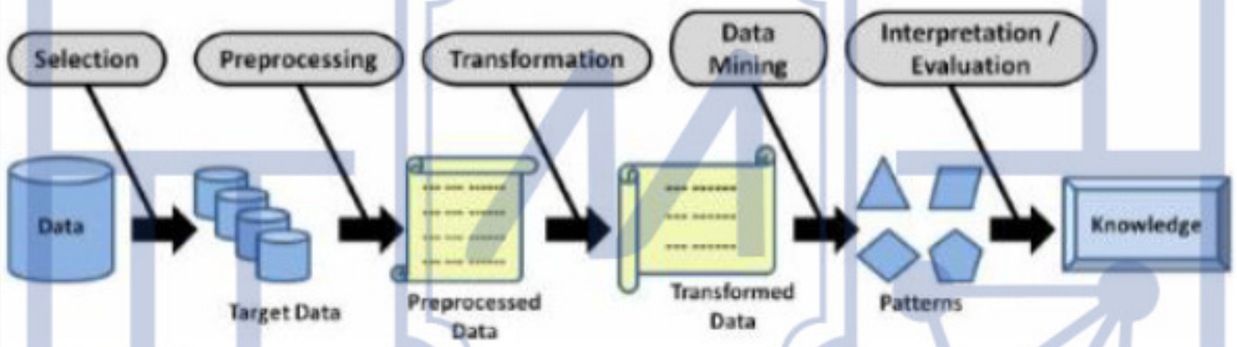
BAB II

KAJIAN LITERATUR

2.1 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) merupakan proses terstruktur dan bersifat iteratif yang bertujuan menemukan pola valid, baru, dan bermakna dari kumpulan data berukuran besar. Dalam praktiknya, KDD memadukan metode dari berbagai bidang seperti statistik, kecerdasan buatan, pembelajaran mesin, dan ilmu komputer, sehingga pengetahuan yang dihasilkan memiliki tingkat keandalan yang tinggi [18]. Nilai utama KDD terletak pada kemampuannya mengubah data mentah menjadi informasi yang bermakna dan dapat dimanfaatkan dalam mendukung pengambilan keputusan serta menghasilkan pengetahuan baru dari berbagai domain data [19]. Tahapan proses KDD secara umum ditunjukkan pada Gambar 2.1



Gambar 2. 1 Proses Knowledge Discovery in Databases (KDD) [20]

Secara garis besar, proses *Knowledge Discovery in Databases (KDD)* terdiri dari lima tahapan yang saling berurutan, yaitu:

1. Seleksi Data, yaitu memilih data yang relevan dari keseluruhan sumber data sehingga diperoleh target data yang sesuai dengan tujuan penelitian.
2. Praproses Data, yakni membersihkan data dari nilai kosong, duplikat, dan inkonsistensi,
3. Transformasi Data, yakni mengonversi dan menormalisasi data ke format yang siap diproses,
4. Data Mining, yakni menerapkan algoritma untuk mengungkap pola tersembunyi,
5. Evaluasi dan Interpretasi, yaitu mengevaluasi pola yang diperoleh untuk memastikan validitas dan relevansinya terhadap tujuan penelitian, kemudian menginterpretasikan pola tersebut menjadi pengetahuan *(knowledge)* yang dapat digunakan sebagai dasar pengambilan keputusan.

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

Tahapan tersebut menggambarkan alur kerja metodologi *Knowledge Discovery in Databases* (KDD) yang diterapkan secara sistematis melalui tahapan seleksi data, pembersihan data, transformasi data, *data mining*, dan evaluasi hasil. Setiap tahapan saling berkaitan untuk menghasilkan informasi dan pengetahuan yang relevan dari data yang diolah. Meskipun demikian, penilaian kualitas data belum memiliki standar yang digunakan secara universal, sehingga setiap penelitian dapat menerapkan metode dan kriteria evaluasi yang berbeda sesuai dengan karakteristik data serta tujuan analisis. Kondisi tersebut menyebabkan hasil pengolahan maupun analisis data dapat bervariasi antarpenelitian [21].

2.2 Data Mining

Data mining adalah bidang ilmu yang berfokus pada penggalian pola, relasi, atau pengetahuan tersembunyi dari kumpulan data berukuran besar dengan memanfaatkan teknik komputasional dan statistik. Bidang ini berada di irisan antara statistik, kecerdasan buatan, dan basis data, sehingga cakupannya cukup luas untuk menangani berbagai permasalahan pada data. Teknik yang umum digunakan dalam *data mining* antara lain klasifikasi, asosiasi, dan clustering. Di antara ketiganya, clustering menjadi teknik yang paling sesuai untuk data tanpa label kategori karena pengelompokan dilakukan berdasarkan kemiripan data itu sendiri. Selain clustering, *data mining* juga mencakup deteksi anomali, yaitu proses mengidentifikasi titik data yang menyimpang secara signifikan dari pola umum [22].

