

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian yang Relevan

Penelitian mengenai pengelompokan data kelistrikan telah dilakukan dengan menggunakan berbagai pendekatan *clustering*. (Wang dkk., 2022) membandingkan algoritma *K-Means* dan DBSCAN untuk mengekstraksi pola permintaan listrik berdimensi tinggi, sedangkan (Okereke dkk., 2023) menerapkan *K-Means* pada data *smart meter* untuk mengelompokkan konsumen berdasarkan fitur penggunaan listrik dalam domain waktu. Penelitian lain menggunakan *Fuzzy C-Means* untuk mengidentifikasi pola perilaku konsumsi listrik dan *Mini Batch K-Means* untuk mengelompokkan data pendistribusian listrik pada tingkat provinsi (Mulyadi dkk., 2024; Zhu dkk., 2024). Rangkaian penelitian tersebut menunjukkan bahwa metode *clustering* dapat digunakan untuk menemukan kelompok berdasarkan kemiripan data kelistrikan, meskipun objek, variabel, dan unit analisis yang digunakan berbeda.

Penerapan *K-Means* juga berkembang pada bidang selain kelistrikan. (Ardana dkk., 2024) menerapkannya untuk segmentasi pelanggan penjualan daring, sedangkan (Anggari dkk., 2024) mengelompokkan negara berdasarkan sejumlah indikator ekonomi. (Wenny, 2024) menggabungkan normalisasi *Min-Max* dan *K-Means* pada data kependudukan, sementara (Akmal dkk., 2025) serta (Ariansah dkk., 2025) menerapkan proses serupa untuk mengelompokkan wilayah berdasarkan sejumlah indikator numerik. Penelitian-penelitian tersebut relevan secara metodologis karena menunjukkan pentingnya pemilihan variabel, penyamaan skala data, penentuan jumlah *cluster*, dan evaluasi hasil pengelompokan.

Untuk memperlihatkan perkembangan metode, perbedaan objek, dan posisi penelitian ini secara sistematis, perbandingan penelitian disajikan pada Tabel 2.1. Tabel tersebut mencakup sepuluh penelitian terdahulu dan satu baris penelitian yang dilakukan oleh Aisyah Tri Nurfajri.

Tabel 2.1 Penelitian Terdahulu

No.	Peneliti (Tahun)	Metode yang Digunakan	Hasil Utama	Perbedaan dengan Penelitian Ini
1	Wang et al. (2022)	Perbandingan <i>Means</i>	<i>K- Kedua</i> dan menghasilkan	algoritma Menggunakan pola permintaan data listrik

No.	Peneliti (Tahun)	Metode yang Digunakan	Hasil Utama	Perbedaan dengan Penelitian Ini
		DBSCAN pada data permintaan listrik berdimensi tinggi.	pengelompokan yang berbeda. Kesesuaian metode dipengaruhi oleh karakteristik, kepadatan, dan dimensi data.	konsumen dengan atribut berdimensi tinggi, bukan data agregat ULP berdasarkan daya terpasang, produksi listrik, dan jumlah pelanggan.
2	Okereke et al. (2023)	<i>K-Means Clustering</i> pada fitur domain waktu yang diperoleh dari data <i>smart meter</i> .	Konsumen listrik dapat dikelompokkan yang berdasarkan kemiripan pola penggunaan energi yang terekam melalui <i>smart meter</i> .	Berfokus pada perilaku konsumsi konsumen individual, sedangkan penelitian ini mengelompokkan unit pelayanan pada tingkat ULP.
3	Zhu et al. (2024)	<i>Fuzzy C-Means</i> dan evaluasi menggunakan <i>Davies-Bouldin Index</i> serta koefisien profil.	Diperoleh empat <i>cluster</i> pola konsumsi dari sejumlah jenis pengguna listrik baru, seperti penyimpanan energi, kendaraan listrik, dan irigasi.	Menggunakan <i>Fuzzy C-Means</i> dan data konsumsi beberapa jenis pengguna, bukan <i>K-Means</i> pada data agregat ULP.
4	Mulyadi et al. (2024)	<i>Mini Batch K-Means</i> dan <i>Silhouette Score</i> pada data provinsi pendistribusian listrik tahun 2014–2022.	Data pendistribusian listrik pada sejumlah provinsi dapat dikelompokkan berdasarkan kesamaan pola distribusinya.	Unit analisis berada pada tingkat provinsi dan dapat menggunakan <i>Mini Batch K-Means</i> , sedangkan penelitian ini menggunakan 16 ULP serta <i>K-Means</i> konvensional.
5	Ardana et al. (2024)	<i>K-Means Clustering</i> pada data pelanggan penjualan daring.	Pengelompokan menghasilkan lima segmen pelanggan berdasarkan kemiripan karakteristik data pelanggan.	Objeknya pelanggan perdagangan daring dan berorientasi pada pemasaran, bukan data kelistrikan tingkat ULP.

No.	Peneliti (Tahun)	Metode yang Digunakan	Hasil Utama	Perbedaan dengan Penelitian Ini
6	Anggari et al. (2024)	<i>Min-Max Scaling, K-Means, Elbow Method, Silhouette Score, DBI, dan Calinski-Harabasz Index.</i>	Negara-negara dikelompokkan menjadi tiga kelompok berdasarkan GDP, ekspor, impor, inflasi, dan pengangguran, dengan nilai Silhouette sebesar 0,52.	Kesamaannya terdapat pada penggunaan beberapa variabel numerik dan normalisasi, tetapi objeknya berupa negara dan indikator ekonomi.
7	Wenny (2024)	Normalisasi <i>Min-Max</i> dan algoritma <i>K-Means</i> pada data kependudukan.	Data kependudukan dimanfaatkan untuk membentuk kelompok ekonomi tingkat masyarakat berdasarkan sejumlah kriteria.	Menggunakan data individu atau kependudukan untuk tingkat klasifikasi ekonomi, bukan data kelistrikan pada unit pelayanan.
8	Fourteen et al. (2025)	<i>K-Means</i> dengan evaluasi <i>Silhouette Score</i> dan <i>inertia</i> pada lima dataset berbeda.	Kualitas hasil pengelompokan berbeda menurut karakteristik dataset, jumlah <i>cluster</i> , parameter, dan proses persiapan data. Nilai <i>Silhouette</i> tertinggi yang diperoleh sebesar 0,6946.	Tidak berfokus pada satu objek penelitian tertentu, tetapi memberikan dasar penting mengenai evaluasi internal dan pengaruh karakteristik data.
9	Akmal et al. (2025)	<i>Min-Max Scaling, K-Means, dan Elbow Method</i> pada indikator pembangunan manusia.	Tiga kelompok kabupaten/kota terbentuk berdasarkan empat indikator Indeks Pembangunan Manusia.	Serupa dalam pengelompokan unit wilayah berdasarkan beberapa variabel, tetapi objek dan indikator yang digunakan bukan kelistrikan.
10	Ariansah et al. (2025)	Pembersihan data, normalisasi, deteksi <i>outlier</i> , <i>K-Means</i> ,	Dua <i>cluster</i> industri kecil dan menengah terbentuk dengan nilai <i>Silhouette</i>	Menggunakan data industri periode 2021–2023, sedangkan penelitian ini

No.	Peneliti (Tahun)	Metode yang Digunakan	Hasil Utama	Perbedaan dengan Penelitian Ini
		<i>Elbow Method</i> , sebesar 0,70 dan <i>Silhouette</i> , dan DBI. sebesar 0,47.		DBI menggunakan data satu tahun pada 16 ULP.
11	Penelitian ini (2026)	Penggabungan dua dataset, <i>Min-Max Scaling</i> , <i>K-Means Clustering</i> , <i>Elbow Method</i> , dan <i>Silhouette Coefficient</i> .	Mengelompokkan 16 ULP menjadi tiga <i>cluster</i> berdasarkan daya dataset tahun 2022 dan produksi berfokus pada segmentasi jumlah tingkat ULP di DKI Model Jakarta. menghasilkan WCSS sebesar 0,9848 dan <i>Silhouette</i> sebesar 0,4005.	Menggunakan kombinasi tiga variabel dari dua data dataset tahun 2022 dan ULP di DKI

Berdasarkan Tabel 2.1, penelitian terdahulu pada bidang kelistrikan lebih banyak menggunakan pola konsumsi atau permintaan listrik pada tingkat pelanggan, data *smart meter*, maupun data distribusi pada skala provinsi. Metode yang digunakan juga beragam, seperti *K-Means*, DBSCAN, *Fuzzy C-Means*, dan *Mini Batch K-Means*. Perbedaan tersebut menunjukkan bahwa hasil *clustering* sangat dipengaruhi oleh objek penelitian, jenis variabel, struktur data, dan algoritma yang digunakan.

Dari sisi metode, penelitian terdahulu memperlihatkan bahwa proses normalisasi dan evaluasi internal merupakan bagian penting dalam penerapan *K-Means*. Normalisasi digunakan untuk mencegah dominasi variabel yang memiliki nilai numerik lebih besar, sedangkan *Elbow Method*, *Silhouette Coefficient*, DBI, dan *inertia* digunakan untuk menilai jumlah serta kualitas *cluster*. Meskipun demikian, tidak seluruh penelitian menggabungkan lebih dari satu ukuran evaluasi atau membahas kualitas setiap anggota *cluster* secara lebih rinci.

