

BAB III

METODE PENELITIAN

3.1 Objek dan Subjek Penelitian

Objek penelitian yang digunakan dalam penelitian ini adalah data transaksi penjualan minimarket tahun 2021–2022 yang diperoleh dari *Kaggle*. *Dataset* tersebut berisi informasi mengenai data penjualan barang yang digunakan untuk proses analisis *clustering*.

Adapun subjek penelitian ini adalah produk atau barang yang terdapat pada data transaksi penjualan minimarket. Produk tersebut akan dianalisis berdasarkan jumlah transaksi dan tingkat penjualannya untuk mengetahui pola penjualan barang.

3.2 Ruang Lingkup Penelitian

Ruang lingkup penelitian ini difokuskan pada proses clustering data transaksi penjualan minimarket menggunakan algoritma K-Means dan K-Medoids. Penelitian dilakukan untuk mengelompokkan data penjualan berdasarkan tingkat penjualan produk sehingga dapat diketahui kategori produk dengan penjualan lambat, sedang, dan cepat.

Selain itu, penelitian ini menggunakan metode *Elbow* untuk menentukan jumlah *cluster* optimal serta *Davies-Bouldin Index (DBI)* untuk mengevaluasi kualitas hasil *clustering*.

3.3 Batasan Penelitian

Adapun Batasan penelitian dalam penelitian ini adalah sebagai berikut:

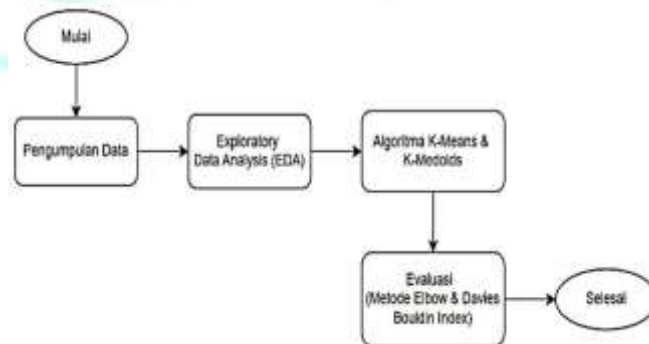
1. Data yang digunakan merupakan data transaksi penjualan minimarket tahun 2021–2022.
2. Penelitian hanya menggunakan algoritma K-Means dan K-Medoids dalam proses *clustering*.
3. Penentuan jumlah klaster menggunakan metode *Elbow*.
4. Evaluasi hasil clustering menggunakan *Davies-Bouldin Index (DBI)*.
5. Penelitian difokuskan pada pengelompokan data penjualan produk berdasarkan tingkat penjualan.

3.4 Prosedur Penelitian

Prosedur penelitian dilakukan melalui beberapa tahapan yang disusun secara sistematis untuk mencapai tujuan penelitian. Tahapan penelitian dimulai dari proses pengumpulan data, analisis data awal menggunakan *Exploratory Data Analysis (EDA)*,

proses *clustering* menggunakan algoritma K-Means dan K-Medoids, hingga evaluasi hasil *clustering* menggunakan metode *Elbow* dan *Davies-Bouldin Index (DBI)*.

Alur tahapan penelitian tersebut digambarkan dalam bentuk *flowchart* untuk mempermudah pemahaman proses penelitian yang dilakukan. *Flowchart* prosedur penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 *Flowchart* Penelitian

Gambar tersebut menunjukkan diagram alir dari proses analisis data untuk *clustering* menggunakan algoritma K-Means dan K-Medoids. Berikut adalah penjelasan langkah-langkah dalam diagram tersebut :

1. Mulai : Proses dimulai dari titik ini.
2. Pengumpulan Data : Langkah pertama adalah mengumpulkan data yang akan digunakan dalam analisis. Data ini bisa berupa data mentah yang akan diproses dan dianalisis lebih lanjut.
3. *Exploratory Data Analysis (EDA)* : Setelah data terkumpul, dilakukan analisis eksploratif untuk memahami karakteristik data, mengidentifikasi pola, menangani nilai yang hilang, serta membersihkan data. Tahap ini penting untuk memastikan bahwa data siap digunakan untuk proses *clustering*.
4. Algoritma K-Means & K-Medoids : Setelah EDA selesai, data dianalisis menggunakan algoritma klustering K-Means dan K-Medoids. K-Means membagi data menjadi beberapa *cluster* berdasarkan jarak ke *centroid*, sementara K-Medoids memilih titik pusat yang benar-benar ada di dalam data sebagai representasi setiap *cluster*.
5. Evaluasi (Metode *Elbow* & *Davies-Bouldin Index*) : Setelah proses *clustering*, evaluasi dilakukan menggunakan Metode *Elbow* dan *Davies-Bouldin Index* untuk menilai kualitas *clustering*. Metode *Elbow* membantu menentukan jumlah *cluster*

optimal dengan melihat titik "siku" pada grafik yang menunjukkan penurunan jarak dalam *cluster*. *Davies-Bouldin Index* digunakan untuk mengukur seberapa baik *cluster* dibentuk dan seberapa jauh antar *cluster*.

