INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

PENERAPAN ALGORITMA K-MEANS UNTUK KLASTERISASI JUMLAH PENUMPANG ANGKUTAN UMUM DI JAKARTA

Arief Rahman Ramadhan[1], Anita Diana[2*]
[1,2]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta
[1,2]Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan, Jakarta 12260
e-mail : ariefrahmanr12@gmail.com[1] , anita.diana@budiluhur.ac.id[2*]

ABSTRAK

Tingginya mobilitas masyarakat di Jakarta menuntut pengelolaan transportasi umum yang efektif, namun seringkali terjadi penumpukan penumpang pada hari-hari sibuk. Masalah ini disebabkan oleh belum adanya sistem segmentasi data yang mampu memetakan pola jumlah penumpang secara spesifik berdasarkan hari dan jenis moda. Akibatnya, perencanaan operasional seperti alokasi armada dan jadwal layanan menjadi tidak responsif terhadap kebutuhan riil di lapangan. Penelitian ini bertujuan menerapkan metode data mining dengan algoritma K-Means Clustering untuk mengelompokkan data jumlah penumpang harian dari berbagai moda transportasi di Jakarta meliputi Transjakarta, Kereta Rel Listrik (KRL), Mass Rapid Trasit (MRT), Light Rail Transit (LRT), dan Kereta Commuter Indonesia (KCI) Bandara. Penelitian ini menggunakan data sekunder periode 1 Januari 2024 hingga 31 Januari 2025 dari portal satudata.jakarta.go.id, dengan total 2.779 baris data. Analisis dilakukan mengikuti metodologi CRISP-DM, yang mencakup tahap persiapan data seperti normalisasi Min-Max, hingga evaluasi model menggunakan Elbow Method dan Davies-Bouldin Index (DBI). Hasil penelitian menunjukkan bahwa algoritma K-Means secara optimal membentuk 5 klaster dengan nilai DBI terendah sebesar 0.794. Setiap klaster menunjukkan karakteristik yaitu, Klaster 0 penumpang sangat tinggi, dominan Transjakarta pada hari Rabu, Klaster 1 penumpang sangat rendah, dominan Kereta Commuter Indonesia (KCI) Bandara pada hari Selasa, Klaster 2 penumpang tinggi, dominan KRL pada hari Rabu, Klaster 3 penumpang menengah, dominan MRT pada hari Senin, dan Klaster 4 penumpang menengah, dominan LRT pada hari Kamis. Penelitian ini menghasilkan rekomendasi kepada Dinas Perhubungan dan pengelola transportasi umum DKI Jakarta untuk pengambilan keputusan operasional berbasis pola jumlah penumpang harian.

Kata kunci *: Data Mining, K-Means Clustering, Transportasi Umum, Jumlah Penumpang, DKI Jakarta*

ABSTRACT

The high mobility of Jakarta's population demands effective public transportation management; however, passenger congestion frequently occurs on busy days. This issue is caused by the absence of a data segmentation system capable of mapping passenger volume patterns specifically by day and mode of transportation. As a result, operational planning, such as fleet allocation and service scheduling, becomes less responsive to actual needs in the field. This study aims to apply data mining techniques using the K-Means Clustering algorithm to group daily passenger data from various public transportation modes in Jakarta, including Transjakarta, Commuter Line (KRL), Mass Rapid Transit (MRT), Light Rail Transit (LRT), and Kereta Commuter Indonesia (KCI) Airport Line. The research uses secondary data from January 1, 2024, to January 31, 2025, obtained from the portal satudata.jakarta.go.id, totaling 2,779 records. The analysis follows the CRISP-DM methodology, covering data preparation steps such as Min-Max normalization and model evaluation using the Elbow Method and Davies-Bouldin Index (DBI). The results show that the K-Means algorithm optimally forms five clusters with the lowest DBI value of 0.794. Each cluster exhibits distinct characteristics: Cluster 0 very high passenger volume, dominated by Transjakarta on Wednesdays; Cluster 1 very low passenger volume, dominated by KCI Airport Line on Tuesdays; Cluster 2 high passenger volume, dominated by KRL on Wednesdays; Cluster 3 medium passenger volume, dominated by MRT on Mondays; and Cluster 4 medium passenger volume, dominated by LRT on

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

Thursdays. This study provides recommendations to the Jakarta Transportation Agency and public transport operators for operational decision-making based on daily passenger volume patterns.