Berdasarkan celah tersebut, penelitian ini melakukan segmentasi pada tingkat Unit Pelayanan Pelanggan di DKI Jakarta melalui penggabungan dua dataset tahun 2022. Penelitian menggunakan tiga variabel, yaitu daya terpasang, produksi listrik, dan jumlah pelanggan, yang dinormalisasi menggunakan *Min-Max Scaling*. Jumlah *cluster* dipertimbangkan melalui *Elbow Method* dan *Silhouette Coefficient*, kemudian karakteristik hasil dianalisis berdasarkan centroid,

rata-rata setiap variabel, keanggotaan ULP, dan nilai Silhouette. Posisi penelitian ini terletak pada kombinasi objek tingkat ULP, penggabungan sumber data, penggunaan tiga variabel, serta evaluasi hasil pengelompokan secara eksploratif.

2.2 Landasan Teori

Landasan teori pada penelitian ini disusun untuk memberikan dasar konseptual terhadap setiap tahapan analisis yang digunakan dalam proses segmentasi Unit Pelayanan Pelanggan (ULP). Pembahasan diawali dengan konsep *data mining* dan kerangka CRISP-DM sebagai dasar pelaksanaan penelitian, kemudian dilanjutkan dengan *flowchart* untuk menjelaskan representasi alur proses, konsep *clustering*, serta algoritma *K-Means* sebagai metode utama dalam pembentukan kelompok data. Selain itu, Python dijelaskan sebagai perangkat pemrograman yang digunakan untuk menjalankan proses pengolahan, pemodelan, evaluasi, dan visualisasi data penelitian.

Selanjutnya, landasan teori membahas *Min-Max Scaling* sebagai metode normalisasi yang digunakan untuk menyamakan rentang variabel daya terpasang, produksi listrik, dan jumlah pelanggan. Ketiga variabel tersebut dijelaskan secara terpisah agar kedudukannya sebagai dasar pengelompokan ULP dapat dipahami secara jelas. Pembahasan kemudian diakhiri dengan teori evaluasi *clustering* melalui *Elbow Method* dan *Silhouette Coefficient*, yang digunakan untuk mempertimbangkan jumlah *cluster* serta menilai kualitas hasil pengelompokan. Dengan susunan tersebut, setiap teori memiliki keterkaitan langsung dengan prosedur penelitian pada Bab III dan hasil analisis yang disajikan pada Bab IV,

2.2.1 Data Mining

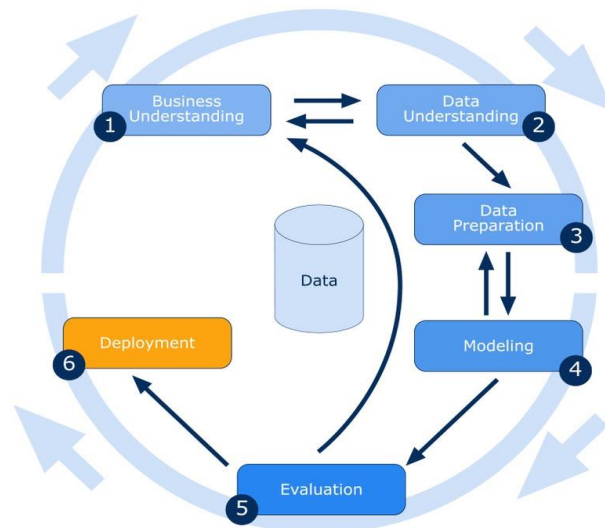
Data mining merupakan proses pengolahan dan analisis data untuk menemukan pola, hubungan, serta informasi yang belum terlihat secara langsung pada data awal. Proses tersebut tidak hanya dilakukan dengan menerapkan algoritma, tetapi juga mencakup pemilihan data, pemeriksaan kualitas, transformasi, pemodelan, dan penafsiran hasil. Dengan demikian, *data mining* digunakan untuk mengubah sekumpulan data menjadi informasi yang lebih terstruktur dan dapat dianalisis sesuai dengan tujuan penelitian. (Annas & Wahab, 2023) menjelaskan bahwa teknologi *data mining* dapat digunakan untuk mengolah kumpulan data dan memperoleh informasi yang bernilai melalui proses pengelompokan berdasarkan kesamaan karakteristik.

Berdasarkan tujuan analisisnya, *data mining* mencakup beberapa fungsi, seperti deskripsi, klasifikasi, prediksi, dan *clustering*. Deskripsi digunakan untuk menjelaskan karakteristik data, sedangkan klasifikasi dan prediksi memerlukan pola atau kelas tertentu untuk memperkirakan kategori maupun nilai. Berbeda dengan fungsi tersebut, *clustering* mengelompokkan objek berdasarkan kemiripan atribut tanpa menggunakan label kelas yang telah ditentukan sebelumnya. Pendekatan ini memungkinkan struktur kelompok ditemukan secara langsung dari pola yang terdapat dalam data. (Kurniawan dkk., 2026) menjelaskan bahwa *clustering* merupakan teknik *machine learning* untuk mengelompokkan data yang mempunyai kemiripan karakteristik sehingga pola segmentasi dapat dianalisis secara lebih sistematis.

Dalam penelitian ini, penerapan *data mining* dimulai dengan menggunakan dua dataset kelistrikan tahun 2022 yang memiliki struktur dan atribut berbeda. Dataset pertama memuat data daya terpasang dan produksi listrik, sedangkan dataset kedua memuat jumlah pelanggan pada setiap Unit Pelayanan Pelanggan (ULP). Kedua dataset diperiksa, dipilih atribut yang sesuai, dirapikan, dan digabungkan berdasarkan nama ULP. Selanjutnya, ketiga variabel dinormalisasi menggunakan *Min-Max Scaling* karena memiliki satuan serta rentang nilai yang berbeda sebelum digunakan dalam pemodelan.

Setelah data dinyatakan siap, proses pengelompokan dilakukan menggunakan algoritma *K-Means Clustering* untuk memperoleh kelompok ULP berdasarkan kemiripan daya terpasang, produksi listrik, dan jumlah pelanggan. Penerapan *data mining* pada penelitian ini tidak diarahkan untuk melakukan prediksi atau klasifikasi, tetapi untuk menemukan pola segmentasi dari 16 ULP tanpa menggunakan label awal. Hasilnya digunakan untuk menjelaskan keanggotaan, pusat, karakteristik, dan kualitas setiap *cluster* secara eksploratif. Oleh karena itu, informasi yang diperoleh dibatasi pada ketiga variabel penelitian dan tidak digunakan secara langsung untuk menilai kinerja teknis maupun menetapkan kebijakan operasional pada masing-masing ULP.

2.2.2 Cross Industry Standard Process for Data Mining (CRISP-DM)



Gambar 2.1 CRISP-DM

Cross-Industry Standard Process for Data Mining atau CRISP-DM merupakan model proses yang digunakan untuk mengarahkan pelaksanaan proyek *data mining* secara sistematis. Model ini bersifat independen terhadap bidang penerapan dan membagi proses analisis ke dalam enam fase, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. (Schröer dkk., 2021) menjelaskan bahwa CRISP-DM masih banyak digunakan sebagai model proses dalam penelitian *data mining* karena memberikan tahapan yang jelas sejak perumusan kebutuhan analisis hingga penyajian hasil.

Keenam fase CRISP-DM tidak selalu berjalan secara satu arah. Hasil pemeriksaan pada suatu fase dapat mengharuskan peneliti kembali ke fase sebelumnya, misalnya ketika ditemukan ketidaksesuaian struktur data pada tahap persiapan atau ketika hasil pemodelan belum memenuhi pertimbangan evaluasi. Sifat iteratif tersebut memungkinkan setiap keputusan mengenai data, metode, dan hasil diperiksa kembali agar tetap sesuai dengan tujuan penelitian. Dalam penelitian ini, CRISP-DM digunakan untuk mengatur proses segmentasi 16 Unit Pelayanan Pelanggan (ULP) di DKI Jakarta berdasarkan daya terpasang, produksi listrik, dan jumlah pelanggan. Tahapan yang diterapkan dijelaskan sebagai berikut:

1. *Business Understanding*

Business understanding merupakan tahap untuk menetapkan permasalahan, tujuan analisis, hasil yang diharapkan, dan batasan penelitian. Tahap ini diperlukan agar proses pengolahan data tidak hanya berorientasi pada penerapan algoritma, tetapi tetap menjawab kebutuhan penelitian yang telah dirumuskan. Dalam penelitian ini, kebutuhan analisis dalam penelitian ini didefinisikan sebagai kebutuhan untuk memperoleh ringkasan profil dan kelompok pembandingan antar-ULP berdasarkan tiga variabel yang tersedia. Kebutuhan tersebut tidak berasal dari permintaan formal PLN, tetapi dirumuskan berdasarkan kondisi data yang menunjukkan adanya perbedaan nilai daya terpasang, produksi listrik, dan jumlah pelanggan antar-ULP. Oleh karena itu, keluaran yang ditetapkan berupa segmentasi eksploratif, bukan rekomendasi kebijakan operasional.

