

BAB I

PENDAHULUAN

1.1 Latar Belakang

Transportasi publik merupakan elemen penting dalam mendukung mobilitas perkotaan dan pembangunan berkelanjutan di wilayah metropolitan [1]. DKI Jakarta sebagai pusat kegiatan ekonomi memiliki kebutuhan mobilitas yang tinggi sehingga memerlukan sistem transportasi massal yang terintegrasi [2]. Transjakarta sebagai sistem Bus Rapid Transit (BRT) utama di Jakarta melayani lebih dari satu juta penumpang per hari dan menjadi salah satu tulang punggung mobilitas masyarakat perkotaan [3]. Data BPS menunjukkan bahwa jumlah penumpang TransJakarta pada periode Januari–Juni 2025 mencapai 189.277.156 orang [4], sedangkan sepanjang tahun 2025 layanan TransJakarta secara kumulatif melampaui 413 juta perjalanan [5].

Tingginya volume penumpang dan transaksi menunjukkan bahwa distribusi layanan TransJakarta belum merata. Beberapa halte utama mengalami kepadatan tinggi pada waktu-waktu tertentu sehingga meningkatkan tingkat keramaian penumpang dan berpotensi menurunkan kenyamanan perjalanan pengguna transportasi publik [6]. Identifikasi halte dengan tingkat kepadatan transaksi tertinggi berdasarkan data menjadi penting untuk mendukung pengambilan keputusan operasional dan perencanaan layanan TransJakarta yang lebih tepat [7]. Besarnya jumlah data transaksi juga memberikan peluang untuk menerapkan teknik data mining dalam mengidentifikasi pola kepadatan pada setiap halte [8].

Penelitian ini menggunakan dataset transaksi TransJakarta yang diperoleh dari Kaggle [9]. Dataset berformat CSV ini memuat informasi seperti kode halte, nama halte, koridor, waktu transaksi, dan koordinat geografis halte. Dari total 4.000.200 baris data, dilakukan proses seleksi dan penyesuaian berdasarkan kelengkapan atribut sehingga diperoleh data yang siap digunakan dalam penelitian. Pada dataset dengan jumlah data yang besar, terdapat kemungkinan adanya data yang memiliki karakteristik berbeda dari sebagian besar data lainnya (anomali), sehingga diperlukan pendekatan komputasional untuk mengidentifikasinya secara objektif dan terukur [10]. Melalui penerapan data mining dengan metode clustering, penelitian ini dapat mengelompokkan halte berdasarkan tingkat kepadatan transaksi sekaligus mengidentifikasi data yang tidak termasuk ke dalam kluster mana pun (noise). Hasil tersebut diharapkan dapat mendukung peningkatan efisiensi operasional, pemerataan layanan antarhalte dan koridor, serta perencanaan pengembangan layanan TransJakarta yang lebih tepat berdasarkan data.

Sebagai solusi, penelitian ini mengimplementasikan algoritma *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN), yang mampu menangani variasi kepadatan data dan mendeteksi *noise* tanpa perlu menentukan jumlah kluster di awal [11]. Berbeda dengan algoritma *clustering* konvensional seperti K-Means dan DBSCAN, HDBSCAN mampu menangani data dengan variasi kepadatan yang berbeda melalui pendekatan hierarkis. Beberapa penelitian terdahulu menunjukkan bahwa HDBSCAN menghasilkan kualitas kluster yang berbeda dibandingkan dengan metode clustering lainnya. Hal ini ditunjukkan oleh nilai *Silhouette Score* sebesar 0,215 yang lebih tinggi dibandingkan dengan K-Means (0,055) dan Agglomerative Clustering (0,028), serta nilai *Davies-Bouldin Index* sebesar 1,557 yang lebih rendah [12]. Penelitian lain menunjukkan bahwa optimasi parameter HDBSCAN menggunakan *Silhouette Coefficient* menghasilkan nilai evaluasi sebesar 0,6507 serta mampu mengidentifikasi data *noise* dalam proses pembentukan kluster [13]. Selain itu, Kim [14] menunjukkan bahwa kondisi cuaca dan peristiwa kalender memengaruhi pola penggunaan transportasi publik berdasarkan kajian data *smart card*. Berdasarkan penelitian-penelitian tersebut, penelitian ini menerapkan algoritma HDBSCAN pada data transaksi TransJakarta dengan menggunakan fitur yang merepresentasikan lokasi halte, waktu transaksi, rasio hari kerja, kondisi cuaca, status hari libur, dan jumlah transaksi untuk mengelompokkan halte berdasarkan tingkat kepadatannya.

Berdasarkan uraian di atas, penulis melakukan penelitian menggunakan pendekatan data mining berbasis algoritma HDBSCAN dengan judul **“Implementasi Algoritma HDBSCAN Untuk Identifikasi Halte TransJakarta Dengan Tingkat Kepadatan Tertinggi Berdasarkan Data Transaksi”**.