***Key Words** : Data Mining, K-Means Clustering, Public Transportation, Passenger Volume, DKI Jakarta*

1. PENDAHULUAN

DKI Jakarta sebagai pusat pemerintahan dan kegiatan ekonomi nasional memiliki tingkat mobilitas masyarakat yang sangat tinggi. Untuk menunjang kebutuhan tersebut, pemerintah menyediakan berbagai moda transportasi umum seperti MRT, LRT, dan Transjakarta. Namun, penggunaan moda-moda ini belum sepenuhnya stabil, melainkan mengalami fluktuasi yang cukup signifikan setiap bulannya. Berdasarkan data dari BPS Provinsi DKI Jakarta dalam Berita Resmi Statistik No. 13/03/31/Th. XXVII, 3 Maret 2025, jumlah penumpang MRT Jakarta pada Januari 2025 tercatat sebesar 3.534.665 orang, naik 12,80% dari bulan Desember 2024 yang berjumlah 3.133.700 orang. Lalu untuk jumlah perjalanan MRT pada Januari 2025 sebesar 8.043 perjalanan, mengalami penurunan sebesar 1,30% jika dibandingkan pada bulan Desember 2024 sebesar 8.149 perjalanan. Lalu penumpang LRT pada bulan Januari 2025 sebesar 99.328 orang jika dibandingkan pada bulan Desember 2024 yang mencapai 101.209 orang. Jumlah perjalanan LRT mengalami penurunan sebesar 0,57% pada Januari 2025 yaitu 6.324 perjalanan jika dibandingkan pada bulan Desember 2024 yang mencapai 6.320 perjalanan. Selain itu, jumlah penumpang Transjakarta pada bulan Januari 2025 sebesar 32.227.028, mengalami peningkatan 4,18% jika dibandingkan pada bulan Januari 2024 yang berjumlah 30.934.492 orang. Total bus Transjakarta yang beroperasi pada 2025 berjumlah 4.341 unit, mengalami penurunan 1,23% jika dibandingkan pada Januari 2024 yang berjumlah 4.287 unit.

Salah satu permasalahan utama yang timbul dari kondisi tersebut adalah terjadinya penumpukan atau antrean penumpang pada hari-hari tertentu, terutama di hari sibuk, akibat tidak seimbangnya antara jumlah armada yang tersedia dengan volume permintaan. Di sisi lain, belum tersedia segmentasi berbasis data yang mampu merepresentasikan pola penggunaan moda transportasi secara spesifik menurut jenis moda dan hari operasionalnya. Perencanaan operasional seperti alokasi armada, jadwal layanan, dan manajemen kepadatan belum sepenuhnya mengacu pada pola penggunaan aktual masyarakat yang bervariasi tiap harinya.

Untuk menjawab permasalahan tersebut, penelitian ini dilakukan dengan pendekatan data mining menggunakan metode *K-Means Clustering*, yang bertujuan untuk mengelompokkan data jumlah penumpang harian berdasarkan hari dan jenis moda transportasi. Dengan metode ini, diharapkan dapat diperoleh pola penggunaan transportasi yang lebih terstruktur dan bermanfaat dalam mendukung perencanaan dan pengambilan keputusan yang lebih efektif oleh pihak pengelola.

Metode *K-Means Clustering* merupakan salah satu algoritma dalam unsupervised learning yang umum digunakan untuk membagi data ke dalam beberapa kelompok (cluster). Dalam konteks transportasi, metode ini dapat dimanfaatkan untuk mengelompokkan hari-hari tertentu berdasarkan jumlah penumpang yang tercatat, sehingga terbentuk kategori hari dengan karakteristik jumlah penumpang yang tinggi, sedang, maupun rendah.

Penerapan metode *K-Means Clustering* telah dibuktikan dalam berbagai penelitian di bidang transportasi untuk menganalisis pola penumpang dan mendukung pengambilan keputusan berbasis data. Penelitian oleh (Wicaksono & Prasetyo, 2024) menunjukkan efektivitas algoritma *K-Means* dalam mengelompokkan jumlah penumpang berdasarkan rute dan halte. Hasil penelitian ini menunjukkan bahwa *K-Means* lebih terukur dan efisien dalam membentuk klaster dibandingkan metode AHC, dan dapat digunakan untuk meningkatkan pelayanan angkutan Trans Jateng.

Selain itu, studi oleh (Wibowo et al., 2021) menganalisa perubahan perilaku penumpang TransJakarta sebelum dan selama pandemi COVID-19. Menggunakan gabungan algoritma *K-Means* dan K-Medoids yang dianalisis kembali dengan metode majority voting, studi ini mampu mengidentifikasi jumlah klaster optimal yang berbeda antar kondisi, sehingga mendukung evaluasi dan penyesuaian layanan transportasi selama kondisi darurat.

Studi serupa dilakukan oleh (Permadi & Wiyaja, 2023) pada penelitian mengelola pola perjalanan berdasarkan atribut operasional seperti jumlah penumpang, ritase, dan jarak tempuh, yang hasilnya dimanfaatkan untuk menyusun strategi operasional yang lebih efektif oleh pengelola layanan Bus Biskita.

Kemudian, (Erni et al., 2019) dalam penelitiannya berhasil membentuk lima klaster dari halte-halte BRT berdasarkan jumlah penumpang. Hasil klasterisasi tersebut menjadi dasar dalam

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

menentukan strategi peningkatan kualitas layanan serta pemerataan distribusi penumpang.

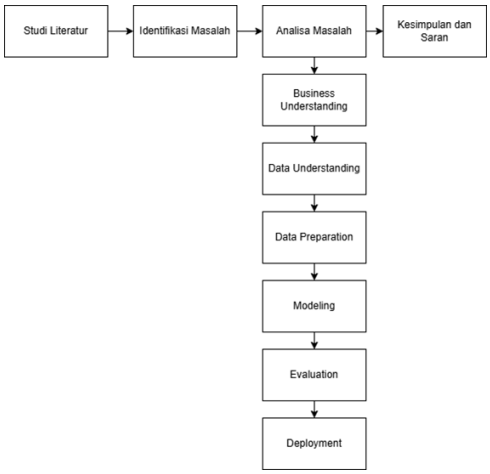
Pada penelitian (Tan et al., 2021) menggabungkan metode *K-Means* dengan GRNN (Generalized Regression Neural Network) untuk memprediksi arus penumpang di Stasiun Chengdu Timur, Tiongkok. Model gabungan ini mampu memetakan dan memprediksi lonjakan penumpang dengan akurasi tinggi yang berguna sebagai sistem peringatan dini.