2. *Data Understanding*

Setelah tujuan ditetapkan, data understanding dilakukan untuk mengenali sumber, struktur, atribut, satuan, dan kualitas data yang tersedia. Tahap ini mencakup pembacaan data, pemeriksaan jumlah baris dan kolom, pengenalan tipe data, serta identifikasi awal terhadap nilai kosong, data ganda, atau perbedaan penulisan identitas objek. Pada penelitian ini, pemeriksaan dilakukan terhadap dua dataset tahun 2022. Dataset pertama memuat informasi kelistrikan pada setiap ULP, sedangkan dataset kedua memuat jumlah pelanggan. Nama ULP digunakan sebagai atribut penghubung untuk memastikan bahwa data dari kedua sumber dapat dipasangkan dengan benar.

3. *Data Preparation*

Berdasarkan pemahaman terhadap data, tahap data preparation dilakukan untuk membentuk dataset akhir yang siap digunakan dalam pemodelan. Kegiatan pada tahap ini meliputi pemilihan atribut, perapian struktur data, pemeriksaan kualitas, penggabungan dataset, dan transformasi nilai. Dalam penelitian ini, atribut yang dipilih terdiri atas daya terpasang, produksi listrik, dan jumlah pelanggan. Baris agregat DKI Jakarta dikeluarkan karena tidak merepresentasikan satu ULP, sedangkan kedua dataset digabungkan berdasarkan kesesuaian nama unit. Ketiga variabel kemudian dinormalisasi menggunakan Min-Max Scaling ke dalam rentang 0 sampai 1 agar perbedaan satuan dan besarnya nilai tidak menyebabkan salah satu variabel mendominasi perhitungan jarak.

4. *Modeling*

Tahap modeling dilakukan dengan memilih dan menerapkan algoritma yang sesuai dengan tujuan analisis. Karena penelitian bertujuan menemukan kelompok tanpa menggunakan label kelas awal, algoritma yang digunakan adalah K-Means Clustering. Pemodelan dilakukan terhadap data hasil normalisasi dengan menggunakan jarak Euclidean sebagai ukuran kedekatan antara data dan pusat kelompok atau centroid. Proses K-Means berlangsung secara iteratif melalui penentuan anggota cluster dan pembaruan centroid sampai susunan kelompok mencapai kondisi konvergen. Konfigurasi pemodelan ditetapkan secara konsisten agar proses dapat dijalankan kembali menggunakan data dan parameter yang sama.

5. *Evaluation*

Setelah pemodelan dilakukan, tahap evaluation digunakan untuk menilai struktur pengelompokan serta memastikan bahwa hasilnya masih sesuai dengan tujuan penelitian. Evaluasi tidak dilakukan menggunakan ukuran klasifikasi seperti accuracy atau confusion matrix karena penelitian tidak memiliki label kelas sebenarnya. Penentuan jumlah cluster dilakukan dengan mempertimbangkan Elbow Method melalui nilai Within-Cluster Sum of Squares dan Silhouette Coefficient melalui tingkat kekompakan serta keterpisahan kelompok. Hasil kedua metode dibandingkan sebelum jumlah cluster pada model akhir ditetapkan. Selain nilai evaluasi keseluruhan, kualitas setiap kelompok dan posisi masing-masing anggota juga diperiksa agar hasil segmentasi dapat dijelaskan secara hati-hati.

6. *Deployment*

Deployment merupakan tahap penyajian hasil analisis agar keluaran pemodelan dapat dipahami dan diperiksa kembali. Dalam penelitian ini, tahap tersebut tidak berupa penerapan model ke dalam aplikasi operasional PLN. Hasil segmentasi disajikan melalui tabel data, grafik evaluasi, nilai centroid, keanggotaan setiap ULP, profil masing-masing cluster, dan visualisasi sebaran data. Keluaran pengolahan juga disimpan dalam format CSV dan PNG, sedangkan rangkaian analisis didokumentasikan melalui notebook Google Colab. Penyajian tersebut digunakan untuk memberikan gambaran eksploratif mengenai karakteristik kelompok berdasarkan ketiga variabel penelitian, bukan sebagai dasar langsung dalam menetapkan kebijakan operasional.

Dengan penerapan keenam fase tersebut, CRISP-DM menjadi penghubung antara tujuan penelitian, kondisi data, proses persiapan, pemodelan, evaluasi, dan penyajian hasil. Setelah kerangka proses penelitian dipahami, pembahasan selanjutnya diarahkan pada konsep *clustering* sebagai teknik pengelompokan yang digunakan pada fase *modeling*.

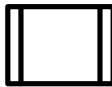
2.2.3 Flowchart

Flowchart merupakan representasi grafis yang digunakan untuk menggambarkan urutan proses, hubungan antaraktivitas, serta keputusan yang terdapat dalam suatu prosedur. Setiap tahapan ditampilkan menggunakan simbol dengan bentuk dan fungsi tertentu, kemudian dihubungkan melalui garis berarah untuk menunjukkan perpindahan dari satu proses ke proses berikutnya. Penyajian secara visual membantu pembaca memahami alur penyelesaian masalah tanpa harus terlebih dahulu membaca seluruh uraian prosedural secara terperinci. Smrti et al. (2023) menjelaskan bahwa *flowchart* dapat digunakan untuk menggambarkan tahapan penyelesaian masalah secara terstruktur sehingga urutan logika lebih mudah dipahami.

Selain berfungsi sebagai media visualisasi, *flowchart* membantu menyusun proses secara logis sebelum tahapan tersebut diterapkan. Hubungan antarproses, urutan pelaksanaan, kondisi percabangan, serta kemungkinan pengulangan dapat diamati melalui diagram yang sama. Zhang et al. (2023) menunjukkan bahwa penyusunan *flowchart* secara bertahap dapat mendukung pemahaman terhadap struktur proses dan pengembangan kemampuan berpikir komputasional. Temuan tersebut memperkuat penggunaan *flowchart* sebagai media untuk menjelaskan proses yang terdiri atas beberapa tahapan dan keputusan yang saling berkaitan.

Dalam penelitian ini, *flowchart* digunakan untuk memvisualisasikan tahapan segmentasi Unit Pelayanan Pelanggan (ULP) berdasarkan kerangka CRISP-DM. Alur tersebut mencakup identifikasi masalah, pemeriksaan dua dataset, persiapan dan penggabungan data, normalisasi, pemodelan menggunakan *K-Means Clustering*, evaluasi jumlah *cluster*, serta penyajian hasil. Diagram juga menunjukkan alur kembali ke tahap sebelumnya apabila data belum memenuhi pemeriksaan atau hasil evaluasi masih memerlukan penyesuaian. Dengan demikian, *flowchart* pada Bab III tidak menggantikan penjelasan metode, tetapi membantu pembaca melihat hubungan antartahapan penelitian secara menyeluruh. Agar simbol yang digunakan dapat dipahami secara konsisten, Tabel 2.2 menjelaskan simbol-simbol yang digunakan dalam *flowchart* penelitian.

Tabel 2.2 Simbol Flowcart

No	Simbol	Nama	Keterangan
1		<i>Flow Direction</i>	Menunjukkan arah jalannya proses dari satu langkah ke langkah berikutnya.
2		<i>Terminator</i>	Digunakan untuk menandai awal dan akhir dari suatu proses.
3		<i>Connector</i>	Berfungsi untuk menghubungkan bagian flowchart yang terpisah agar terstruktur
4		<i>Process</i>	Menunjukkan langkah pengolahan data atau aktivitas yang dilakukan dalam sistem.
5		<i>Decision</i>	Menyatakan kondisi tertentu yang menghasilkan percabangan alur (ya/tidak).
6		<i>Input/Output</i>	Digunakan untuk menunjukkan proses input data atau output hasil
7		<i>Preparation</i>	Menunjukkan tahap persiapan sebelum proses utama dilakukan.
8		<i>Predefine Proses</i>	Menunjukkan proses tambahan atau prosedur yang sudah didefinisikan sebelumnya.
9		<i>Display</i>	Menunjukkan hasil yang ditampilkan ke layar atau media output.
10		<i>Data Storage</i>	Menunjukkan penyimpanan data dalam sistem atau database.

Berdasarkan Tabel 2.2, simbol *terminator* digunakan sebagai batas yang memperjelas awal dan akhir diagram. Di antara kedua batas tersebut, simbol *process* menjadi simbol yang paling banyak digunakan karena sebagian besar alur penelitian berupa kegiatan pengolahan data. Sementara itu, simbol *input/output* membedakan data yang masuk dan keluaran yang dihasilkan dari kegiatan pengolahan, sehingga pembaca dapat mengenali posisi dataset serta hasil penelitian secara lebih jelas.

Selanjutnya, simbol *decision* diperlukan karena alur penelitian tidak seluruhnya berjalan secara linear. Pemeriksaan kualitas data dapat menghasilkan keputusan untuk melanjutkan proses

atau kembali melakukan perbaikan. Hal yang sama berlaku pada tahap evaluasi jumlah *cluster*, ketika hasil *Elbow Method* dan *Silhouette Coefficient* perlu diperiksa sebelum model akhir ditetapkan. Setiap cabang pada simbol keputusan harus diberi keterangan “Ya” dan “Tidak” agar arah proses tidak menimbulkan penafsiran ganda.

