah jumlah dari referensi *bootstraps*.

c) W_κ adalah *within-cluster dispersion* untuk κ *cluster*.

d) W_κ^b adalah *within-cluster dispersion* dari *bootstrap* b.

Within-cluster dispersion W_κ:

$$W_\kappa = \sum_{i=1}^{\kappa} \frac{1}{2|C_i|} \sum_{x,y \epsilon C_i} ||x - y||^2 \quad (2.5)$$

Dimana:

a) C_i adalah *cluster* i.

b) $|C_i|$ adalah jumlah elemen dalam *cluster* C_i.

c) $||x - y||^2$ adalah jarak kuadrat antara dua titik data x dan y.

1.1.3 Algoritma *K-Medoids*

K-Medoids atau *Partitioning Around Medoids* (PAM) adalah algoritma *clustering* yang mirip dengan *K-Means*. Perbedaan dari kedua algoritma tersebut yaitu PAM menggunakan objek sebagai perwakilan (*medoid*) sebagai pusat *cluster* untuk setiap *cluster*, sedangkan *K-Means* pusat *cluster* menggunakan nilai rata-rata. (Gunawan et al., 2020)

Algoritma *K-Medoids* memiliki kelebihan yaitu untuk mengatasi kelemahan dari *K-Means* yang sensitif terhadap *noise* dan *outlier*, dimana objek dengan nilai yang besar yang memungkinkan menyimpang dari distribusi data. Kelebihan yang lainnya yaitu hasil dari *clustering* tidak bergantung pada urutan dataset (Kamila et al., 2019). Menurut (Dyang et al., 2017) Langkah-langkah algoritma *K-Medoids* yaitu:

1. Inisialisasi atau tentukan pusat *cluster* sebanyak *k* (jumlah *cluster*)
2. Hitung jarak setiap objek *cluster* terdekat dengan menggunakan persamaan *Euclidean Distance*

$$\boldsymbol{d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}} \quad (2.6)$$

3. Pilih secara acak objek *cluster* sebagai *medoid* baru.
4. Hitung jarak setiap objek *cluster* dengan *medoids* baru.

5. Hitung total simpangan (*S*) dengan menghitung *distance* baru – total *distance* lama. Jika S<0, maka tukar objek dengan data *cluster* untuk membentuk sekumpulan *k* objek baru sebagai *medoid*.
6. Ulangi Langkah 3 sampai 5 hingga tidak terjadi perubahan *medoid*, sehingga didapatkan *cluster* beserta anggota *cluster* masing-masing.

1.1.4 Penjualan dan Pemasaran

Penjualan adalah proses pertukaran produk atau jasa dengan uang atau nilai lainnya antara penjual dan pembeli. Ini adalah salah satu fungsi bisnis utama yang bertujuan untuk menghasilkan pendapatan dan mempromosikan pertumbuhan perusahaan. Definisi ini dapat ditemukan dalam berbagai sumber, termasuk buku-buku teks tentang manajemen bisnis, situs web perusahaan, dan publikasi industri (Kotler & Armstrong, n.d.)

Seorang pakar pemasaran (Philip et al., 2021) mendefinisikan pemasaran sebagai proses sosial dan manajerial dimana individu dan kelompok mendapatkan apa yang mereka butuhkan dan inginkan melalui penciptaan, penawaran, dan pertukaran produk yang bernilai dengan pihak lain.

1.1.5 Produk

Produk adalah segala sesuatu yang dapat ditawarkan kepada pasar untuk memenuhi kebutuhan, mengingat bahwa produk tersebut dapat menjadi fisik (barang) atau non-fisik (jasa) (Kotler & Armstrong, n.d.)

1.1.6 *Python*

Python adalah *scripting language* yang berorientasi objek (Manalu & Gunadi, 2022). *Python* memiliki struktur data tingkat tinggi yang efisien dan

pendekatan yang sederhana namun efektif untuk pemrograman berorientasi objek (Python Software Foundation, 2024).

2.2 Penelitian Terkait

2.2.1 State Of The Art

Berdasarkan rumusan masalah dan tujuan penelitian yang telah dibuat, maka dilakukan penyusunan *literature review* dari penelitian sebelumnya yang berkaitan dengan *data mining* metode *clustering* dan algoritma *K-Medoids*. Beberapa penelitian sebelumnya yang dilakukan adalah, sebagai berikut:

Tabel 2. 1 Penelitian Sebelumnya

No	Penulis dan Tahun	Judul	Hasil
1	(Syahfitri et al., 2023)	Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode *Elbow D*an *K-Medoid*	Pada penelitian ini mengelompokkan persediaan barang untuk memudahkan pengelolaan stok toko, ditemukan 7 *cluster* optimal. Tanpa *elbow* data dibagi menjadi 3 *cluster* berdasarkan kuantitas terjual dan tersedia.
2	(Yustika et al., 2021)	Analisis Metode *K-Medoids* Pada Penjualan *Smartphone Vivo* Di Kota Pematangsiantar	Dengan hasil yang sama antara perhitungan manual dengan algoritma K-Medoids dapat diterapkan dalam pengelompokkan *smartphone* Vivo berdasarkan data penjualan dengan persentase keakuratan sebesar 100%.