Melalui berbagai penelitian tersebut, terbukti bahwa metode *K-Means Clustering* mampu menggambarkan karakteristik pola penumpang secara efektif dan dapat digunakan dalam berbagai konteks layanan transportasi. Oleh karena itu, penelitian ini akan menerapkan metode *K-Means* untuk mengklasterisasi jumlah penumpang harian angkutan umum di Jakarta berdasarkan moda transportasi dan hari, guna mendukung strategi perencanaan transportasi yang lebih tepat sasaran.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk memberikan rekomendasi berbasis data kepada Dinas Perhubungan (Dishub) DKI Jakarta dalam mengoptimalkan perencanaan dan pengelolaan transportasi umum harian. Dengan menerapkan metode data mining menggunakan algoritma *K-Means Clustering*, penelitian ini membentuk segmentasi jumlah penumpang berdasarkan moda transportasi dan hari operasional, sehingga dapat mengungkap pola mobilitas masyarakat secara lebih akurat. Informasi ini diharapkan mampu menjadi dasar dalam pengambilan keputusan strategis, seperti penyesuaian alokasi armada dan jadwal layanan, untuk mengurangi penumpukan penumpang pada hari-hari tertentu dan meningkatkan efisiensi layanan transportasi di lapangan.

2. METODE PENELITIAN

Penelitian ini mengadopsi metodologi *Cross Industry Standard Process Model for Data Mining* (CRISP-DM) untuk memastikan proses yang terstruktur dan sistematis dalam setiap langkah analisis data. CRISP-DM terdiri dari enam tahap utama, yang diilustrasikan pada Gambar 1 berikut: (Ibrahim & Usino, 2024).



Gambar 1. Tahapan Penelitian

Berikut adalah penjelasan dari masing-masing tahapan yang digambarkan pada Gambar 1.

a) Studi Literatur
Studi literatur merupakan pondasi awal penelitian, di mana peneliti mengumpulkan dan menganalisis berbagai sumber seperti jurnal ilmiah, buku teks, dan laporan penelitian terdahulu. Tahap ini membantu memperluas pemahaman terhadap teori *clustering*, penerapan metode *K-Means*, serta tantangan di sektor transportasi umum. Hasil studi literatur juga digunakan untuk merumuskan kerangka pemikiran dan mengidentifikasi kesenjangan penelitian.

b) Identifikasi Masalah
Berdasarkan studi literatur dan kondisi lapangan, peneliti mengidentifikasi bahwa belum ada segmentasi hari-hari penggunaan angkutan umum di Jakarta yang konkret untuk mendukung perencanaan layanan. Masalah utama dirumuskan pada kebutuhan untuk mengetahui pola hari ramai, sedang, dan sepi demi efisiensi armada dan jadwal operasional.

c) Analisa Masalah
Analisa masalah dilakukan melalui penerapan enam sub-tahap CRISP-DM:

1. *Business Understanding*
   Memahami kebutuhan pemangku kebijakan, seperti Dinas Perhubungan, untuk memiliki data segmentasi hari sebagai dasar pengambilan keputusan operasional.
2. *Data Understanding*
   Mengeksplorasi data awal jumlah penumpang harian dari berbagai moda transportasi di Jakarta, termasuk validasi kualitas,

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

pemeriksaan nilai ekstrem, dan pemetaan atribut.

3. *Data Preparation*
   Melakukan seleksi variabel penting (tanggal, moda, jumlah penumpang), pembersihan data (menangani data hilang dan outlier), serta transformasi (normalisasi dan encoding) agar cocok untuk algoritma *K-Means*.
4. *Modeling*
   Menentukan algoritma *K-Means* dan melakukan iterasi parameter, seperti jumlah klaster, hingga diperoleh model dengan cluster yang bermakna.
5. *Evaluation*
   Mengevaluasi hasil klasterisasi menggunakan *Elbow Method* dan validasi visual untuk memastikan klaster sesuai karakteristik yang diharapkan.
6. *Deployment*
   Menyusun laporan hasil, visualisasi cluster dalam grafik dan peta interaktif, serta memberikan rekomendasi implementasi kepada pemangku kebijakan.

d) Kesimpulan dan Saran
Bagian ini merangkum temuan utama dari seluruh proses CRISP-DM, termasuk pola klaster yang terbentuk. Dihasilkan rekomendasi praktis untuk pengelola transportasi dalam mengoptimalkan jadwal dan alokasi armada berdasarkan kategori hari.