Adapun simbol *flow direction* menghubungkan seluruh tahapan dan menunjukkan arah pelaksanaan penelitian. Arah panah yang bergerak maju menunjukkan kelanjutan proses, sedangkan panah yang kembali menunjukkan adanya pemeriksaan atau pengulangan.

2.2.4 Clustering

Clustering merupakan teknik dalam *data mining* yang digunakan untuk mengelompokkan sejumlah objek berdasarkan kemiripan karakteristik yang terdapat pada data. Berbeda dengan klasifikasi, proses *clustering* tidak menggunakan label kelas yang telah ditentukan sebelumnya sehingga termasuk dalam pendekatan *unsupervised learning*. Kelompok dibentuk secara langsung dari pola kedekatan antarobjek berdasarkan variabel yang dianalisis. (Alasali & Ortakci, 2024) menjelaskan bahwa *clustering* digunakan untuk mengidentifikasi struktur kelompok dalam data dengan menempatkan objek yang memiliki karakteristik serupa ke dalam kelompok yang sama.

Tujuan utama *clustering* adalah menghasilkan kelompok dengan tingkat kemiripan yang tinggi di dalam satu *cluster* dan perbedaan yang jelas antar-*cluster*. Kedekatan objek dalam kelompok disebut *intra-cluster similarity*, sedangkan keterpisahan antar kelompok disebut *inter-cluster dissimilarity*. Dengan prinsip tersebut, hasil pengelompokan yang baik diharapkan memiliki anggota yang relatif dekat dengan kelompoknya sendiri dan relatif jauh dari kelompok lainnya. Meskipun demikian, kualitas kelompok tetap dipengaruhi oleh karakteristik data, skala variabel, ukuran jarak, jumlah *cluster*, dan algoritma yang digunakan. Karakteristik utama *clustering* dapat dijelaskan sebagai berikut:

1. Kesamaan dalam cluster (*intra-cluster similarity*)

Objek yang ditempatkan dalam satu cluster memiliki nilai atau pola karakteristik yang relatif serupa. Semakin dekat posisi antarobjek dalam kelompok yang sama, semakin kompak kelompok tersebut.

2. Perbedaan antar-cluster (*inter-cluster dissimilarity*)

Setiap cluster diharapkan memiliki karakteristik yang dapat dibedakan dari kelompok lainnya. Perbedaan tersebut diperlukan agar kelompok yang terbentuk tidak saling tumpang tindih secara berlebihan.

3. Tidak menggunakan label awal

Keanggotaan kelompok tidak ditentukan berdasarkan kategori yang telah tersedia sebelumnya. Algoritma mengidentifikasi kelompok dari struktur dan hubungan yang terdapat pada data numerik yang dianalisis.

4. Bersifat eksploratif

Hasil clustering digunakan untuk menemukan pola dan struktur kelompok yang sebelumnya belum diketahui. Oleh karena itu, nama dan interpretasi setiap cluster diberikan setelah karakteristik anggotanya dianalisis, bukan sebelum proses pengelompokan dilakukan.

Berdasarkan cara pembentukan kelompoknya, metode *clustering* dapat dibedakan menjadi metode berbasis partisi, hierarki, kepadatan, dan model lainnya. Metode berbasis partisi membagi data ke dalam sejumlah kelompok yang telah ditentukan, seperti pada algoritma *K-Means*. Metode hierarki membentuk hubungan bertingkat antarobjek yang umumnya disajikan melalui *dendrogram*, sedangkan metode berbasis kepadatan membentuk kelompok berdasarkan wilayah data yang memiliki kepadatan tinggi. (Anjarwati dkk., 2025) menunjukkan bahwa pemilihan metode *clustering* perlu mempertimbangkan pola distribusi, kepadatan, dan kemungkinan adanya *outlier*, karena setiap algoritma memiliki karakteristik yang berbeda dalam membentuk kelompok.

Dalam bidang kelistrikan, teknik *clustering* dapat diterapkan untuk mengenali kelompok data yang memiliki profil serupa. (Kusuma dkk., 2025) menggunakan pendekatan *clustering* untuk menganalisis pola konsumsi energi listrik dan menunjukkan bahwa proses pengelompokan dapat membantu memperlihatkan perbedaan karakteristik antardata. Namun, hasil kelompok tetap harus diinterpretasikan sesuai dengan variabel dan unit analisis yang digunakan serta tidak dapat langsung digunakan untuk menyimpulkan kondisi teknis yang tidak tersedia dalam data.

Penerapan metode *clustering* pada data kelistrikan juga dilakukan oleh (Kusuma dkk., 2024) menggunakan metode *Fuzzy C-Means*. Penelitian tersebut mengelompokkan pola penggunaan energi pada pelanggan sosial, seperti tempat ibadah, lembaga pendidikan, dan fasilitas kesehatan, berdasarkan variabel daya aktif dan daya semu. Meskipun metode serta objek

penelitiannya berbeda, hasil tersebut menunjukkan bahwa teknik *clustering* dapat digunakan untuk mengenali perbedaan karakteristik kelompok pelanggan listrik berdasarkan variabel numerik yang tersedia.

Dalam penelitian ini, *clustering* diterapkan untuk mengelompokkan 16 Unit Pelayanan Pelanggan (ULP) di DKI Jakarta berdasarkan tiga variabel, yaitu daya terpasang, produksi listrik, dan jumlah pelanggan. Ketiga variabel tersebut digunakan secara bersamaan untuk menemukan ULP yang memiliki profil data relatif serupa. Proses pengelompokan tidak ditujukan untuk memprediksi kelas, mengukur efisiensi, atau menilai kinerja operasional ULP, melainkan menghasilkan segmentasi eksploratif yang selanjutnya dianalisis berdasarkan nilai *centroid*, keanggotaan, karakteristik, dan kualitas setiap *cluster*.

Dengan demikian, konsep *clustering* menjadi dasar bagi pemilihan algoritma *K-Means* dalam penelitian ini. Sebelum membahas proses normalisasi dan pemodelan, subbab berikutnya menjelaskan daya terpasang, produksi listrik, dan jumlah pelanggan sebagai variabel yang digunakan dalam pembentukan kelompok.

2.2.5 Daya Terpasang (kVA)

Daya terpasang merupakan kapasitas daya listrik yang tercatat pada suatu instalasi atau wilayah pelayanan dan dinyatakan dalam satuan volt-ampere (VA) atau kilovolt-ampere (kVA). Dalam sistem listrik arus bolak-balik, kVA digunakan untuk menunjukkan kapasitas daya semu yang mencakup komponen daya aktif dan daya reaktif. Penggunaan satuan kVA juga umum ditemukan pada kapasitas transformator dan kapasitas sambungan listrik karena menunjukkan kemampuan peralatan atau sistem dalam menyalurkan daya pada batas tertentu (Sutama & Candra, 2023).

Daya terpasang perlu dibedakan dari energi listrik yang dinyatakan dalam kilowatt-hour (kWh). Nilai kVA menunjukkan kapasitas daya, sedangkan kWh menunjukkan jumlah energi listrik yang tercatat selama suatu periode. Oleh karena itu, ULP dengan daya terpasang tinggi tidak selalu memiliki produksi listrik atau jumlah pelanggan yang tinggi. Nilai tersebut juga tidak dapat digunakan secara langsung untuk menyimpulkan tingkat konsumsi, efisiensi, beban puncak, maupun kondisi teknis jaringan karena kesimpulan tersebut memerlukan variabel tambahan.

Pada tingkat Unit Pelayanan Pelanggan, daya terpasang dapat digunakan sebagai salah satu karakteristik numerik untuk membandingkan skala kapasitas yang tercatat antarunit. Penelitian

(Hardiyan & Wahyudin, 2021) juga menggunakan data daya terpasang bersama informasi produksi listrik dan pelanggan pada ULP di DKI Jakarta, yang menunjukkan bahwa variabel tersebut dapat digunakan dalam analisis berbasis data pada tingkat unit pelayanan. Namun, pemaknaan daya terpasang tetap harus disesuaikan dengan tujuan dan metode penelitian yang digunakan.

Dalam penelitian ini, daya terpasang digunakan sebagai salah satu dari tiga variabel pembentuk *cluster*, bersama produksi listrik dan jumlah pelanggan. Nilai yang digunakan merupakan total daya terpasang dalam satuan kVA pada masing-masing dari 16 ULP berdasarkan dataset tahun 2022. Variabel tersebut diperlakukan sebagai atribut numerik yang menggambarkan perbedaan kapasitas daya antar-ULP dan tidak digunakan secara terpisah untuk menilai kinerja masing-masing unit.

Karena nilai daya terpasang memiliki rentang yang berbeda dengan produksi listrik dan jumlah pelanggan, data tersebut dinormalisasi menggunakan *Min-Max Scaling* sebelum diterapkan pada algoritma *K-Means Clustering*. Normalisasi diperlukan agar daya terpasang memberikan kontribusi yang sebanding dengan dua variabel lainnya dalam perhitungan jarak. Setelah daya terpasang dijelaskan sebagai variabel kapasitas, subbab berikutnya membahas produksi listrik sebagai variabel energi yang tercatat pada masing-masing ULP.