Di bidang transportasi publik khususnya, data mining terbukti bermanfaat untuk mengelompokkan halte atau rute berdasarkan karakteristik permintaan penumpang. Sejalan dengan itu, sejumlah algoritma clustering berbasis kepadatan bahkan mampu menjalankan fungsi clustering dan deteksi anomali sekaligus, karena titik yang tidak masuk ke dalam klaster mana pun secara otomatis dikategorikan sebagai noise yang dapat diinterpretasikan sebagai anomali [23].

Pertumbuhan data digital mendorong perkembangan data mining yang semakin pesat. Namun, penerapannya pada transportasi publik, khususnya di negara berkembang, masih menghadapi tantangan berupa keterbatasan ketersediaan data, kualitas data, serta integrasi data antarsistem [24]. Selain itu, penelitian yang menggabungkan penemuan pola (pattern discovery) dan deteksi anomali terus berkembang karena kombinasi kedua pendekatan tersebut mampu memberikan informasi yang lebih komprehensif mengenai pola normal maupun penyimpangan pada data [25].

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

2.3 Clustering

Clustering adalah salah satu teknik dalam *data mining* yang bekerja dengan cara mengelompokkan data berdasarkan seberapa mirip karakteristik yang dimiliki setiap objek. Data yang serupa dikumpulkan dalam satu kelompok, sementara data yang berbeda dipisahkan. Karena proses ini berjalan tanpa label atau kategori yang ditentukan sebelumnya, *clustering* masuk ke dalam kelompok metode *unsupervised learning*. Metode ini banyak digunakan ketika peneliti ingin menemukan pola atau struktur tersembunyi dalam data yang kompleks, tanpa harus mengetahui terlebih dahulu kategori apa yang akan terbentuk [26].

Secara umum, metode clustering terbagi ke dalam beberapa pendekatan utama. Partitioning clustering membagi data ke dalam sejumlah cluster yang telah ditentukan, dengan K-Means sebagai algoritma yang paling umum digunakan. Hierarchical clustering membangun struktur pengelompokan secara bertingkat yang hasilnya biasanya ditampilkan dalam bentuk dendrogram. Density-based clustering membentuk cluster berdasarkan tingkat kepadatan data di suatu area tertentu. Adapun grid-based clustering terlebih dahulu membagi ruang data ke dalam sejumlah grid sebelum proses pengelompokan dilakukan [27].

Di antara pendekatan tersebut, *density-based clustering* bekerja dengan prinsip bahwa data yang berkumpul di area padat akan membentuk satu *cluster*, sedangkan data yang berada di area sepi dianggap sebagai *noise* atau *outlier*. Pendekatan ini unggul dalam menangani klaster yang bentuknya tidak beraturan (*arbitrary shape*), sesuatu yang sulit dilakukan oleh metode lain. Beberapa algoritma yang termasuk dalam kategori ini antara lain *DBSCAN*, *OPTICS*, dan *HDBSCAN* [28].

Clustering memiliki sejumlah kelebihan yang membuatnya banyak digunakan. Metode ini tidak memerlukan label data sehingga fleksibel untuk berbagai jenis dataset, mampu mengungkap pola tersembunyi yang tidak terlihat secara kasatmata, serta membantu menyederhanakan data kompleks menjadi kelompok yang lebih mudah dipahami [29]. Namun demikian, *clustering* juga memiliki keterbatasan. Kualitas hasil pengelompokan sangat bergantung pada karakteristik data dan parameter yang dipilih, beberapa metode cukup sensitif terhadap *noise* dan *outlier*, dan interpretasi hasilnya kerap bersifat subjektif karena nilai optimal bergantung pada karakteristik *dataset* [30]. Oleh karena itu, pemilihan metode *clustering* yang tepat menjadi langkah penting agar hasil yang diperoleh benar-benar bermakna dan dapat dipertanggungjawabkan.

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

2.4 Praproses Data

Praproses data merupakan tahap awal yang dilakukan untuk membersihkan dan menyiapkan data mentah sebelum proses analisis dilakukan. Tahap ini penting karena data yang mengandung nilai hilang, duplikasi, inkonsistensi format, atau kesalahan pencatatan dapat memengaruhi kualitas hasil analisis dan menghasilkan model yang kurang akurat [31].