2.1 Pengumpulan Data

Penelitian ini menggunakan data publik mengenai jumlah penumpang harian angkutan umum di Provinsi DKI Jakarta. Data diperoleh secara online melalui situs resmi Satu Data Jakarta satudata.jakarta.go.id/. Dataset ini berisi data harian penumpang dari beberapa moda transportasi umum yang beroperasi di wilayah DKI Jakarta yang berjumlah 2.779 baris. Rentang waktu data mencakup bulan Januri 2024 sampai Januari 2025. Data tersebut memuat jumlah penumpang dari masing-masing moda transportasi, seperti Transjakarta, KRL, MRT, LRT, kapal, dan bus sekolah. Informasi ini diperoleh dari sistem pemantauan terintegrasi yang disediakan oleh Dinas Perhubungan Provinsi DKI Jakarta.

periode_data	tanggal	jenis_moda	jumlah_penumpang_per_hari
202401	01/01/2024	transjakarta	632778
202401	01/01/2024	bus sekolah	0
202401	01/01/2024	krl	870760
202401	01/01/2024	mrt	57856
202401	01/01/2024	lrt	3166
202401	01/01/2024	KCI Commuter Bandara	6723
202401	01/01/2024	kapal	14321
202401	02/01/2024	transjakarta	1045214
202401	02/01/2024	bus sekolah	40401
202401	02/01/2024	krl	988279
202401	02/01/2024	mrt	99620
202401	02/01/2024	lrt	2901
202401	02/01/2024	KCI Commuter Bandara	6920
202401	02/01/2024	kapal	4932

Gambar 2. Tabel Dataset

Adapun atribut-atribut dalam dataset tersebut adalah sebagai berikut:

a. periode_data: Informasi tahun dan bulan data dicatat, yang menunjukkan periode waktu pengambilan data jumlah penumpang.
b. tanggal: Tanggal pencatatan jumlah penumpang.
c. jenis_moda: Menyatakan jenis moda transportasi umum yang digunakan, seperti Transjakarta, bus sekolah, MRT, KRL, LRT, KCI Commuter Bandara dan kapal.
d. total_penumpang: Jumlah keseluruhan penumpang dari seluruh moda transportasi per hari (kolom hasil akumulasi).

2.2 Data Pra-processing

Sebelum dilakukan pengolahan data, diperlukan proses data preprocessing atau pra-pemrosesan data untuk memperoleh hasil yang optimal dari proses data mining. Pada penelitian ini, proses data *preprocessing* mencakup *data cleaning* jumlah penumpang harian angkutan umum, *data integration*, serta transformasi data agar dapat digunakan pada tahap pemodelan klasterisasi. Tahapan ini penting untuk memastikan kualitas dan kesiapan data sebelum dianalisis menggunakan algoritma *K-Means* (Sudarsono et al., 2021).

1. *Data Cleaning*
   Pada tahap data cleaning, dilakukan serangkaian proses untuk memastikan kualitas data. Pengecekan terhadap nilai kosong (*missing values*) menunjukkan bahwa seluruh atribut terisi lengkap. Kolom periode_data dihapus karena tidak berpengaruh langsung pada analisis klasterisasi, mengingat informasi waktu sudah terwakili oleh atribut tanggal. Data dengan jenis moda "kapal" dihapus karena tidak termasuk dalam cakupan transportasi umum darat di DKI Jakarta, sedangkan moda "bus sekolah" dihapus karena mayoritas nilainya nol, terutama pada

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

akhir pekan saat layanan tidak beroperasi. Nilai nol dalam jumlah besar dapat menurunkan kualitas analisis dan memicu pembentukan klaster berdasarkan ketidakaktifan, sehingga penghapusan kedua moda ini dilakukan agar klasterisasi lebih bersih, fokus, dan relevan dengan tujuan penelitian. Proses *data cleaning* ini dilakukan menggunakan bahasa pemrograman Python melalui platform Google Colab, yaitu layanan berbasis cloud yang memungkinkan penulisan dan eksekusi kode Python langsung melalui peramban, dilengkapi dukungan GPU/TPU serta fitur kolaboratif sehingga memudahkan peneliti dalam melakukan data preprocessing secara fleksibel (Ridwan Nazar, 2024).

2. *Data Integration*
   Data intergration adalah menggabungkan dataset dari berbagai moda menjadi satu tabel terpadu. Tetapi, Pada penelitian ini tidak dilakukan proses penggabungan data karena seluruh informasi sudah tersedia dalam satu dataset terpadu dan tidak berasal dari sumber atau tabel berbeda.
3. *Data Transformation*
   Kemudian pada tahap data transformation, dilakukan beberapa penyesuaian agar data dapat digunakan dalam pemodelan algoritma *K-Means*. Pertama, kolom baru ditambahkan untuk menampilkan nama hari dari atribut tanggal. Setelah itu, data kategorikal pada atribut hari dan jenis moda diubah ke dalam bentuk numerik menggunakan teknik encoding.

Jenis Moda Transportasi	Nilai Numerik
KCI Commuter Bandara	0
KRL	1
LRT	2
MRT	3
Transjakarta	4

Gambar 3. Transformasi Kategori Transportasi

Jenis Moda Transportasi	Nilai Numerik
Minggu	0
Senin	1
Selasa	2
Rabu	3
Kamis	4
Jumat	5
Sabtu	6

Gambar 4. Transformasi Kategori Hari

Terakhir, dilakukan proses normalisasi menggunakan metode *Min-Max* terhadap atribut jumlah penumpang per hari, hari yang sudah dikonversi ke numerik, serta jenis moda yang juga telah berbentuk numerik. Langkah ini dilakukan agar semua atribut berada dalam skala yang sama dan tidak mendominasi proses klasterisasi.