2.2.6 Produksi Listrik (kWh)

Produksi listrik merupakan jumlah energi listrik yang dicatat selama periode tertentu dan dinyatakan dalam satuan kilowatt-hour (kWh). Dalam penelitian ini, istilah produksi listrik mengikuti nama variabel pada dataset statistik kelistrikan Unit Pelayanan Pelanggan (ULP) di DKI Jakarta. Oleh karena itu, variabel tersebut dipahami sebagai nilai produksi listrik yang tercatat pada masing-masing ULP selama tahun 2022, bukan sebagai kapasitas pembangkit atau jumlah energi yang secara langsung dikonsumsi oleh setiap pelanggan. Tabel statistik BPS juga menempatkan produksi listrik sebagai variabel yang berbeda dari listrik terjual dan listrik yang dipakai sendiri.

Satuan kWh menunjukkan besarnya energi listrik yang terbentuk dari penggunaan daya aktif selama jangka waktu tertentu. Secara konseptual, hubungan antara energi, daya aktif, dan waktu dapat dinyatakan melalui persamaan berikut:

$$E = P \times t \quad (2.1)$$

Keterangan:

- A. E = energi listrik dalam kilowatt-hour (kWh);
- B. P = daya aktif dalam kilowatt (kW);
- C. t = waktu penggunaan dalam jam.

Persamaan tersebut menunjukkan bahwa kWh merupakan satuan energi, sehingga berbeda dengan kVA yang digunakan untuk menyatakan kapasitas daya semu. Sebagai contoh konseptual, peralatan dengan daya aktif 1 kW yang digunakan selama 2 jam memerlukan energi sebesar 2 kWh. Meskipun demikian, penelitian ini tidak menghitung produksi listrik menggunakan persamaan tersebut karena nilai produksi telah tersedia secara langsung dalam dataset.

Pada tingkat ULP, nilai produksi listrik dapat digunakan untuk memperlihatkan perbedaan besarnya energi listrik yang tercatat antarunit pelayanan dalam periode yang sama. (Hardiyan & Wahyudin, 2021) menggunakan produksi listrik bersama variabel kelistrikan lainnya dalam analisis unit pelayanan di wilayah DKI Jakarta. Hal tersebut menunjukkan bahwa produksi listrik dapat digunakan sebagai atribut numerik dalam membandingkan karakteristik antar-ULP. Namun, pemaknaannya tetap harus dibatasi sesuai data yang tersedia karena produksi listrik tidak secara langsung menjelaskan jenis pelanggan, pola pemakaian harian, beban puncak, kondisi jaringan, maupun efisiensi suatu ULP.

Produksi listrik juga tidak dapat langsung disamakan dengan konsumsi pelanggan. Konsumsi merupakan energi yang digunakan oleh pelanggan, sedangkan dataset penelitian mencantumkan produksi listrik, listrik terjual, dan listrik yang dipakai sendiri sebagai atribut yang berbeda. Dalam pemodelan penelitian ini, hanya variabel produksi listrik yang digunakan, sedangkan listrik terjual dan listrik yang dipakai sendiri berada di luar ruang lingkup analisis. Oleh karena itu, hasil pengelompokan tidak digunakan untuk menghitung konsumsi per pelanggan atau menilai tingkat pemanfaatan kapasitas listrik.

Dalam penelitian ini, produksi listrik menjadi salah satu dari tiga variabel pembentuk *cluster*, bersama daya terpasang dan jumlah pelanggan. Nilai produksi listrik setiap ULP diperlakukan sebagai data numerik yang menggambarkan salah satu karakteristik kelistrikan selama tahun 2022. Karena rentang nilainya mencapai ratusan juta hingga miliaran kWh dan jauh lebih besar daripada dua variabel lainnya, produksi listrik dinormalisasi menggunakan *Min-Max Scaling*. Proses tersebut dilakukan agar besarnya nilai numerik produksi listrik tidak mendominasi perhitungan jarak pada algoritma *K-Means Clustering*.

Dengan demikian, produksi listrik digunakan untuk membedakan profil ULP berdasarkan nilai energi yang tercatat, bukan untuk memberikan penilaian teknis terhadap proses pembangkitan, konsumsi, atau efisiensi pelayanan. Setelah variabel daya terpasang dan produksi listrik dijelaskan, subbab berikutnya membahas jumlah pelanggan sebagai variabel ketiga yang digunakan dalam proses segmentasi.

2.2.7 Jumlah Pelanggan

Jumlah pelanggan merupakan banyaknya pelanggan listrik yang tercatat pada suatu Unit Pelayanan Pelanggan (ULP) dalam periode tertentu. Dalam penelitian ini, variabel tersebut mengacu pada data tahun 2022 dalam dataset “Jumlah Pelanggan Listrik Menurut Unit Pelayanan Pelanggan di Provinsi DKI Jakarta, 2022”. Nilainya dinyatakan dalam satuan pelanggan dan digunakan untuk menunjukkan perbedaan cakupan pelanggan pada masing-masing ULP.

Jumlah pelanggan perlu dibedakan dari jumlah penduduk atau jumlah pengguna listrik secara langsung karena satu pelanggan dapat merepresentasikan satu sambungan yang digunakan oleh lebih dari satu individu. Data tersebut juga tidak memuat informasi mengenai golongan tarif, jenis pelanggan, kapasitas sambungan, maupun pola konsumsi. Oleh karena itu, jumlah pelanggan diperlakukan sebagai data agregat dan tidak digunakan untuk menjelaskan kondisi sosial, konsumsi energi, atau kinerja pelayanan secara langsung.

Dalam analisis kelistrikan, jumlah pelanggan dapat digunakan sebagai variabel numerik untuk membedakan karakteristik antarunit pelayanan. (Hardiyan & Wahyudin, 2021) menggunakan jumlah pelanggan bersama variabel kelistrikan lainnya pada analisis tingkat ULP di DKI Jakarta. Meskipun demikian, jumlah pelanggan perlu dianalisis bersama daya terpasang dan produksi listrik karena ULP dengan pelanggan lebih banyak belum tentu memiliki kapasitas daya atau produksi listrik yang lebih tinggi.

Dalam penelitian ini, jumlah pelanggan digunakan bersama daya terpasang dan produksi listrik sebagai variabel pembentuk *cluster*. Sebelum diterapkan pada algoritma *K-Means Clustering*, ketiga variabel dinormalisasi menggunakan *Min-Max Scaling* agar memiliki rentang yang sebanding dalam perhitungan jarak. Dengan demikian, jumlah pelanggan berfungsi untuk menggambarkan cakupan pelanggan setiap ULP, sedangkan proses penyamaan skala ketiga variabel dibahas pada subbab normalisasi data.

2.2.8 Normalisasi Data

Normalisasi data merupakan proses transformasi nilai numerik ke dalam rentang yang sebanding tanpa mengubah urutan relatif data pada setiap variabel. Tahap ini diperlukan ketika variabel yang dianalisis memiliki satuan dan rentang nilai yang berbeda, terutama pada algoritma berbasis jarak seperti *K-Means Clustering*. Tanpa normalisasi, variabel dengan nilai numerik lebih besar dapat memberikan pengaruh lebih dominan dalam perhitungan jarak, meskipun tingkat kepentingannya tidak lebih tinggi daripada variabel lainnya. Wenny (2024) menunjukkan bahwa normalisasi *Min-Max* dapat digunakan sebelum proses *K-Means* untuk menyamakan skala beberapa atribut numerik.

Salah satu metode yang umum digunakan adalah *Min-Max Scaling*, yaitu mengubah nilai setiap variabel ke dalam rentang 0 sampai 1. Proses transformasi dilakukan secara terpisah pada masing-masing variabel menggunakan nilai minimum dan maksimum yang terdapat dalam data. Rumus *Min-Max Scaling* dinyatakan sebagai berikut:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (2.2)$$

Nilai normalisasi sebesar 0 menunjukkan bahwa data tersebut merupakan nilai terendah pada variabel yang dianalisis, sedangkan nilai 1 menunjukkan nilai tertinggi. Nilai tersebut tidak berarti bahwa objek tidak memiliki daya terpasang, produksi listrik, atau pelanggan, tetapi hanya menunjukkan posisi relatifnya terhadap data lain. Transformasi ini tetap mempertahankan urutan nilai asli, meskipun hasilnya dapat dipengaruhi oleh keberadaan nilai ekstrem karena batas skala ditentukan oleh nilai minimum dan maksimum. Anggari et al. (2024) menerapkan *Min-Max Scaling* sebelum proses *K-Means* agar sejumlah indikator dengan skala berbeda dapat dianalisis secara sebanding.

Dalam penelitian ini, normalisasi diterapkan secara terpisah pada variabel daya terpasang, produksi listrik, dan jumlah pelanggan dari 16 ULP. Ketiga variabel tersebut memiliki satuan serta rentang yang berbeda, sehingga seluruh nilai diubah ke dalam rentang 0 sampai 1 sebelum digunakan dalam penentuan jumlah *cluster* dan pemodelan *K-Means*. Perhitungan model menggunakan nilai normalisasi dengan presisi penuh, sedangkan pembulatan hanya dilakukan ketika hasil disajikan dalam tabel. Setelah ketiga variabel memiliki skala yang sebanding, proses penelitian dilanjutkan dengan penerapan algoritma *K-Means Clustering*.