Secara umum, praproses data mencakup beberapa langkah utama, yaitu seleksi atribut, pembersihan data, normalisasi, dan transformasi data. Data yang telah melalui proses praproses dengan baik cenderung menghasilkan model clustering yang lebih optimal dibandingkan dengan data yang langsung digunakan tanpa proses pembersihan dan penyesuaian [32].

Tahapan praproses data yang umum diterapkan meliputi:

1. Seleksi atribut, yaitu pemilihan variabel yang relevan sesuai kebutuhan kajian,
2. Pembersihan data, yaitu penghapusan data yang memiliki nilai hilang, duplikat, atau tidak valid,
3. Normalisasi fitur menggunakan metode *Min-Max Normalization* yang mentransformasi nilai ke rentang [0, 1] dengan rumus sebagai berikut:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \text{..... (1)}$$

Keterangan variabel:

x' = nilai hasil normalisasi

x = nilai asli data

x_{min} = nilai minimum fitur

x_{max} = nilai maksimum fitur

Proses ini bertujuan agar tidak ada satu variabel yang mendominasi perhitungan jarak antar titik data,

4. *Encoding* variabel kategorikal, yaitu mengubah data kategorikal ke dalam bentuk numerik agar dapat diproses oleh algoritma *clustering*.

2.5 HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) merupakan algoritma clustering berbasis kepadatan yang menjadi dasar pengembangan HDBSCAN. DBSCAN menggunakan dua parameter utama, yaitu epsilon (ε) sebagai radius pencarian tetangga dan MinPts sebagai jumlah minimum titik dalam radius tersebut, sehingga setiap data dapat dikategorikan sebagai core point, border point, atau noise point. Penggunaan

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

parameter epsilon yang bersifat global menyebabkan DBSCAN memiliki keterbatasan dalam mengidentifikasi klaster pada data yang memiliki tingkat kepadatan yang bervariasi [33].

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) dikembangkan untuk mengatasi keterbatasan tersebut melalui pendekatan hierarkis tanpa memerlukan parameter epsilon global. Proses HDBSCAN meliputi perhitungan mutual reachability distance, pembentukan minimum spanning tree, penyusunan hierarki klaster dalam bentuk dendrogram, serta ekstraksi klaster yang stabil menggunakan metode Framework for Optimal Selection of Clusters (FOSC). Parameter utama yang digunakan adalah min_cluster_size dan min_samples untuk mengendalikan pembentukan klaster berdasarkan struktur kepadatan data [34].

Sebelum menghitung *mutual reachability distance*, *HDBSCAN* terlebih dahulu menghitung jarak dasar antara dua titik data menggunakan *Euclidean distance*, yaitu jarak lurus yang dihitung berdasarkan seluruh atribut yang dimiliki oleh masing-masing titik. Secara matematis, jarak *Euclidean* antara dua titik a dan b dirumuskan sebagai:

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad \text{.....(2)}$$

Keterangan variabel:

d(a,b) = jarak *Euclidean* antara titik a dan b

a_i = nilai atribut ke-i pada titik a

b_i = nilai atribut ke-i pada titik b

n = jumlah atribut

Selain jarak *Euclidean*, HDBSCAN juga memperhitungkan tingkat kepadatan lokal di sekitar setiap titik data melalui *core distance*. *Core distance* suatu titik didefinisikan sebagai jarak titik tersebut ke tetangga ke-k terdekat, dengan k merupakan nilai parameter min_samples. Secara matematis, core distance dirumuskan sebagai:

$$core\_dist_k(p) = d(p, N_k(p)) \quad \text{.....(3)}$$

Keterangan variabel:

$core\_dist_k(p)$ = core distance titik p

$N_k(p)$ = tetangga ke-k terdekat dari titik p

k = nilai min_samples

Mutual reachability distance merupakan konsep penting dalam HDBSCAN yang mendefinisikan jarak antar titik dengan mempertimbangkan kepadatan lokal. Secara matematis, jarak antara dua titik a dan b dinyatakan sebagai:

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

$$mrd(a,b) = max\big(core\_dist(a), core\_dist(b), dist(a,b)\big) \ldots\ldots\ldots (4)$$

Keterangan variabel:

mrd(a,b) = *mutual reachability distance* antara titik a dan titik b

core_dist(a) = *core distance* titik a (jarak ke tetangga ke-min_samples terdekat dari a)

core_dist(b) = *core distance* titik b (jarak ke tetangga ke-min_samples terdekat dari b)

dist(a,b) = jarak *Euclidean* antara titik a dan titik b

Nilai mrd(a, b) selalu lebih besar atau sama dengan jarak Euclidean asli. Hal ini memungkinkan HDBSCAN mengenali klaster dengan kepadatan yang berbeda dalam satu proses tanpa bergantung pada parameter global [35].