2.3 Algoritma *K-Means*

Algoritma *K-Means clustering* merupakan metode analisis cluster yang memecah objek menjadi k *cluster* berdasarkan kedekatan ke pusat *cluster* (*centroid*). Algoritma ini bersifat sederhana dan mudah dipahami (Dinata et al., 2020). Secara lebih spesifik *K-Means* dapat dijelaskan sebagai berikut:

1. Menetapkan K sebagai jumlah *cluster* yang optimal.
2. Menentukan atau menghasilkan nilai acak untuk pusat cluster awal (*centroid*) sebanyak k.
3. Menggunakan rumus jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak *Euclidean Distance* hingga ditemukan jarak yang paling pendek dari setiap data dengan *centroid*. persamaan *Euclidean Distance* antara lain :

$$d(xi, \mu j) = \sqrt{\sum(xi - \mu j)^2} \quad (1)$$

   Keterangan:
   xi: data kriteria
   μj: *centroid* pada *cluster* ke-j
4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terdekat). Memperbaharui nilai *centroid*. Menurut (Ibrahim & Usino, 2024), nilai *centroid* baru diperoleh dari rata-rata cluster yang bersangkutan dengan menggunakan rumus:

$$C_k = \frac{1}{n_k}\sum d_i \quad (2)$$

   Keterangan:
   nk :Jumlah data dalam *cluster* k
   di : Jumlah nilai jarak yang masuk dalam masing-masing *cluster*
5. Melakukan perulangan dari langkah 2 hingga 4 sampai anggota tiap cluster tidak ada yang berubah.
6. Jika langkah terakhir telah terpenuhi, maka nilai pusat cluster (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

2.4 Penentuan Data Latih dan Data Uji

Penelitian ini menggunakan algoritma *K-Means Clustering* yang merupakan bagian dari *unsupervised learning.* Karena tidak terdapat variabel target atau labeling, maka seluruh data yang tersedia digunakan secara langsung dalam proses pemodelan tanpa pemisahan eksplisit antara data latih dan data uji seperti pada *supervised learning*.

Dataset yang digunakan terdiri dari 2.779 baris data yang merepresentasikan jumlah penumpang harian angkutan umum di Jakarta. Seluruh data ini digunakan untuk membentuk pola pengelompokan berdasarkan karakteristik jumlah penumpangnya masing-masing.

2.5 Rapid Miner

Proses pemodelan dan evaluasi pada penelitian ini menggunakan perangkat lunak RapidMiner, yaitu salah satu aplikasi *open source* yang banyak dimanfaatkan untuk analisis data, text mining, dan prediksi. RapidMiner menyediakan lebih dari 500 operator untuk berbagai keperluan, mulai dari *input, output, preprocessing*, hingga visualisasi. Dengan antarmuka berbasis *Graphical User Interface* (GUI), RapidMiner memudahkan pengguna dalam merancang proses analisis secara sistematis tanpa harus menulis kode pemrograman secara manual (Ubaidillah et al., 2024).

3. HASIL DAN PEMBAHASAN

3.1 Hasil Komparasi Model

Komparasi model dilakukan untuk menentukan jumlah klaster yang paling optimal dengan membandingkan nilai *Davies Bouldin Index* (DBI). Nilai DBI yang lebih rendah menunjukkan kualitas klasterisasi yang lebih baik, karena mencerminkan jarak antar klaster yang besar dan kekompakan dalam klaster yang tinggi. Pada penelitian ini, dilakukan perbandingan dua algoritma, yaitu *K-Means* dan K-Medoids, dengan berbagai jumlah klaster. Proses ini dilakukan menggunakan RapidMiner. Hasil evaluasi ditampilkan dalam tabel berikut sebagai dasar pemilihan model terbaik.

1. Pemodelan dengan Algoritma *K-Means*

   Pemodelan awal dilakukan dengan menggunakan algoritma *K-Means*. Penentuan jumlah cluster yang optimal diuji dengan 4 variasi jumlah cluster, yaitu 2, 3, 4, dan 5 cluster. Nilai DBI dari masing-masing jumlah cluster ditampilkan pada Gambar 5 berikut.

Jumlah *Cluster*	Nilai DBI
2 *cluster*	1,274
3 *cluster*	1,136
4 *cluster*	0,925
5 *cluster*	0,794

Gambar 5. Data Hasil Davies Bouldin Index pada Algoritma K-means

Berdasarkan hasil tersebut, diketahui bahwa nilai DBI terendah terdapat pada jumlah 5 cluster, yaitu sebesar 0.794. Karena DBI yang paling kecil menunjukkan kualitas klasterisasi terbaik, maka dapat disimpulkan bahwa model *K-Means* paling optimal jika menggunakan 5 cluster.