6. Selesai : Proses analisis data selesai. Diagram ini menunjukkan alur sistematis untuk melakukan klustering, dari pengumpulan data hingga evaluasi hasil *clustering* untuk memastikan bahwa hasil akhir memiliki kualitas yang baik.

3.5 Pengumpulan Data

Pada tahap pengumpulan data, akan digunakan data yang relevan dengan penelitian. Dataset pada penelitian ini yaitu data transaksi penjualan di sebuah minimarket pada tahun 2021-2022 dan data ini diakses pada tanggal 22 januari pada pukul 11.00 WIB melalui *platform* *web* di <https://www.kaggle.com/datasets/ipunguhbpwt/clsuterpenjualan>. Dataset ini terdiri dari 7403 baris dan memiliki 5 variabel yaitu kode_barang, nama_barang, jumlah_transaksi, total_penjualan dan rata_rata.

3.6 Exploratory Data Analysis (EDA)

Data yang telah diperoleh kemudian dianalisis melalui proses *Exploratory Data Analysis (EDA)*. Proses ini mencakup pencarian pola, memeriksa apakah terdapat *missing value*, *outlier* dan duplikat dalam data. Langkah selanjutnya adalah melakukan visualisasi data untuk memperoleh informasi sebelum data tersebut diproses lebih lanjut.

3.7 Clustering (Algoritma K-Means)

Pada penelitian ini, melibatkan penerapan model *machine learning* dengan memanfaatkan algoritma *K-Means Clustering*, yang termasuk salah satu teknik *unsupervised learning*. Pada tahapan ini, hasil klasterisasi di evaluasi dengan membagi pelanggan ke dalam segmen-segmen tertentu. Proses segmentasi pelanggan dimulai dari persiapan data, dilanjutkan dengan seleksi data yang akan di klaster, dan menetapkan jumlah klaster dengan menggunakan metode *Elbow*. Berikut adalah tahapan-tahapan dalam penerapan algoritma K-Means :

- a. Menentukan nilai k untuk total klaster yang akan dihasilkan
- b. Menentukan pusat klaster awal yang akan dibentuk.

c. Mengukur jarak antara objek pada *centroid* menggunakan *Euclidean Distance*.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

Dimana D_e merupakan *Euclidean Distance*, i merupakan banyaknya data, (x, y) merupakan titik data dan (s, t) merupakan titik pusat. Menghitung jarak *Euclidean* antara dua titik dalam ruang dua dimensi, yaitu titik pertama dengan koordinat (x_i, y_i) dan titik kedua dengan koordinat (s_i, t_i) . Dalam rumus ini D_e merepresentasikan jarak *Euclidean* antara kedua titik tersebut. Komponen $(x_i - s_i)^2$ dan $(y_i - t_i)^2$ masing-masing mewakili perbedaan kuadrat antara koordinat x dan y dari kedua titik. Setelah menghitung perbedaan kuadrat antara koordinat x dan y , hasilnya dijumlahkan, kemudian diakarkan untuk mendapatkan jarak lurus antara kedua titik. Jarak *Euclidean* ini sering digunakan dalam analisis data dan pemrosesan citra untuk mengukur kedekatan atau perbedaan antara dua titik dalam ruang dua dimensi, yang sering kali berguna dalam klasifikasi, klustering, dan pengenalan pola.

d. Kelompokkan data ke dalam klaster berdasarkan jarak terdekat.

e. Melakukan perulangan, lalu menentukan posisi titik pusat (*centroid*) baru dengan menggunakan rumus berikut :

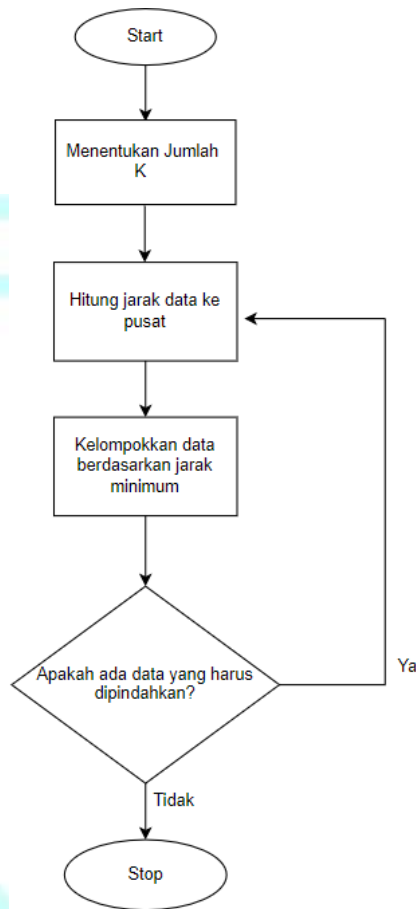
$$\bar{v} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj}$$