1.2 Rumusan Masalah

Berdasarkan hal tersebut, penelitian ini merumuskan masalah sebagai berikut:

1. Distribusi penumpang antarhalte TransJakarta belum merata sehingga beberapa halte memiliki tingkat kepadatan penumpang yang lebih tinggi dibandingkan dengan halte lainnya. Selain itu, tingkat kepadatan penumpang juga berbeda antara hari kerja dan hari libur nasional.
2. Besarnya volume data transaksi TransJakarta, keberadaan data anomali, serta perbedaan tingkat kepadatan pada setiap halte menjadi tantangan dalam menentukan halte dengan tingkat kepadatan penumpang tertinggi.

1.3 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Mengidentifikasi halte TransJakarta dengan tingkat kepadatan penumpang tertinggi berdasarkan data transaksi.
2. Mengelompokkan halte TransJakarta berdasarkan tingkat kepadatan transaksi.

1.4 Manfaat

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Bagi PT Transportasi Jakarta (TransJakarta): hasil identifikasi halte dengan tingkat kepadatan transaksi tertinggi dapat dimanfaatkan oleh pihak manajemen operasional sebagai landasan dalam menyusun kebijakan penjadwalan, penyesuaian frekuensi keberangkatan, serta redistribusi armada yang lebih proporsional dan berbasis data transaksi nyata.
2. Bagi Pemerintah Provinsi DKI Jakarta dan perencana transportasi: informasi sebaran kepadatan halte secara spasial dan temporal yang dihasilkan oleh penelitian ini dapat digunakan sebagai bahan kajian dalam proses perencanaan dan pengembangan infrastruktur transportasi publik yang lebih merata, terukur, dan berorientasi pada kebutuhan masyarakat.
3. Bagi pengembangan ilmu pengetahuan di bidang informatika dan transportasi: penelitian ini berkontribusi dalam memperkaya literatur ilmiah terkait penerapan algoritma HDBSCAN pada data transaksi transportasi publik berskala besar, sekaligus dapat dijadikan acuan bagi penelitian selanjutnya yang mengkaji persoalan serupa di bidang data mining transportasi.

1.5 Ruang Lingkup

Ruang lingkup penelitian mengenai implementasi algoritma HDBSCAN untuk identifikasi halte TransJakarta dengan tingkat kepadatan tertinggi berdasarkan data transaksi dibatasi sebagai berikut:

1. Objek penelitian adalah halte-halte bus TransJakarta yang tercatat dalam dataset transaksi penumpang pada periode Januari hingga Desember 2023. Data yang digunakan hanya data tap-in, yaitu penumpang yang masuk ke halte, sedangkan data *tap-out* tidak digunakan sehingga kepadatan yang diukur adalah kepadatan penumpang masuk pada setiap halte.

2. Dataset transaksi TransJakarta diperoleh dari Kaggle dalam format CSV dengan total 4.000.200 baris data. Berdasarkan keterangan penyedia dataset, data tersebut merupakan dummy data yang dihasilkan menggunakan library Faker sehingga digunakan untuk keperluan pembelajaran dan penelitian.
3. Data eksternal yang digunakan terdiri dari dua sumber, yaitu data hari libur nasional tahun 2023 dari SKB 3 Menteri [15] dalam bentuk daftar tanggal resmi untuk membentuk fitur *is_holiday*, serta data curah hujan harian Jakarta tahun 2023 yang diperoleh melalui NASA POWER [16] menggunakan *endpoint temporal daily point API* [17] dalam format JSON untuk membentuk fitur cuaca. Kedua data tersebut kemudian diintegrasikan ke dalam dataset utama berdasarkan kesesuaian tanggal transaksi.
4. Sebelum proses clustering, dataset diproses melalui tahap *preprocessing* yang meliputi pembersihan data, penanganan *missing values*, normalisasi *Min-Max*, serta penambahan fitur eksternal.
5. Fitur yang digunakan dalam proses clustering berjumlah tujuh, yaitu:
 - a) *tapInStopsLat*
 - b) *tapInStopsLon*
 - c) *tapInTime*
 - d) *weekday_ratio*
 - e) *is_holiday*
 - f) cuaca
 - g) *transaction_count*Seluruh fitur dinormalisasi ke rentang [0,1] sebelum diproses menggunakan algoritma HDBSCAN.
6. Luaran penelitian berupa pengelompokan halte TransJakarta ke dalam beberapa kluster berdasarkan tingkat kepadatan transaksi, identifikasi halte dengan tingkat kepadatan tertinggi, serta deteksi transaksi yang tergolong sebagai anomali (noise).
7. Tahapan penelitian mengacu pada metodologi *Knowledge Discovery in Databases* (KDD), yaitu:
 - a) Seleksi data (*Data Selection*)
 - b) Praproses data (*Preprocessing*)
 - c) Transformasi data (*Transformation*)
 - d) Data mining menggunakan algoritma HDBSCAN
 - e) Evaluasi dan interpretasi hasil *clustering*.