2. Pemodelan dengan Algoritma *K-Medoids*

   Selanjutnya dilakukan pemodelan menggunakan algoritma *K-Medoids*. Uji coba dilakukan pada jumlah cluster yang sama, yaitu 2, 3, 4, dan 5 cluster. Hasil nilai DBI untuk masing-masing cluster ditampilkan pada Gambar 6 berikut:

Jumlah *Cluster*	Nilai DBI
2 *cluster*	1,647
3 *cluster*	1,625
4 *cluster*	1,368
5 *cluster*	1,292

Gambar 6. Nilai Davies Bouldin Index K-Medoids

Dari tabel di atas, dapat dilihat bahwa nilai DBI terendah terdapat pada 5 cluster, yaitu sebesar 1.292, namun angka ini masih lebih tinggi dibandingkan seluruh nilai DBI pada algoritma *K-Means.*

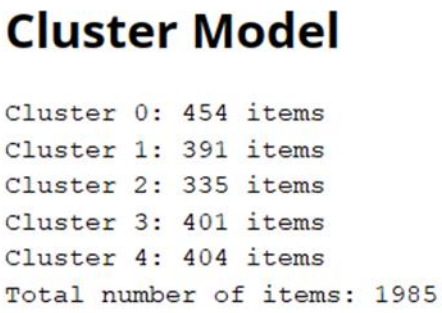
Dari hasil perbandingan kedua algoritma, *K-Means* dengan 5 cluster menghasilkan nilai DBI paling rendah yaitu sebesar 0,794 dibandingkan algoritma *K-Medoids.* Oleh karena itu, algoritma *K-Means* dipilih sebagai metode klasterisasi terbaik untuk penelitian ini, dengan jumlah cluster optimal sebanyak 5 cluster.

3.2 Penyajian Model Terbaik

Setelah dilakukan proses evaluasi pemodelan menggunakan algoritma *K-Means* dan dibandingkan dengan algoritma K-Medoids, diperoleh bahwa model terbaik adalah algoritma *K-Means* dengan jumlah 5 cluster. Hal ini didasarkan pada nilai *Davies Bouldin Index* (DBI) terendah, yaitu sebesar 0.794, yang menunjukkan bahwa model ini memiliki pemisahan antar cluster yang baik serta kekompakan internal yang tinggi pada setiap cluster yang terbentuk. Proses klasterisasi dilakukan terhadap data

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

yang telah melalui tahap normalisasi menggunakan metode *Min-Max Normalization*. Adapun atribut yang digunakan dalam proses pemodelan ini terdiri dari tiga atribut, yaitu Jumlah Penumpang per Hari, Hari (yang telah dikonversi dalam bentuk numerik), Jenis Moda Transportasi (yang telah dikodekan dalam bentuk numerik). Setelah proses klasterisasi dilakukan, diperoleh hasil distribusi data ke dalam empat cluster, dengan rincian Gambar 7 sebagai berikut



Gambar 7. Hasil Jumlah *Cluster*

Untuk memahami karakteristik dari masing-masing cluster, dilakukan analisis lebih lanjut berdasarkan nilai minimum, rata-rata, dan maksimum dari setiap atribut (jumlah penumpang per hari, hari, dan jenis moda transportasi). Hasil analisis tersebut disajikan dalam tabel-tabel pada Gambar 8 berikut.

		jumlah_penumpang_per_hari	hari	jenis_moda
Cluster 0	Min	471864	0	4
	Rata-rata	1050548,956	3,017	4
	Max	1307318	6	4
Cluster 1	Min	0	0	0
	Rata-rata	6974,805	2,997	0,008
	Max	390448	6	1
Cluster 2	Min	451270	0	1
	Rata-rata	897143,941	3,007	1
	Max	1209506	6	1
Cluster 3	Min	0	0	1
	Rata-rata	59555,834	1,5	2,506
	Max	550347	3	4
Cluster 4	Min	2122	4	1
	Rata-rata	58005,510	4,997	2,495
	Max	557040	6	3

Gambar 8. Data Jumlah Minimal, Maksimal, dan Rata-rata Pada Setiap Atribut Data

Berdasarkan hasil pemodelan klasterisasi menggunakan algoritma *K-Means*, diperoleh sebanyak 5 klaster dari total 1.985 data. Masing-masing klaster dibentuk berdasarkan atribut jumlah penumpang per hari, hari, dan jenis moda transportasi yang telah melalui proses transformasi dan normalisasi. Analisis dilakukan untuk mengetahui karakteristik masing-masing klaster, sehingga dapat memberikan gambaran pola penggunaan angkutan umum di DKI Jakarta sebagai berikut:

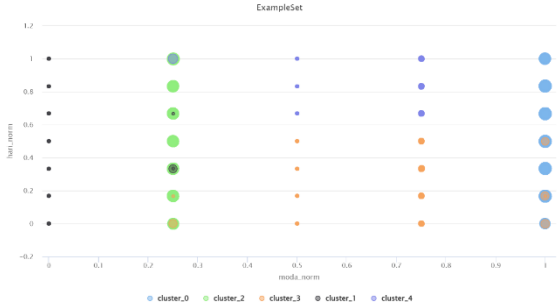
1. *Cluster* 0
   Cluster ini menunjukkan jumlah penumpang sangat tinggi, dengan rata-rata mencapai 1.050.548 penumpang per hari. Moda transportasi dominan pada cluster ini adalah Transjakarta, dan hari dominan berada pada nilai 3,017 jika dibulatkan menjadi nilai 3, yang menunjukkan hari Rabu. Pola ini merepresentasikan hari kerja puncak di tengah pekan, di mana mobilitas masyarakat tinggi dan penggunaan moda utama meningkat tajam.
2. *Cluster* 1
   Cluster ini memiliki rata-rata penumpang paling rendah, yakni 6.974 penumpang per hari. Moda dominan berada di nilai 0,008, mengarah kuat pada KCI Bandara. Hari dominan berada di angka 2,997 jika dibulatkan menjadi nilai 3, yang merujuk pada hari Rabu. *Cluster* ini mencerminkan hari-hari dengan mobilitas rendah dan moda yang digunakan cenderung ringan serta bersifat pendukung, bukan utama.
3. *Cluster* 2
   Rata-rata penumpang harian dalam cluster ini berada di angka 897.143, tergolong tinggi namun tidak setinggi cluster 0. Moda dominan adalah KRL (nilai moda: 1), dan hari dominan berada di 3,007 jika dibulatkan menjadi nilai 3, yaitu Rabu. Pola ini menunjukkan bahwa pada hari-hari menjelang akhir pekan, terjadi peningkatan aktivitas yang cukup tinggi dengan penggunaan moda yang lebih ringan atau menengah seperti KRL.
4. *Cluster* 3
   Cluster ini memiliki rata-rata penumpang sebesar 59.555 per hari. Moda dominan berada di nilai 2,506, yang mencerminkan moda MRT dan LRT, dan hari dominan berada pada nilai 1,5 jika dibulatkan menjadi nilai 2, yang berarti Selasa. Ini menunjukkan hari kerja awal minggu dengan mobilitas sedang dan moda kelas menengah yang digunakan masyarakat untuk beraktivitas.
5. *Cluster* 4
   Rata-rata penumpang per hari di cluster ini mencapai 58.005, tergolong menengah. Moda dominan berada di angka 2,495, yaitu LRT dan MRT, sementara hari dominan berada di 4,997 jika dibulatkan menjadi nilai 5, mendekati Jumat. *Cluster* ini mengindikasikan lonjakan aktivitas menjelang akhir pekan dengan penggunaan moda transportasi menengah.

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 2, Desember 2025
eISSN : 2580-3042
pISSN : 1979-0694

Cluster	Jumlah Penumpang	Moda Dominan	Hari Dominan
0	Sangat Tinggi	Transjakarta	Rabu
1	Rendah	KCI Bandara	Rabu
2	Tinggi	KRL	Rabu
3	Menengah	MRT	Selasa
4	Menengah	LRT	Jumat

Gambar 9. Karakteristik Setiap *Cluster*

3.3 Visualisasi Data



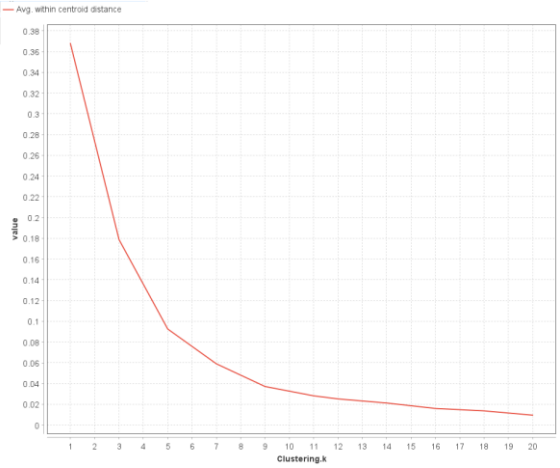
Gambar 10. Visualisasi Scatter Plot

Visualisasi pada Gambar 10 merupakan hasil scatter plot dari proses klasterisasi dengan enam klaster menggunakan algoritma *K-Means*. Grafik ini memetakan data berdasarkan variabel moda_norm pada sumbu X dan hari_norm pada sumbu Y, dengan warna yang membedakan tiap klaster dan ukuran titik mewakili jumlah penumpang.

Visualisasi *scatter plot* klasterisasi memperlihatkan distribusi data berdasarkan nilai normalisasi atribut moda transportasi (moda_norm) dan hari (hari_norm). Setiap klaster menunjukkan karakteristik penggunaan moda transportasi yang berbeda-beda berdasarkan intensitas penggunaan dan waktu operasionalnya. *Cluster* 0 (ditandai warna biru muda) berada pada nilai moda_norm dan hari_norm yang tinggi, mencerminkan dominasi moda utama seperti Transjakarta yang banyak digunakan pada hari kerja padat, khususnya hari Rabu, dengan volume penumpang sangat tinggi. Sebaliknya, *Cluster* 1 (berwarna hitam) memiliki nilai moda_norm dan hari_norm yang relatif rendah, menunjukkan dominasi moda ringan seperti KCI Bandara yang lebih aktif digunakan pada hari dengan aktivitas rendah seperti hari Selasa, serta jumlah penumpang yang rendah. *Cluster* 2 (warna hijau) terletak pada nilai moda_norm dan hari_norm di tingkat menengah, yang menunjukkan dominasi moda KRL pada hari Rabu dengan jumlah penumpang yang tinggi namun tidak sepadat cluster 0. Adapun *Cluster* 3 (berwarna oranye) tersebar pada nilai moda_norm menengah ke bawah dan hari_norm rendah, merepresentasikan moda MRT yang digunakan pada awal minggu, terutama hari Senin, dengan jumlah penumpang yang tergolong menengah. Terakhir, *Cluster* 4 (berwarna ungu) terkonsentrasi pada nilai moda_norm dan hari_norm menengah hingga tinggi, mencerminkan dominasi moda LRT pada hari Kamis dengan volume penumpang yang juga menengah.