2.2.9 Algoritma *K-Means*

Algoritma *K-Means* merupakan metode *partitioning clustering* yang membagi sekumpulan data ke dalam sejumlah *cluster* berdasarkan kemiripan nilai antarobjek. Jumlah kelompok, yang dinyatakan dengan k , ditentukan sebelum proses pemodelan dilakukan. Setiap *cluster* direpresentasikan oleh sebuah pusat kelompok atau *centroid*, sedangkan setiap objek ditempatkan ke dalam *cluster* yang memiliki jarak terdekat terhadap *centroid*. (Annas & Wahab, 2023) menjelaskan bahwa *K-Means* banyak digunakan karena memiliki tahapan yang relatif sederhana dan dapat diterapkan pada data numerik untuk menemukan pola pengelompokan tanpa menggunakan label kelas awal.

Proses *K-Means* bertujuan membentuk kelompok yang anggotanya memiliki jarak relatif dekat terhadap *centroid* masing-masing. Sebelum pemodelan dilakukan, variabel yang digunakan perlu memiliki skala sebanding karena perbedaan rentang nilai dapat memengaruhi hasil perhitungan jarak. Oleh karena itu, dalam penelitian ini daya terpasang, produksi listrik, dan jumlah pelanggan terlebih dahulu dinormalisasi menggunakan *Min-Max Scaling*. Data hasil normalisasi kemudian diproses menggunakan ukuran jarak Euclidean untuk menentukan kedekatan setiap ULP terhadap seluruh *centroid*.

Jarak Euclidean antara data ke- i dan *centroid cluster* ke- j dapat dihitung menggunakan persamaan berikut:

$$d(x_i, \mu_j) = \sqrt{\sum_{p=1}^m (x_{ip} - \mu_{jp})^2} \quad (2.3)$$

Keterangan:

- A. $d(x_i, \mu_j)$ = jarak data ke- i terhadap *centroid cluster* ke- j ;
- B. x_{ip} = nilai variabel ke- p pada data ke- i ;
- C. μ_{jp} = nilai variabel ke- p pada *centroid cluster* ke- j ;
- D. m = jumlah variabel yang digunakan.

Dalam penelitian ini, nilai m adalah tiga karena proses pengelompokan menggunakan daya terpasang, produksi listrik, dan jumlah pelanggan. Setiap ULP dihitung jaraknya terhadap seluruh *centroid*, kemudian ditempatkan ke dalam kelompok dengan nilai jarak terkecil. Dengan demikian,

penentuan anggota *cluster* tidak dilakukan berdasarkan satu variabel tertentu, tetapi berdasarkan kombinasi posisi ketiga variabel dalam ruang data hasil normalisasi.

Setelah seluruh data memperoleh kelompok, *centroid* diperbarui dengan menghitung nilai rata-rata setiap variabel dari seluruh anggota dalam *cluster* tersebut. Pembaruan nilai *centroid* dapat dinyatakan melalui persamaan berikut:

$$\mu_{jp} = \frac{1}{n_j} \sum_{x_i \in C_j} x_{ip} \quad (2.4)$$

Keterangan:

- A. μ_{jp} = nilai *centroid* pada *cluster* ke- j untuk variabel ke- p ;
- B. n_j = jumlah anggota pada *cluster* ke- j ;
- C. C_j = himpunan anggota *cluster* ke- j ;
- D. x_{ip} = nilai variabel ke- p pada data ke- i .

Perhitungan jarak dan pembaruan *centroid* dilakukan secara berulang hingga model mencapai kondisi konvergen. Kondisi tersebut terjadi ketika susunan keanggotaan dan posisi *centroid* tidak lagi mengalami perubahan yang berarti atau ketika jumlah iterasi telah mencapai batas maksimum. Secara sistematis, proses algoritma *K-Means* terdiri atas tahapan berikut:

1. Menentukan jumlah cluster
Nilai k ditentukan sesuai kebutuhan analisis dan hasil evaluasi beberapa alternatif jumlah kelompok. Dalam penelitian ini, jumlah cluster dipertimbangkan melalui Elbow Method dan Silhouette Coefficient.
2. Menentukan centroid awal
Pusat kelompok awal ditentukan sebelum proses pengalokasian data. Penelitian ini menggunakan metode K-Means++ agar posisi centroid awal tersebar lebih terarah dan tidak terkumpul pada bagian data yang sama.
3. Menghitung jarak data terhadap centroid
Jarak setiap data dihitung terhadap seluruh centroid menggunakan Euclidean Distance. Perhitungan dilakukan menggunakan nilai normalisasi dengan presisi penuh.
4. Menentukan keanggotaan cluster

Setiap data dialokasikan ke dalam cluster yang memiliki jarak paling kecil. Tahap ini disebut sebagai proses assignment.

5. Memperbarui nilai centroid

Setelah seluruh data memperoleh kelompok, nilai centroid dihitung kembali berdasarkan rata-rata setiap variabel dari anggota cluster.

6. Mengulangi proses hingga konvergen

Tahap penghitungan jarak, penentuan anggota, dan pembaruan centroid dilakukan kembali sampai susunan kelompok mencapai kondisi stabil.

Tujuan matematis dari algoritma *K-Means* adalah meminimalkan jumlah kuadrat jarak antara setiap data dan *centroid cluster* yang ditempatinya. Ukuran tersebut disebut *Within-Cluster Sum of Squares* (WCSS) dan dirumuskan sebagai berikut:

$$WCSS = \sum_{j=1}^k \sum_{x_i \in C_j} d(x_i, \mu_j)^2 \quad (2.5)$$

Keterangan:

- A. $WCSS$ = jumlah kuadrat jarak dalam seluruh *cluster*;
- B. k = jumlah *cluster*;
- C. C_j = himpunan data pada *cluster* ke- j ;
- D. x_i = data ke- i ;
- E. μ_j = *centroid cluster* ke- j .

Nilai WCSS yang lebih kecil menunjukkan bahwa data berada semakin dekat dengan *centroid* kelompoknya. Meskipun demikian, nilai WCSS akan terus menurun ketika jumlah *cluster* ditambah, sehingga nilai terkecil tidak dapat langsung digunakan sebagai satu-satunya dasar dalam menentukan jumlah kelompok. Oleh karena itu, perubahan WCSS perlu dianalisis melalui *Elbow Method* dan dibandingkan dengan ukuran evaluasi lain, seperti *Silhouette Coefficient*.

Algoritma *K-Means* memiliki kelebihan berupa proses yang relatif mudah dipahami, efisien pada data numerik, dan menghasilkan *centroid* yang dapat digunakan untuk menggambarkan profil setiap kelompok. Namun, metode ini mengharuskan jumlah *cluster* ditentukan sebelum pemodelan serta sensitif terhadap perbedaan skala, pemilihan *centroid* awal, nilai ekstrem, dan bentuk penyebaran data. Alasali dan Ortakcı (2024) menjelaskan bahwa pemilihan teknik *clustering* perlu

disesuaikan dengan struktur data karena setiap metode memiliki asumsi dan keterbatasan yang berbeda. (Anjarwati dkk., 2025) juga menunjukkan bahwa karakteristik distribusi dan keberadaan *outlier* perlu diperhatikan ketika memilih algoritma pengelompokan.

Dalam penelitian ini, algoritma *K-Means* diterapkan pada data hasil normalisasi dari 16 ULP dengan tiga variabel, yaitu daya terpasang, produksi listrik, dan jumlah pelanggan. Hasil pemodelan digunakan untuk memperoleh *centroid*, anggota, dan profil setiap *cluster* secara eksploratif. Pengelompokan tersebut tidak digunakan untuk melakukan prediksi maupun menyimpulkan kinerja teknis setiap ULP. Setelah mekanisme *K-Means* dijelaskan, pembahasan selanjutnya diarahkan pada metode evaluasi *clustering* yang digunakan untuk mempertimbangkan jumlah kelompok dan menilai kualitas hasil pemodelan.

2.2.10 Evaluasi Clustering

Evaluasi *clustering* merupakan tahap untuk menilai tingkat kekompakan kelompok dan keterpisahan antar-*cluster* yang terbentuk. Berbeda dengan klasifikasi, proses *clustering* tidak menggunakan label kelas sebenarnya sebagai pembanding. Oleh karena itu, evaluasi dilakukan menggunakan ukuran internal yang memanfaatkan hubungan jarak antarobjek dan pusat kelompok. Alasali dan Ortakçı (2024) menjelaskan bahwa pemilihan ukuran evaluasi perlu disesuaikan dengan karakteristik metode dan struktur data karena setiap ukuran menilai aspek kualitas *cluster* yang berbeda.

Dalam penelitian ini, evaluasi dilakukan menggunakan *Elbow Method* dan *Silhouette Coefficient*. *Elbow Method* menilai perubahan tingkat kekompakan kelompok melalui nilai *Within-Cluster Sum of Squares* (WCSS) pada beberapa alternatif jumlah *cluster*. Sementara itu, *Silhouette Coefficient* mengukur kedekatan setiap data terhadap anggota dalam kelompok yang sama serta keterpisahannya dari kelompok lain. Penggunaan kedua metode secara bersama-sama memberikan pertimbangan yang lebih lengkap dibandingkan penggunaan satu ukuran saja. Fourteen et al. (2025) menunjukkan bahwa nilai *inertia* dan *Silhouette* dapat menghasilkan informasi evaluasi yang berbeda karena kualitas pengelompokan dipengaruhi oleh jumlah *cluster* dan karakteristik dataset.