Dimana $v_i j$ merupakan rata-rata titik pusat pada klaster ke- i untuk variabel j , N_i merupakan banyaknya anggota dalam klaster ke- i , i, k merupakan Indikator dari klaster, j merupakan Indikator dari variabel dan x_{kj} merupakan Nilai data ke- k pada klaster tersebut untuk variabel ke- j

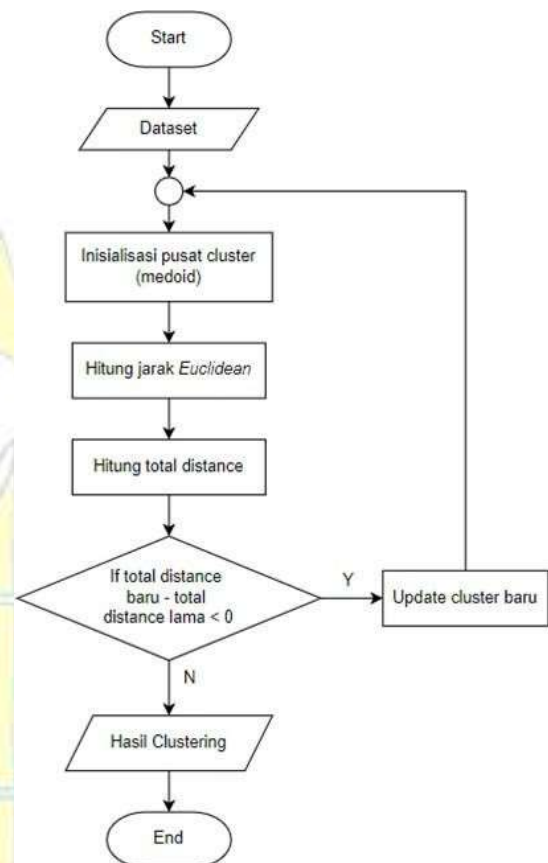
Rumus diatas digunakan untuk menghitung rata-rata nilai dari sekumpulan data tertentu. Dalam rumus ini, $\bar{v}$ melambangkan nilai rata-rata dari semua data yang diamati, sedangkan menunjukkan jumlah total elemen atau data dalam himpunan tersebut. Simbol $\sum_{k=0}^n x_{kj}$ menyatakan penjumlahan dari setiap nilai x_{kj} dalam himpunan, di mana adalah

indeks yang menunjuk pada setiap elemen mulai dari 0 hingga N_i . Setelah semua nilai dalam himpunan dijumlahkan, hasilnya dibagi dengan jumlah elemen untuk memperoleh rata-rata $\bar{v}$. Rumus ini sering digunakan dalam analisis statistik atau data untuk menemukan nilai rata-rata dari sejumlah besar data, sehingga memudahkan dalam memahami karakteristik umum dari data tersebut.

f. Ulangi langkah 3 jika posisi *centroid* yang baru berbeda.



Gambar 3.2 Flowchart K-Means



Gambar 3.3 Flowchart K-Medoid

Gambar 3.2 menunjukkan alur kerja algoritma K-Means Clustering. Proses dimulai dengan menentukan jumlah kluster (K) yang diinginkan. Selanjutnya, jarak tiap data ke pusat kluster dihitung, dan data dikelompokkan berdasarkan jarak minimum ke pusat kluster tersebut. Algoritma kemudian mengecek apakah ada data yang perlu dipindahkan ke kluster lain. Jika ada, pusat kluster diperbarui, dan proses diulang. Jika tidak ada data yang perlu dipindahkan lagi, klustering selesai.

Lalu, pada Gambar 3.3 menunjukkan alur proses algoritma K-Medoids untuk klustering data. Proses dimulai dengan dataset yang sudah disiapkan. Pertama, pusat kluster (medoid) diinisialisasi secara acak. Kemudian, jarak *Euclidean* dari setiap titik data ke medoid dihitung, diikuti dengan perhitungan total jarak (total *distance*). Jika total jarak baru lebih kecil dari jarak sebelumnya, maka kluster diperbarui

dengan medoid baru dan perhitungan diulang. Proses ini terus berulang sampai tidak ada pengurangan dalam total jarak, menghasilkan klustering akhir yang optimal.