HDBSCAN telah banyak digunakan dalam berbagai permasalahan karena kemampuannya menangani variasi kepadatan data serta mengidentifikasi noise secara otomatis [36]. Penelitian mengenai penerapan HDBSCAN pada berbagai domain juga menunjukkan bahwa algoritma ini mampu mengelompokkan data dengan distribusi yang kompleks tanpa memerlukan jumlah klaster di awal [37].

Salah satu karakteristik HDBSCAN adalah pembentukan condensed cluster tree yang merepresentasikan hierarki klaster berdasarkan tingkat kepadatan, sehingga stabilitas setiap klaster dapat dievaluasi [38]. Berdasarkan struktur tersebut, metode Framework for Optimal Selection of Clusters (FOSC) digunakan untuk memilih klaster yang memiliki tingkat stabilitas tertinggi tanpa menetapkan jumlah klaster sejak awal [39].

Klaster akhir pada HDBSCAN dipilih melalui algoritma FOSC berdasarkan nilai stabilitasnya pada *condensed cluster tree*. Stabilitas suatu klaster dihitung dari selisih nilai *lambda* pada saat setiap titik keluar dan pada saat titik tersebut mulai menjadi anggota klaster. Secara matematis, stabilitas klaster dirumuskan sebagai:

$$Stability(C) = \sum_{p \in C}(\lambda_{keluar}(p) - \lambda_{masuk}(p)) \ldots\ldots\ldots\ldots\ldots (5)$$

Keterangan variabel:

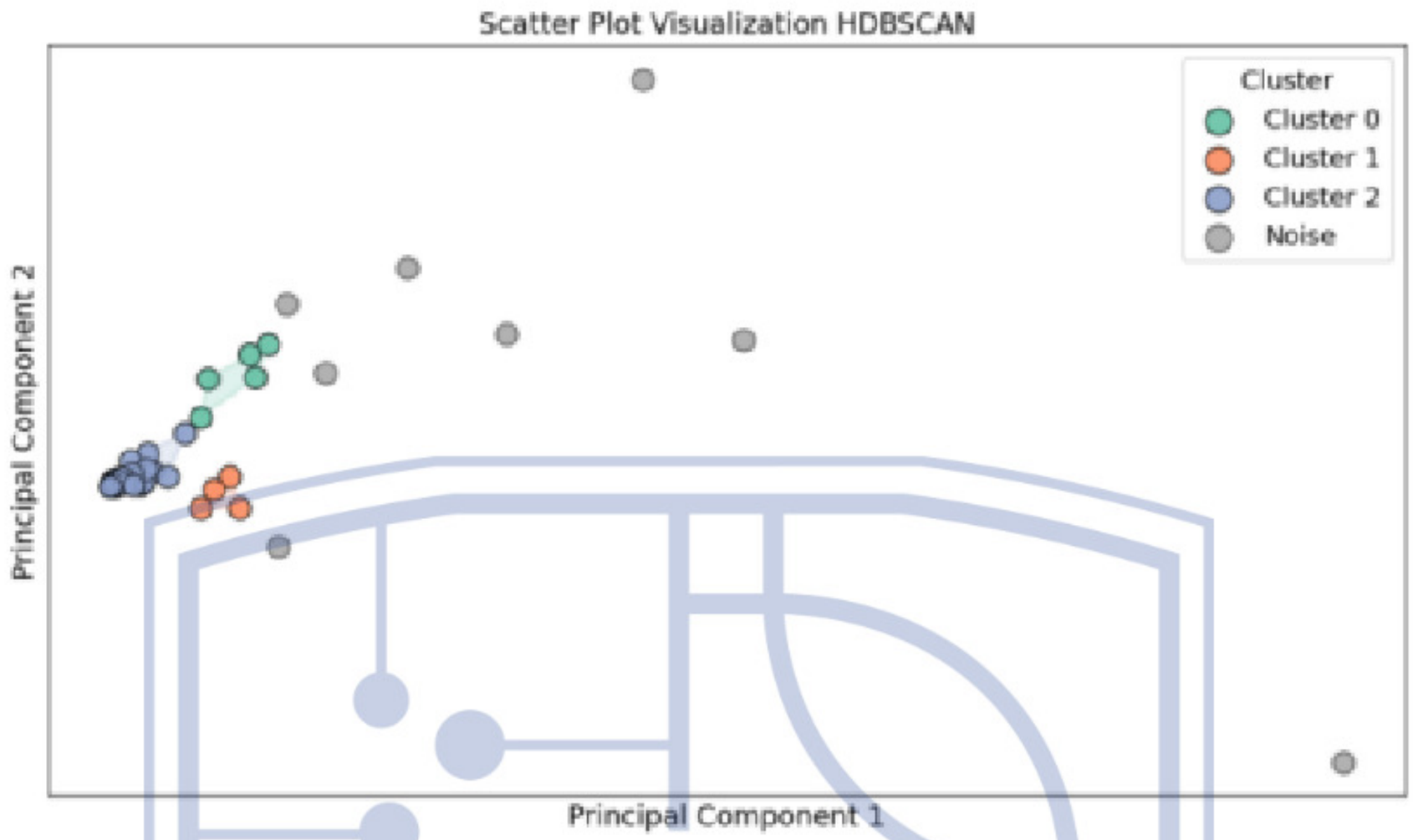
Stability(C) = stabilitas klaster

λ keluar = nilai lambda saat titik keluar dari klaster

λ masuk = nilai lambda saat titik mulai menjadi anggota klaster

p = titik data dalam klaster

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.



Gambar 2. 2 Ilustrasi Hasil Clustering Menggunakan Algoritma HDBSCAN [40]

Gambar 2.2 menunjukkan contoh hasil klasterisasi HDBSCAN yang divisualisasikan dalam bentuk scatter plot. Terlihat adanya kelompok data yang lebih padat dan saling berdekatan (Cluster 0, 1, dan 2), sementara sebagian titik data lainnya tersebar lebih renggang dan jauh dari kelompok-kelompok tersebut. Kondisi ini menggambarkan data dengan kepadatan yang tidak seragam. Variasi kepadatan tersebut menjadi tantangan bagi algoritma berbasis kepadatan seperti DBSCAN yang menggunakan parameter global, sehingga sulit membedakan klaster padat dari titik-titik yang tersebar renggang dalam satu proses. HDBSCAN mampu mengatasi hal ini dengan menyesuaikan kepadatan secara adaptif sehingga klaster dengan kepadatan yang berbeda dapat diidentifikasi dalam satu proses. Selain itu, data yang tidak termasuk dalam klaster mana pun, karena tidak memenuhi kriteria kepadatan minimum, akan dikategorikan sebagai *noise*.

2.6 Evaluasi Clustering

Evaluasi kualitas hasil *clustering* merupakan tahap penting untuk menilai sejauh mana algoritma mampu membentuk kelompok data yang baik. Pada *unsupervised learning*, evaluasi dilakukan menggunakan metrik internal yang menilai struktur klaster tanpa menggunakan label kelas. Salah satu metrik yang paling umum digunakan adalah *Silhouette Coefficient* [41].

Silhouette Coefficient mengukur tingkat kedekatan setiap titik data terhadap klasternya sendiri dibandingkan dengan klaster terdekat lainnya. Untuk setiap titik i, nilai

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

a(i) merupakan rata-rata jarak ke semua titik dalam klaster yang sama, sedangkan b(i) adalah rata-rata jarak minimum ke klaster terdekat yang berbeda [42]. Nilai *Silhouette* dihitung dengan rumus sebagai berikut:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \text{..... (6)}$$

Keterangan variabel:

s(i) = nilai *Silhouette* untuk titik data ke-i, rentang [-1, +1]

a(i) = rata-rata jarak titik i ke semua titik lain dalam klaster yang sama

b(i) = rata-rata jarak minimum titik i ke klaster terdekat yang berbeda

Nilai yang dihasilkan berada pada rentang -1 hingga +1. Nilai yang mendekati +1 menunjukkan bahwa titik data telah berada pada klaster yang tepat, nilai mendekati 0 menunjukkan posisi di batas antarklaster, sedangkan nilai negatif menunjukkan kemungkinan kesalahan pengelompokan [43].

Silhouette Coefficient banyak digunakan karena mudah diinterpretasikan dan mampu memberikan gambaran kualitas klaster secara keseluruhan. Nilai rata-rata *Silhouette* yang lebih tinggi menunjukkan bahwa klaster yang terbentuk memiliki tingkat kohesi dan separasi yang lebih baik [44].