Pada penelitian ini, *Elbow Method* diterapkan pada alternatif $K = 1$ sampai $K = 6$, sedangkan *Silhouette Coefficient* dihitung pada $K = 2$ sampai $K = 6$. *Silhouette* tidak dihitung pada $K = 1$ karena evaluasi tersebut memerlukan sekurang-kurangnya dua kelompok untuk

membandingkan jarak data terhadap *cluster* sendiri dan *cluster* terdekat lainnya. Hasil dari kedua metode kemudian dipertimbangkan untuk menentukan jumlah *cluster* pada model akhir. Pembahasan lebih lanjut mengenai prinsip, perhitungan, dan interpretasi kedua metode tersebut dijelaskan pada Subbab 2.2.10.1 dan Subbab 2.2.10.2.

2.2.7 Normalisasi Data

Normalisasi data merupakan salah satu tahapan penting dalam proses *preprocessing* pada *data mining* yang bertujuan untuk menyamakan skala nilai antar variabel agar dapat dibandingkan secara adil dalam proses analisis. Data yang memiliki rentang nilai yang berbeda secara signifikan dapat menyebabkan bias pada algoritma berbasis jarak, seperti *K-Means*, karena variabel dengan skala yang lebih besar akan lebih dominan dalam perhitungan jarak. Joyeta et.al (2020) dalam penelitiannya menyatakan bahwa normalisasi data berperan dalam meningkatkan akurasi model serta membantu algoritma dalam mengenali pola data secara lebih optimal. Selain itu, penelitian oleh Annad & Wahab (2023) menyatakan bahwa proses normalisasi sangat penting dalam pengolahan data karena dapat meningkatkan performa algoritma *clustering* dan mengurangi pengaruh perbedaan skala antar atribut.

Secara umum, terdapat beberapa metode normalisasi yang sering digunakan dalam *data mining*, antara lain:

1. *Min-Max Normalization*

Metode ini mengubah nilai data ke dalam rentang tertentu, biasanya antara 0 hingga 1. Metode ini sering digunakan karena mampu mempertahankan distribusi data asli tanpa mengubah bentuk pola data.

$$x' = (x - x_{min}) / (x_{max} - x_{min}) \quad (2.6)$$

2. *Z-Score Normalization (Standardization)*

Metode ini mengubah data sehingga memiliki rata-rata (mean) 0 dan standar deviasi 1. Metode ini cocok digunakan jika data memiliki distribusi yang bervariasi.

$$x' = (x - \mu) / \quad (2.7)$$

3. *Decimal Scaling*

Metode ini dilakukan dengan cara membagi nilai data dengan faktor tertentu berdasarkan jumlah digit terbesar dalam data.

$$x' = x/10^j \quad (2.8)$$

Dimana :

- a) x = nilai data asli
- b) x' = nilai hasil normalisasi
- c) x_{min}, x_{max} = nilai minimum dan maksimum
- d) μ = rata-rata (mean)
- e) σ = standar deviasi
- f) j = jumlah digit

Normalisasi data memiliki peran yang sangat penting terutama dalam algoritma berbasis jarak seperti *K-Means*, karena perhitungan jarak sangat sensitif terhadap perbedaan skala data. Tanpa normalisasi, variabel dengan nilai besar seperti produksi listrik (kWh) dapat mendominasi variabel lain seperti daya terpasang (KVA), sehingga hasil clustering menjadi kurang optimal. Oleh karena itu, dalam penelitian ini digunakan metode *Min-Max Normalization* untuk menjaga distribusi data tetap proporsional dan memastikan bahwa setiap variabel memiliki kontribusi yang seimbang dalam proses *clustering*.

2.2.8 Daya Terpasang (kVA)

Daya terpasang merupakan besaran kapasitas daya listrik maksimum yang disediakan oleh penyedia listrik kepada pelanggan sesuai dengan perjanjian yang telah ditetapkan. Daya ini dinyatakan dalam satuan kilo *Volt Ampere* (kVA) dan mencerminkan kemampuan maksimum sistem kelistrikan dalam menyuplai energi listrik kepada pelanggan. Dalam sistem distribusi tenaga listrik, daya terpasang menjadi indikator penting dalam perencanaan kapasitas jaringan, karena berkaitan langsung dengan kemampuan gardu distribusi, transformator, serta infrastruktur pendukung lainnya. Rifqi & Aldisa (2023) dalam penelitiannya menyatakan kapasitas daya listrik merupakan faktor utama dalam memastikan keandalan sistem distribusi serta menjaga keseimbangan antara permintaan dan pasokan energi listrik dan data daya terpasang dapat digunakan sebagai salah satu variabel penting dalam analisis data untuk memahami karakteristik penggunaan energi listrik.

Secara teknis, daya terpasang berkaitan dengan konsep daya semu (*apparent power*), yang merupakan kombinasi antara daya aktif dan daya reaktif dalam sistem kelistrikan. Hubungan antara daya semu, Daya aktif, dan daya reaktif dapat dinyatakan dalam rumus sebagai berikut :

$$S = \sqrt{P^2 + Q^2} \quad (2.9)$$

Keterangan:

- a) S = daya semu (kVA)
- b) P = daya aktif (kW)
- c) Q = daya reaktif (kVAR)

Daya terpasang memiliki karakteristik yang relatif stabil dalam jangka waktu tertentu karena ditentukan berdasarkan kontrak antara pelanggan dan penyedia listrik. Namun, dalam praktiknya, tidak semua daya yang terpasang akan digunakan secara maksimal oleh pelanggan. Hal ini menyebabkan adanya perbedaan antara kapasitas yang tersedia dengan energi listrik yang benar-benar digunakan. Oleh karena itu, analisis terhadap daya terpasang sangat penting untuk mengetahui tingkat pemanfaatan kapasitas listrik serta efisiensi penggunaan energi.

Dalam konteks penelitian ini, daya terpasang (KVA) digunakan sebagai variabel yang merepresentasikan kapasitas maksimum energi listrik yang tersedia pada setiap Unit Pelayanan Pelanggan (ULP). Dengan mengombinasikan data daya terpasang dengan data produksi listrik (kWh), dapat dilakukan analisis lebih lanjut untuk mengetahui pola penggunaan energi, serta mengidentifikasi kondisi seperti *under-utilization* (kapasitas besar tetapi penggunaan rendah) dan *over-utilization* (penggunaan mendekati kapasitas maksimum). Hal ini sangat penting dalam mendukung pengambilan keputusan terkait perencanaan dan pengelolaan sistem distribusi listrik yang lebih efisien.

2.2.9 Produksi Listrik (kWh)

Produksi listrik merupakan jumlah energi listrik yang dihasilkan atau disalurkan kepada pelanggan dalam periode waktu tertentu yang dinyatakan dalam satuan kilo *Watt hour* (kWh). Energi listrik ini mencerminkan tingkat konsumsi aktual pelanggan terhadap daya listrik yang tersedia. Dalam sistem ketenagalistrikan, produksi listrik menjadi indikator penting dalam mengukur tingkat penggunaan energi serta efisiensi pemanfaatan daya listrik oleh pelanggan. Ramadani et.al (2023) dalam penelitiannya menyatakan bahwa konsumsi energi listrik merupakan

salah satu parameter utama dalam analisis sistem energi karena berkaitan langsung dengan kebutuhan energi masyarakat serta perencanaan pasokan listrik. Data konsumsi energi listrik dapat digunakan untuk mengidentifikasi pola penggunaan energi serta membantu dalam pengambilan keputusan yang berbasis data.

Secara teknis, energi listrik (kWh) merupakan hasil dari penggunaan daya listrik dalam periode waktu tertentu. Hubungan antara energi, daya, dan waktu dapat dinyatakan dalam rumus sebagai berikut:

$$E = P \times t \quad (2.10)$$

Keterangan:

- a) E = energi listrik (kWh)
- b) P = daya listrik (kW)
- c) t = waktu (jam)

Produksi listrik bersifat dinamis karena dipengaruhi oleh berbagai faktor, seperti aktivitas pelanggan, kondisi ekonomi, cuaca, serta jenis penggunaan listrik (rumah tangga, industri, atau bisnis). Fluktuasi penggunaan energi listrik ini menyebabkan perbedaan tingkat konsumsi pada setiap wilayah atau Unit Pelayanan Pelanggan (ULP). Oleh karena itu, analisis terhadap data produksi listrik sangat penting untuk memahami pola konsumsi energi serta mengidentifikasi kecenderungan penggunaan listrik di suatu wilayah.

Dalam konteks penelitian ini, produksi listrik (kWh) digunakan sebagai variabel yang merepresentasikan tingkat pemanfaatan energi listrik secara nyata oleh pelanggan pada setiap ULP. Dengan mengombinasikan variabel ini dengan daya terpasang (KVA), dapat dilakukan analisis yang lebih komprehensif untuk mengetahui hubungan antara kapasitas listrik yang tersedia dengan tingkat penggunaannya. Hasil analisis ini diharapkan dapat memberikan gambaran mengenai efisiensi penggunaan energi serta membantu dalam proses segmentasi ULP menggunakan metode *K-Means*, sehingga dapat mendukung pengambilan keputusan dalam pengelolaan distribusi energi listrik secara lebih optimal.