3.8 Clustering (Algoritma K-Medoids)

Algoritma K-Medoids adalah metode klasik dalam teknik partisi *clustering* yang membagi dataset yang terdiri dari n objek menjadi k *cluster* yang sudah ditentukan sebelumnya. Algoritma ini bekerja berdasarkan prinsip untuk meminimalkan kesamaan antara setiap objek dengan titik acuan yang tepat. Berikut adalah langkah-langkah untuk menyelesaikan algoritma K-Medoids :

- a. Menginisialisasi pusat kluster dengan jumlah k (jumlah kluster).
- b. Menentukan atau menghitung jarak (cost) memakai rumus *Euclidean Distance* sebagai berikut :

$$d_{\text{Euclidean}}(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$
- c. Tentukan objek secara acak di setiap kluster sebagai pilihan medoid baru.
- d. Menghitung jarak antara masing-masing objek di setiap kluster dengan kandidat medoid yang baru.
- e. Tentukan jumlah simpangan (S) dengan cara menghitung selisih jumlah *distance* yang baru dan jumlah *distance* yang lama. Apabila $S < 0$, ubah objek dengan data kluster untuk membentuk k objek baru sebagai medoid.
- f. Kembali ke langkah 3 sampai 5 hingga medoid tidak lagi berubah, sehingga dihasilkan kluster beserta anggotanya masing-masing.

3.9 Evaluasi

3.9.1 Metode *Elbow*

Metode *Elbow* diterapkan untuk menginterpretasikan dan menguji konsistensi performa jumlah kluster yang optimal dengan mempertimbangkan nilai SSE. Dengan memanfaatkan hasil perbandingan jumlah kluster yang terbentuk di titik tertentu, metode *elbow* berpotensi untuk membantu menentukan jumlah kluster yang terbaik (Sari et al., 2023). Pada titik tertentu, grafik akan menunjukkan penurunan yang signifikan dengan adanya suatu lekukan yang dikenal sebagai "siku". Penentuan nilai k yang optimal dilakukan dengan membandingkan nilai SSE (*Sum of Square Error*)

yang ditampilkan pada grafik (Yulistiani et al., 2024). Nilai tersebut kemudian menjadi jumlah kluster optimal atau k yang terbaik. Dengan meningkatnya jumlah kluster K , nilai SSE cenderung menurun. Rumus SSE untuk algoritma K-Means dapat ditemukan dalam formula 2.

$$SSE = \sum_{K=1}^k \sum_{X_i \in C_k} |X_i - C_k|^2$$

Keterangan K merupakan jumlah kluster yang digunakan, X_i merupakan jumlah data ke- i , C_k merupakan banyaknya kluster i pada kluster ke- k , $|$ merupakan menghitung jarak *Euclidean*.

3.9.2 Davies-Bouldin Index (DBI)

Evaluasi yang diterapkan untuk menilai seberapa baik kluster yang dihasilkan oleh algoritma K-Means & K-Medoids adalah *Davies-Bouldin Index*. Semakin kecil nilai dari DBI, maka semakin baik hasil *cluster* tersebut (Kamilah et al., 2024). Evaluasi dengan menggunakan *Davies-Bouldin Index* melibatkan skema evaluasi kluster internal, dimana kualitas kluster dinilai berdasarkan jumlah serta kedekatan antara *cluster* hasil. *Davies-Bouldin Index* adalah salah satu teknik yang diterapkan untuk mengukur keakuratan hasil kluster dari sebuah teknik *clustering*. *Cohesion* didefinisikan sebagai jumlah data yang berhubungan dengan titik pusat kluster yang diikuti oleh kluster tersebut. Sementara itu, pemisahan diukur berdasarkan jarak antara titik pusat kluster satu dengan yang lainnya. Penilaian dengan *Davies-Bouldin Index* ini berusaha mengoptimalkan jarak antar kluster, yaitu antara kluster C_i dan C_j , secara bersamaan berusaha meminimalkan jarak antara titik dalam satu kluster. Apabila jarak antara kluster berada pada tingkat maksimal, ini menunjukkan bahwa persamaan karakteristik di antara kluster adalah rendah, sehingga perbedaan di antara kluster menjadi lebih terlihat. Sebaliknya, apabila jarak dalam satu kluster minimal, hal ini menunjukkan bahwa setiap objek dalam kluster tersebut memiliki tingkat kesamaan karakteristik yang tinggi (Mughnyanti et al., 2020)