Selain menggunakan *Silhouette Coefficient* untuk mengevaluasi kualitas hasil *clustering*, hasil pengelompokan juga dapat diklasifikasikan menggunakan metode *Natural Breaks Jenks*. Metode ini merupakan metode klasifikasi data yang digunakan untuk menentukan batas setiap kategori berdasarkan karakteristik data, sehingga hasil klasifikasi lebih mudah diinterpretasikan [45]. Hasil klasifikasi tersebut digunakan untuk mengelompokkan hasil *clustering* HDBSCAN ke dalam tingkat kepadatan, yaitu rendah, sedang, tinggi, dan sangat tinggi. Pengelompokan ini bertujuan untuk mempermudah interpretasi hasil *clustering* sehingga perbedaan tingkat kepadatan antarhalte lebih mudah dipahami.

2.7 Sistem Transportasi

Transportasi publik secara sederhana dapat dipahami sebagai layanan angkutan yang dioperasikan untuk melayani masyarakat secara bersama-sama, dengan rute, jadwal, dan tarif yang telah ditetapkan. Berbeda dengan kendaraan pribadi yang hanya melayani satu atau beberapa orang, transportasi publik dirancang untuk mengangkut banyak penumpang sekaligus sehingga mobilitas di dalam kota dapat berlangsung lebih efisien [46].

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

Transportasi publik merupakan pilar penting dalam mewujudkan sistem transportasi yang berkelanjutan, karena mampu mengurangi dampak lingkungan yang dihasilkan oleh sektor transportasi. Pengembangan transportasi publik juga menjadi bagian dari upaya menciptakan sistem mobilitas yang lebih ramah lingkungan dan berkelanjutan [47].

Beberapa ciri utama transportasi publik yang membedakannya dari moda transportasi lain adalah kapasitas angkut yang besar, jadwal operasional yang terstruktur, serta jangkauan layanan yang luas. Keberadaan sistem ini dapat meningkatkan efisiensi penggunaan ruang jalan dan mendukung pengembangan sistem transportasi perkotaan yang berkelanjutan [48].

Dalam praktiknya, transportasi publik hadir dalam berbagai bentuk, mulai dari bus, kereta api, hingga berbagai jenis angkutan massal lainnya. Salah satu sistem yang banyak diterapkan di kota-kota berkembang adalah Bus Rapid Transit (BRT), yaitu sistem bus yang beroperasi di jalur khusus, dilengkapi halte yang terintegrasi, dan menggunakan sistem pembayaran yang terorganisir. Penerapan BRT dapat meningkatkan kualitas layanan transportasi sekaligus mendukung efisiensi mobilitas perkotaan [49].

Di luar fungsinya sebagai sarana perpindahan, transportasi publik juga memiliki peran yang lebih luas. Layanan yang mudah diakses memungkinkan masyarakat menjangkau tempat kerja, sekolah, fasilitas kesehatan, dan pusat kegiatan lainnya tanpa harus memiliki kendaraan pribadi. Aksesibilitas yang baik melalui transportasi publik berkontribusi terhadap peningkatan mobilitas masyarakat dan pemerataan akses terhadap berbagai layanan perkotaan [50].

2.8 Konsep Kepadatan Penumpang

Kepadatan penumpang menggambarkan seberapa banyak pengguna layanan transportasi yang berada pada suatu lokasi atau dalam rentang waktu tertentu. Dalam konteks transportasi publik, kepadatan penumpang mencerminkan tingkat pemanfaatan layanan serta menunjukkan lokasi dan waktu ketika permintaan terhadap layanan transportasi mencapai tingkat yang tinggi. Informasi mengenai kepadatan penumpang sangat penting dalam pengelolaan transportasi publik karena berkaitan dengan kapasitas layanan, kenyamanan pengguna, dan efisiensi operasional [51].

Secara umum, kepadatan penumpang diukur berdasarkan jumlah pengguna yang berada pada suatu titik layanan dalam periode waktu tertentu [52]. Ketika kepadatan meningkat, berarti permintaan terhadap layanan transportasi juga meningkat. Sebaliknya, kepadatan yang rendah menunjukkan bahwa kapasitas layanan yang tersedia belum

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

dimanfaatkan secara optimal. Kedua kondisi tersebut perlu dipahami karena memberikan implikasi yang berbeda terhadap pengelolaan layanan transportasi.

Dari sudut pandang pengguna, kepadatan penumpang yang terlalu tinggi dapat mengurangi tingkat kenyamanan selama menggunakan layanan transportasi. Kondisi tersebut dapat menyebabkan ruang gerak menjadi terbatas, waktu tunggu meningkat, dan kualitas pelayanan yang dirasakan pengguna menurun [53]. Di sisi lain, tingkat kepadatan yang terlalu rendah juga dapat mengindikasikan bahwa layanan belum dimanfaatkan secara optimal sehingga diperlukan evaluasi terhadap pengelolaan layanan yang tersedia.