2.2.10 Evaluasi Clustering

Evaluasi *clustering* merupakan tahap penting dalam proses analisis data berbasis *unsupervised learning* yang bertujuan untuk menilai kualitas hasil pengelompokan data. Berbeda

dengan metode *supervised learning*, *clustering* tidak memiliki label acuan (*ground truth*), sehingga diperlukan metode evaluasi khusus untuk mengukur seberapa baik data telah dikelompokkan. Menurut Annas & Wahab (2023), evaluasi *clustering* digunakan untuk memastikan bahwa hasil pengelompokkan memiliki tingkat kesamaan yang tinggi dalam satu *cluster* dan perbedaan yang jelas antar *cluster*. Selain itu, Alasali et.al (2024) menjelaskan bahwa evaluasi *clustering* sangat penting untuk menentukan jumlah *cluster* yang optimal serta menghindari kesalahan dalam interpretasi hasil analisis.

Dalam penelitian ini, evaluasi *clustering* dilakukan menggunakan beberapa metode yang umum digunakan, yaitu sebagai berikut:

1. *Elbow Method*

Elbow Method merupakan metode yang digunakan untuk menentukan jumlah *cluster* (k) yang optimal dengan cara melihat nilai *Within Cluster Sum of Squares* (WCSS). WCSS merupakan jumlah total jarak kuadrat antara data dengan *centroid* dalam satu *cluster*. Semakin kecil nilai WCSS, maka *cluster* yang terbentuk semakin baik.

$$WCSS = \sum_{j=1}^k \sum_{i=1}^n |x_i^{(j)} - c_j|^2 \quad (2.11)$$

Metode ini bekerja dengan cara menghitung nilai WCSS untuk beberapa jumlah *cluster*, kemudian divisualisasikan dalam bentuk grafik. Titik optimal ditentukan pada bagian grafik yang membentuk sudut siku (*elbow*), yaitu saat penurunan nilai WCSS mulai melambat. Titik tersebut menunjukkan bahwa penambahan jumlah *cluster* tidak memberikan peningkatan yang signifikan terhadap kualitas *clustering*.

2. *Silhouette Coefficient*

Silhouette Coefficient merupakan metode evaluasi *clustering* yang mengukur seberapa baik suatu data ditempatkan dalam cluster tertentu berdasarkan tingkat kedekatan dengan data dalam *cluster* yang sama dan jaraknya terhadap *cluster* lain. Nilai *silhouette* berada pada rentang -1 hingga 1.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2.12)$$

Keterangan:

- a) $a(i)$ = rata-rata jarak data ke data lain dalam satu cluster
- b) $b(i)$ = rata-rata jarak data ke cluster terdekat lainnya

Interpretasi nilai silhouette:

- a) Mendekati 1 $\rightarrow$ data terkelompok dengan baik
- b) Mendekati 0 $\rightarrow$ data berada di antara dua cluster
- c) Bernilai negatif $\rightarrow$ kemungkinan salah pengelompokan

Metode ini memberikan gambaran yang lebih akurat mengenai kualitas *cluster* karena mempertimbangkan kohesi (kedekatan dalam cluster) dan separasi (jarak antar *cluster*) secara bersamaan. data, sehingga tahap *preprocessing* menjadi sangat vital.

2.3 Kerangka Pemikiran

Kerangka pemikiran dalam penelitian ini disusun untuk menunjukkan hubungan antara data yang digunakan, tahapan pengolahan, metode pengelompokan, serta keluaran yang dihasilkan. Penyusunannya mengacu pada proses analisis yang telah ditetapkan dalam kerangka CRISP-DM, sehingga setiap bagian menggambarkan alur penelitian secara berurutan mulai dari pengumpulan dan pemahaman data hingga evaluasi hasil *clustering*. Dengan demikian, kerangka pemikiran tidak hanya menunjukkan urutan kegiatan, tetapi juga memperjelas keterkaitan antara permasalahan penelitian, tiga variabel yang dianalisis, dan metode yang digunakan.

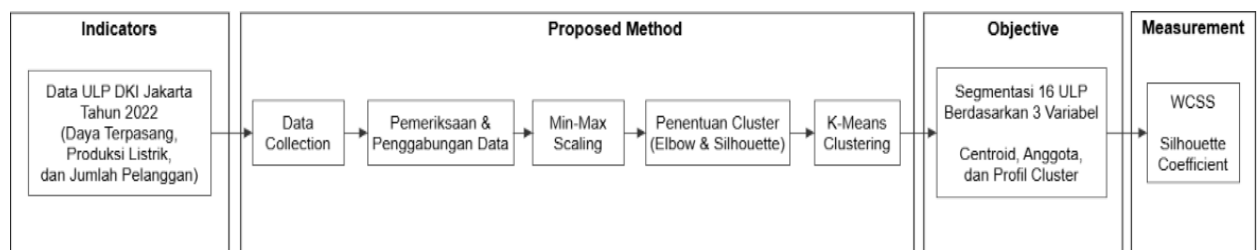
Pada tahap masukan, penelitian menggunakan dua dataset sekunder kelistrikan di DKI Jakarta tahun 2022. Dataset pertama memuat daya terpasang dan produksi listrik pada setiap Unit Pelayanan Pelanggan (ULP), sedangkan dataset kedua memuat jumlah pelanggan. Kedua dataset memiliki nama ULP sebagai atribut penghubung. Setelah baris agregat DKI Jakarta dikeluarkan, data yang dianalisis terdiri atas 16 ULP dengan tiga variabel, yaitu daya terpasang dalam satuan kVA, produksi listrik dalam satuan kWh, dan jumlah pelanggan.

Data tersebut selanjutnya memasuki tahap proses yang mencakup pemeriksaan struktur dan kualitas data, pemilihan atribut, penyeragaman nama ULP, konversi tipe data, penghapusan baris agregat, serta penggabungan kedua dataset. Setelah dataset akhir terbentuk, ketiga variabel dinormalisasi menggunakan *Min-Max Scaling* agar memiliki rentang nilai yang sebanding. Jumlah *cluster* kemudian dipertimbangkan menggunakan *Elbow Method* berdasarkan perubahan nilai WCSS dan *Silhouette Coefficient* berdasarkan tingkat kekompakan serta keterpisahan kelompok. Hasil

pertimbangan tersebut digunakan dalam pemodelan menggunakan algoritma *K-Means Clustering* dengan ukuran jarak Euclidean.

Tahap keluaran menghasilkan segmentasi 16 ULP berdasarkan kombinasi daya terpasang, produksi listrik, dan jumlah pelanggan. Hasil pengelompokan disajikan dalam bentuk keanggotaan ULP, nilai *centroid*, profil setiap *cluster*, serta visualisasi sebaran data. Kualitas hasil pemodelan diukur menggunakan WCSS dan *Silhouette Coefficient*, termasuk pemeriksaan nilai Silhouette pada setiap kelompok dan setiap anggota. Hasil tersebut digunakan untuk menggambarkan perbedaan profil antar-ULP secara eksploratif dan tidak digunakan secara langsung untuk menilai kinerja teknis, tingkat efisiensi, maupun kondisi operasional masing-masing unit.

Berdasarkan alur tersebut, kerangka pemikiran menghubungkan dua sumber data, tiga variabel penelitian, tahap persiapan data, proses penentuan jumlah *cluster*, pemodelan *K-Means*, dan evaluasi hasil dalam satu rangkaian yang sistematis. Kerangka ini menjadi dasar konseptual bagi tahapan penelitian yang dijelaskan secara lebih operasional pada Bab III.



Gambar 2.2 Kerangka Pemikiran Penelitian

Berdasarkan Gambar 2.2, kerangka pemikiran penelitian terdiri atas empat bagian utama, yaitu *indicators*, *proposed method*, *objective*, dan *measurement*. Bagian *indicators* memuat data Unit Pelayanan Pelanggan (ULP) di DKI Jakarta tahun 2022 yang mencakup daya terpasang, produksi listrik, dan jumlah pelanggan. Ketiga variabel tersebut diperoleh dari penggabungan dataset kelistrikan dan dataset jumlah pelanggan sebagai dasar pembentukan profil setiap ULP.

Selanjutnya, bagian *proposed method* menunjukkan tahapan analisis yang dimulai dari *data collection*, pemeriksaan dan penggabungan data, serta normalisasi menggunakan *Min-Max Scaling*. Setelah data memiliki rentang yang sebanding, jumlah *cluster* dipertimbangkan menggunakan *Elbow Method* dan *Silhouette Coefficient*, kemudian pengelompokan dilakukan dengan algoritma *K-Means Clustering*.

Pada bagian *objective*, hasil analisis berupa segmentasi 16 ULP berdasarkan ketiga variabel penelitian, yang selanjutnya dijelaskan melalui nilai *centroid*, anggota, dan profil setiap *cluster*. Adapun bagian *measurement* menggunakan WCSS untuk menilai kekompakan data terhadap *centroid* serta *Silhouette Coefficient* untuk menilai kedekatan anggota dalam kelompok dan keterpisahan antar-*cluster*. Dengan demikian, kerangka pemikiran ini menggambarkan alur penelitian secara menyeluruh, mulai dari data awal, proses pengolahan, pembentukan kelompok, hingga evaluasi hasil segmentasi.