3.4 Pengujian

Bagian ini menjelaskan proses evaluasi jumlah klaster optimal dalam klasterisasi data. Evaluasi dilakukan dengan menggunakan *Elbow Method* dengan menggunakan perangkat lunak RapidMiner. Hasil visualisasi dari *Elbow Method* ditunjukkan pada Gambar 11.



Gambar 11. Grafik Elbow Method

Gambar 11 menunjukkan hasil *Elbow Method* yang menggambarkan hubungan antara jumlah klaster (k) dan nilai *average within centroid distance*. Grafik tersebut memperlihatkan penurunan nilai yang tajam dari klaster 2 hingga klaster 5. Setelah jumlah klaster 5, penurunan nilai menjadi lebih landai, yang mengindikasikan bahwa penambahan jumlah klaster selanjutnya tidak lagi memberikan pengurangan signifikan terhadap jarak rata-rata ke pusat klaster. Titik "siku" atau elbow yang cukup jelas terlihat pada k = 5, sehingga berdasarkan metode ini, jumlah klaster yang optimal adalah 5 klaster.

4. KESIMPULAN

Penelitian ini menggunakan algoritma *K-Means Clustering* pada data jumlah penumpang harian angkutan umum di Jakarta periode Januari 2024–Januari 2025 dengan bantuan RapidMiner. Model terbaik diperoleh pada jumlah klaster 5 dengan nilai *Davies Bouldin Index* (DBI) sebesar 0,794, menunjukkan pemisahan klaster yang cukup

baik. Hasil klasterisasi mengungkap bahwa hari Rabu menjadi puncak aktivitas penumpang di berbagai moda, khususnya Transjakarta dan KRL. Rekomendasi yang dihasilkan meliputi penambahan armada dan optimalisasi jadwal pada moda dengan volume tinggi (misalnya Transjakarta dan KRL pada hari Rabu), efisiensi operasional pada moda dengan volume rendah (seperti KCI Bandara), serta penyesuaian jadwal atau kapasitas pada moda dengan volume menengah (MRT dan LRT) agar layanan lebih responsif terhadap pola mobilitas masyarakat. Temuan ini dapat menjadi dasar bagi pengambil kebijakan, khususnya Dinas Perhubungan DKI Jakarta, untuk mengoptimalkan alokasi armada dan jadwal layanan.

5. REFERENSI

Dinata, R. K., Safwandi, S., Hasdyna, N., & Azizah, N. (2020). Analisis K-Means Clustering pada Data Sepeda Motor. *INFORMAL: Informatics Journal*, *5*(1), 10. https://doi.org/10.19184/isj.v5i1.17071

Erni, D., Putri, P. Y., & Yulia, S. (2019). Klustering Jumlah Penumpang pada Halte Bus Rapid Transit Kota Tangerang. In *Jurnal Sistem Cerdas* (Vol. 02, Issue 03).

Ibrahim, I., & Usino, W. (2024). KLASTERISASI TINGKAT KELAYAKAN PROVINSI DALAM PEMBANGUNAN KAWASAN INDUSTRI MENGGUNAKAN ALGORITMA K-MEANS. *Prosiding Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, *3*(2), 324–333.

Jurnal, H., Ubaidillah, ruf, & Fatah, Z. (2024). JURNAL ILMIAH MULTIDISIPLIN ILMU IMPLEMENTASI RAPIDMINER PADA KLASTERISASI GEMPA BUMI DI INDONESIA BERDASARKAN KEDALAMAN MENGGUNAKAN K-MEANS. *JURNAL ILMIAH MULTIDISIPLIN ILMU*, *6*, 84–91. https://doi.org/10.69714/w0m9zv32

Permadi, A., & Wiyaja, Y. A. (2023). Pengelompokan Terbaik Menggunakan Algoritma K-Means Pada Dataset Bus Biskita Bogor. *INTERNAL (Information System Journal)*, *6*(1), 88–100.

Ridwan Nazar. (2024). IMPLEMENTASI PEMROGRAMAN PYTHON MENGGUNAKAN GOOGLE COLAB. *Jurnal Informatika Dan Komputer*, *15*, 50–56.

Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). ANALISIS DATA MINING DATA NETFLIX MENGGUNAKAN APLIKASI RAPID MINER. *JBASE - Journal of Business and Audit Information Systems*, *4*(1). https://doi.org/10.30813/jbase.v4i1.2729

Tan, Y., Liu, H., Pu, Y., Wu, X., & Jiao, Y. (2021). Passenger Flow Prediction of Integrated Passenger Terminal Based on K-Means–GRNN. *Journal of Advanced Transportation*, *2021*(1), 1055910.

Wibowo, A., Moh Makruf, Inge Virdyna, & Farah Chikita Venna. (2021). Penentuan Klaster Koridor TransJakarta dengan Metode Majority Voting pada Algoritma Data Mining. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(3), 565–575. https://doi.org/10.29207/resti.v5i3.3041

Wicaksono, M. R. J., & Prasetyo, S. Y. J. (2024). Analisis Pengelompokan Jumlah Penumpang Bus Trans Jateng Menggunakan Metode Clustering K-Means Dan Agglomerative Hierarchical Clustering (AHC). *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer Dan Manajemen)*, *5*(3), 1346–1354.