Lebih jauh, data kepadatan penumpang digunakan sebagai dasar pengambilan berbagai keputusan operasional, seperti penyesuaian kapasitas layanan, pengaturan jadwal perjalanan, evaluasi kinerja layanan, dan perencanaan pengembangan sistem transportasi. Dengan memahami pola kepadatan penumpang berdasarkan lokasi dan waktu, pengelola dapat mengelola permintaan layanan secara lebih efisien [54], [55].

2.9 Faktor-Faktor yang Mempengaruhi Kepadatan Penumpang

Kepadatan penumpang tidak muncul secara acak. Ada berbagai faktor yang bekerja secara bersamaan dan membentuk pola distribusi penumpang yang dapat sangat berbeda dari satu tempat ke tempat lain, maupun dari satu waktu ke waktu lainnya. Faktor-faktor tersebut meliputi:

1. **Faktor Spasial (Lokasi)**

   Lokasi merupakan salah satu faktor yang memengaruhi kepadatan penumpang. Titik layanan transportasi yang berada di sekitar pusat bisnis, kawasan perdagangan, kampus, atau permukiman padat penduduk cenderung memiliki jumlah penumpang yang lebih tinggi. Kondisi tata guna lahan di sekitar halte atau stasiun memiliki hubungan yang erat dengan tingkat penggunaan layanan transportasi publik [56].

2. **Faktor Temporal (Waktu)**

   Perbedaan waktu berpengaruh terhadap jumlah penumpang. Periode pagi dan sore hari, terutama pada jam berangkat dan pulang kerja atau sekolah, umumnya merupakan waktu dengan tingkat kepadatan penumpang yang tinggi. Selain itu, hari kerja dan hari libur menghasilkan pola mobilitas yang berbeda, di mana hari kerja cenderung menunjukkan pola perjalanan yang lebih teratur, sedangkan hari libur lebih dipengaruhi oleh aktivitas rekreasi dan sosial [57].

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

3. **Faktor Lingkungan (Cuaca)**

   Kondisi cuaca merupakan salah satu faktor yang dapat memengaruhi penggunaan transportasi publik. Curah hujan, suhu, dan kondisi cuaca lainnya dapat memengaruhi pilihan moda transportasi serta jumlah pengguna pada waktu tertentu. Oleh karena itu, faktor lingkungan perlu dipertimbangkan dalam perencanaan dan pengelolaan sistem transportasi publik [58].

4. **Faktor Kualitas Layanan**

   Kualitas layanan juga memengaruhi tingkat penggunaan transportasi publik. Frekuensi perjalanan, ketepatan waktu, keterjangkauan tarif, keamanan, kenyamanan, serta kemudahan perpindahan antarmoda merupakan faktor-faktor yang memengaruhi keputusan masyarakat dalam menggunakan transportasi publik. Peningkatan kualitas layanan dapat mendorong peningkatan jumlah pengguna transportasi publik [59].

Dengan memahami keempat faktor tersebut, pengelola transportasi dapat memperoleh gambaran yang lebih menyeluruh mengenai faktor-faktor yang memengaruhi penggunaan layanan transportasi publik, sehingga kebijakan dan pengelolaan layanan dapat dilakukan secara lebih tepat.

2.10 TransJakarta

TransJakarta adalah sistem *Bus Rapid Transit* (BRT) pertama di Asia Tenggara, yang mulai beroperasi secara resmi pada 15 Januari 2004 di Jakarta. Pengelolaan sistem ini berada di bawah PT Transportasi Jakarta, sebuah Badan Usaha Milik Daerah (BUMD) milik Pemerintah Provinsi DKI Jakarta. Kehadiran TransJakarta pada awalnya ditujukan untuk mengatasi persoalan kemacetan dan memenuhi kebutuhan transportasi massal yang terintegrasi di kawasan metropolitan Jakarta [60]. Sebagai BRT pertama di Asia Tenggara, TransJakarta kerap dijadikan objek studi komparatif dalam kajian tata kelola dan integrasi transportasi publik di kawasan metropolitan Asia Tenggara [61].

Dari sisi jaringan, hingga pertengahan 2026, TransJakarta mengoperasikan 233 rute yang didukung 271 halte BRT di seluruh wilayah DKI Jakarta dan daerah penyangganya, dengan sistem integrasi antarmoda yang menghubungkan TransJakarta, MRT, LRT, hingga layanan pengumpan (feeder) dan Mikrotrans [62]. Sepanjang tahun 2025, total penumpang

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

yang dilayani TransJakarta mencapai lebih dari 413 juta perjalanan, dengan rata-rata sekitar 1,4 juta perjalanan per hari [63].