3	(Robiansyah & Nurcahyo, 2021)	Akurasi Pemberian Intensif Menggunakan Algoritma K- Medoids Terhadap TingkatKedisiplinan Pegawai	Penelitian denganmenggunakan metode ini mendapatkan hasil berupa pengelompokkan 3 yaitu tingkat kedisiplinan baik berjumlah 12 pegawai, tingkat kedisiplinan cukup berjumlah 8 pegawai, tingkat kedisiplinan kurang berjumlah 5 pegawai.
4	(Arifandi et al., 2021)	Implementasi algoritma K-Medoids untuk *clustering* terinfeksi kasus COVID-19 Di DKI Jakarta	Berdasarkan kajian penelitian tersebut dapat disimpulkan bahwa pada pengelompokkan total kasus terinfeksi COVID 19 Berdasarkan kelurahan terdapat 3 *cluster*, yaitu *cluster* 0, *cluster* 1, dan *cluster* 2.
5	(Berliana et al., 2023)	*Clustering* Data Persediaan Barang Menggunakan Metode Elbow dan DBSCAN	Pada penelitian ini metode DBSCAN dengan menggunakan metode Elbow menghasilkan 0,1eps, 1 minPts dan memperoleh hasil 144 *cluster* dan 0 data noise. Sebanyak 144 set tersebut dikumpulkan berdasarkan inventori yang besar. Oleh karena itu, pengujian metode DBSCAN dilakukan pada 0,1eps, 13 minPts tanpa attachment dan memperoleh 3 *cluster* dan 959 data noise. Ketiga kelompok yang terbentuk berbeda-beda pada setiap kelompoknya. Kelompok 1 merupakan produk dengan penjualan

			atau inventori terendah, Kelompok 0 merupakan produk dengan penjualan dan inventori di bawah tengah dibandingkan dengan Kelompok 0, dan Kelompok 2 merupakan produk dengan penjualan dan inventori tertinggi.
6	(Wira et al., 2019)	Implementasi metode *K-Medoids Clustering* untuk mengetahui pola pemilihan program studi mahasiswa baru tahun 2018 di Universitas Kanjuruhan Malang	Penelitian ini menggunakan metode k-medoids agar dapat diketahui pola pemilihan program studi bagi mahasiswa baru di lingkungan Universitas Kanjuruhan Malang.
7	(Maulana & Sundari, 2022)	Penerapan Algoritma K-Medoids Dalam *Cluster*isasi Penyebaran Tempat Ibadah Di Sumatera Utara	Penelitian ini melakukan hasil bahwa proses *clustering* penyebaran rumah ibadah di sumatera utara yang dilakukan dari tahun 2011 sampai dengan tahun 2020 yang dibagi menjadi 5 *cluster* menghasilkan pada *cluster* 1 berjumlah 44 anggota, *cluster* 2 berjumlah 23 anggota, *cluster* 3 berjumlah 49 anggota, *cluster* 4 berjumlah 64, dan *cluster* 5 berjumlah 150 anggota.

8	(Oktavianti Hermadi et al., n.d.)	Implementasi algoritma K-Medoids *Clustering* untuk mencari keuntungan sementara dalam laporan keuangan	Penelitian ini untuk mengelompokkan keuntungan sementara agar nantinya dapat mempermudah untuk memperoleh keputusan dan dipergunakan untuk kepentingan instansi tersebut.
9	(Nurlaela et al., 2020)	Algoritma K-Medoids Untuk *Clustering* Penyakit Maag Di Kabupaten Karawang	Hasil yang didapat pada penelitian ini yaitu dataset penyakit maag di kabupaten karawang pada tahun 2017 sampai 2019 memiliki *cluster* optimal sebanyak 2 *cluster*, dimana *cluster* 1 dengan 35 daerah dikategorikan rendah, dan pada *cluster* 2 dengan 15 daerah yang dikategorikan tinggi, dan menghasilkan nilai silhouette coefficient sebesar 0,5561.
10	(Gustrianda & Mulyana, 2022)	Penerapan *Data mining* dalam pemilihan produk unggulan dengan metode algoritma K-Means dan K-Medoids	Hasil penelitian menunjukkan bahwa nilai Davies Bouldin dari algoritma K-Means adalah -0,430, sedangkan nilai Davies Bouldin K-Means adalah -1,392 yang menunjukkan bahwa nilai Davies Bouldin dari metode K-Means merupakan nilai Davies Bouldin yang paling kecil, sehingga penggunaan metode K-Means untuk