Secara konseptual, setiap perjalanan penumpang pada sistem BRT berbasis kartu elektronik dapat dicatat melalui mekanisme *tap* kartu berbasis *Radio Frequency Identification* (RFID), dengan rekaman transaksi yang berpotensi memuat informasi seperti kode dan nama halte asal-tujuan, waktu transaksi, nomor koridor, serta koordinat geografis halte [64].

Besarnya volume penumpang menunjukkan potensi terjadinya kepadatan pada halte dan periode waktu tertentu sehingga diperlukan pengelolaan operasional yang efektif [65].

Gambar 2. 3 Bus Listrik TransJakarta Koridor 1 Blok M–Kota[66]

Gambar 2.3 menampilkan armada bus listrik TransJakarta yang beroperasi pada Koridor 1 rute Blok M–Kota. Bus listrik ini merupakan bagian dari modernisasi armada TransJakarta yang bertujuan mengurangi emisi karbon dan meningkatkan kualitas layanan transportasi publik di DKI Jakarta. Koridor 1 sendiri merupakan salah satu koridor utama BRT dengan jalur busway eksklusif yang menghubungkan kawasan bisnis Jakarta Selatan dengan pusat kota.

Gambar 2. 4 Mekanisme Tap-On-Bus (TOB) TransJakarta[67]

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

Gambar 2.4 mengilustrasikan mekanisme *Tap-On-Bus (TOB)* yang digunakan oleh penumpang TransJakarta. Pada sistem ini, penumpang melakukan tap kartu elektronik berbasis RFID langsung di dalam bus, bukan di gerbang halte. Setiap transaksi *tap-in* dan *tap-out* secara otomatis merekam data seperti nomor kartu, kode halte, waktu transaksi, dan nomor koridor, sehingga menghasilkan dataset transaksi berskala besar yang menjadi objek kajian pada studi berbasis data transaksi TransJakarta.

2.11 Kerangka Pemikiran

Penelitian ini dilakukan karena adanya perbedaan tingkat kepadatan penumpang pada halte-halte TransJakarta. Studi mengenai pola ridership pada sistem transportasi publik menunjukkan bahwa beberapa titik mengalami jumlah penumpang yang tinggi pada waktu tertentu, sedangkan titik lainnya memiliki tingkat aktivitas yang lebih rendah, sehingga distribusi penumpang belum merata di seluruh jaringan halte. Namun, hingga saat ini belum terdapat pemetaan yang dilakukan secara sistematis untuk mengelompokkan halte berdasarkan karakteristik kepadatan penumpangnya [68].

Untuk mengatasi permasalahan tersebut, penelitian ini menggunakan metode Knowledge Discovery in Databases (KDD) yang terdiri atas tahapan seleksi data, praproses data, transformasi data, data mining, dan evaluasi. Pada tahap data mining digunakan metode clustering karena mampu mengelompokkan data berdasarkan kemiripan karakteristik tanpa memerlukan label kelas sebelumnya. Algoritma HDBSCAN dipilih karena merupakan metode clustering berbasis kepadatan (density-based) yang mampu menangani data dengan tingkat kepadatan yang berbeda-beda serta mengidentifikasi data yang dianggap sebagai noise atau anomali [69].

Variabel yang digunakan dalam penelitian ini terdiri atas faktor spasial, temporal, kalender, dan lingkungan. Faktor spasial direpresentasikan oleh koordinat halte, faktor temporal direpresentasikan oleh waktu transaksi dan proporsi hari kerja, faktor kalender direpresentasikan oleh status hari libur, sedangkan faktor lingkungan direpresentasikan oleh curah hujan. Pemilihan variabel tersebut didasarkan pada temuan bahwa faktor temporal dan kondisi cuaca memengaruhi pola penggunaan transportasi publik [70].

Seluruh data melalui tahap praproses dan normalisasi untuk memastikan setiap variabel memiliki skala yang sebanding. Selanjutnya, dilakukan evaluasi terhadap 24 kombinasi parameter yang diperoleh melalui proses grid search guna memperoleh performa klaster terbaik. Berdasarkan hasil evaluasi tersebut, dipilih parameter dengan nilai Silhouette

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

Coefficient tertinggi sebagai salah satu indeks validitas klaster yang umum digunakan untuk menilai kohesi dan pemisahan antarklaster [71].

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.