			mengelompokkan hasil merupakan pendekatan yang lebih tepat digunakan untuk permasalahan ini.
11	(Ayu et al., 2019)	Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma *Clustering* K-Medoids dalam Pengelompokan Produksi Kerajinan Bali	Analisis perbandingan algoritma k-Medoid *cluster*ing metode Elbow dan Silhouette menunjukkan bahwa wilayah Bali merupakan wilayah dengan kegiatan kerajinan terbanyak. Pada tahun 2017, jumlah industri air rumah tangga di wilayah Bali sebanyak 5.574. Proses *cluster*ing K-Medoid dengan metode penghitungan Silhouette menunjukkan hasil yang lebih baik. Hal ini dikarenakan nilai DBI proses perakitan K-Medoid memiliki rata-rata koefisien *silhouette* sebesar 1,06, lebih rendah dibandingkan nilai DBI proses perakitan K-Medoid dengan metode Elbow sebesar 1,10.
12	(Prasetyaningrum & Susanti, 2023)	Perbandingan Algoritma K-Means Dan K-Medoids Untuk Pemetaan Hasil Produksi Buah-Buahan	Hasil yang didapatkan pada penelitian ini yaitu algoritma K-means mendapat hasil nilai yang lebih kecil dengan nilai 0,296 sedangkan hasil algoritma K-Medoids dengan nilai lebih besar 0,507.

			Algoritma terbaik untuk *cluster*isasi hasil produksi buah-buahan yang ada di kabupaten Kotawaringin Timur adalah algoritma K-Means berdasarkan hasil nilai DBI yang diperoleh.
13	(Nurina Sari & Primajaya, 2019)	Penerapan *Clustering DBSCAN* Untuk Pertanian Padi Di Kabupaten Karawang	Ketika teknologi klasterisasi DBSCAN diaplikasikan pada lahan sawah, dihasilkan dua *cluster* dengan karakteristik yang berbeda. Hasil klasterisasi menunjukkan adanya perbedaan curah hujan, luas lahan yang mempengaruhi produksi, jumlah serangan hama, dan jenis hama yang menyerang lahan pertanian. Perbandingan evaluasi kinerja teknologi *cluster*isasi DBSCAN dilakukan dengan berfokus pada nilai rata-rata lebar bayangan. Ukuran bidang siluet menggambarkan kualitas massa yang terbentuk. Hasil percobaan penelitian ini menunjukkan hasil tertinggi sebesar 0,74 yang terbagi menjadi dua kelompok. Hal ini menunjukkan bahwa blok yang terbentuk memiliki struktur yang kuat.
14	(Pratikto & Damastuti, 2018)	*Cluster*isasi Menggunakan Agglomerative Hierarchical	Pada penelitian ini menggunakan metode *Agglomerative Hierarchical Clustering* (AHC) untuk memodelkan wilayah banjir

		Clustering Untuk Memodelkan Wilayah Banjir	di jawa timur, terdapat 3 wilayah banjir di jawa timur, terdapat 3 *cluster* berdasarkan potensi terjadinya banjir yaitu rendah, sedang dan tinggi, kinerja *cluster* dievaluasi menggunakan *cophenetic correlation coefficient*, dengan metode *average linkage* memberikan solusi terbaik hasil studi divisualisasikan dalam bentuk Sistem Informasi Geografis (gis), agar memudahkan identifikasi daerah berpotensi banjir.

2.2.2 Matriks Penelitian

Tabel 2.2 merupakan matriks penelitian yang berfokus untuk menyelesaikan masalah pada metode dan algoritma. Selain itu, matriks ini dapat memberikan informasi tentang perbedaan penelitian yang akan dilakukan dan penelitian terdahulu.

Tabel 2. 2 Matriks Penelitian

No	**Penulis dan Tahun**	**Ruang Lingkup**									
		Algoritma		**Objek Penelitian**				**Metode**			
		K-Means	***K-Medoids***	***DBSCAN***	***Hierarchical Clustering***	**Karyawan**	**Implementasi**	***Elbow Method***	***Silhouette Analysis***	***Gap Statistic***	***Davies Bouldin Index***

1	(Rindiawan, 2024)	-	√	-	-	√	-	√	√	√	-
2	(Yustika et al., 2021)	-	√	-	-	-	√	-	-	√	-
3	(Syahfitri et al., 2023)	-	√	-	-	√	-	√	-	-	-
4	(Robiansyah & Nurcahyo, 2021)	-	√	-	-	-	√	-	-	√	-
5	(Arifandi et al., 2021)	-	√	-	-	-	√	√	-	√	-
6	(Berliana et al., 2023)	-	-	√	-	-	√	√	-	-	-
7	(Wira et al., 2019)	-	√	-	-	√	-	-	√	-	-
8	(Fialine et al., 2021)	-	√	-	-	-	√	-	√	-	-
9	(Oktavianti Hermadi et al., n.d.)	-	√	-	-	-	√	-	-	-	√
10	(Nasari et al., 2023)	-	√	-	-	-	√	-	-	-	√
11	(Ayu et al., 2019)	-	√	-	-	√	-	√	√	-	√
12	(Prasetyaningrum & Susanti, 2023)	√	√	-	-	-	√	-	-	-	√
13	(Nurina Sari & Primajaya, 2019)	-	-	√	-	-	√	-	√	-	-
14	(Pratikto & Damastuti, 2018)	-	-	-	√	-	√	√	-	-	-
15	(Gustrianda & Mulyana, 2022)	-	√	-	-	√	-	-	-	-	√