

Penerapan GMM dan ARM untuk Analisis Mobilitas Penumpang TransJakarta

Arya Ade Wiguna[#], Rudy Cahyadi^{*}

[#] Program Studi Informatika, Fakultas Teknik Industri, Universitas Pembangunan Nasional “Veteran” Yogyakarta,
Jl. Babarsari No. 2, Tambakbayan, Caturtunggal, Depok, Sleman, DIY, Indonesia
E-mail: 123220058[at]student.upnyk.ac.id; RudyCahyadi[at]upnyk.ac.id

ABSTRACTS

TransJakarta passenger mobility shows dynamic patterns influenced by travel time, duration, service-use intensity, and route connectivity. This study constructs probabilistic passenger mobility segments and analyzes inter-corridor travel associations within each segment. The data consist of TransJakarta tap-in and tap-out transactions processed through data quality assessment, time transformation, filtering, feature engineering, missing value imputation, data cleaning, and variable normalization. Passenger segmentation was performed using Gaussian Mixture Model with model selection based on Bayesian Information Criterion, while travel association patterns were analyzed using Association Rule Mining. The final dataset consisted of 168,132 observations, and the selected GMM formed five mobility clusters. The ARM results show that the most numerous and stable association rules appeared in commuter segments, especially Evening Commuter and Early Morning Commuter. These findings indicate that combining GMM and ARM provides a structured understanding of passenger behavior segmentation and inter-corridor travel relationships.

Manuscript received Jun 20, 2026;
revised Jul 27, 2026; accepted Sep
20, 2026 Date of publication Sep
30, 2026. International Journal,
JITSI : Jurnal Ilmiah Teknologi
Sistem Informasi licensed under a
Creative Commons Attribution-
Share Alike 4.0 International
License



ABSTRAK

Mobilitas penumpang TransJakarta memiliki pola yang dinamis karena dipengaruhi oleh waktu perjalanan, durasi, intensitas penggunaan layanan, dan keterhubungan rute. Penelitian ini bertujuan membentuk segmentasi mobilitas penumpang secara probabilistik dan menganalisis keterkaitan perjalanan antar koridor pada setiap segmen. Data yang digunakan berupa transaksi tap-in dan tap-out TransJakarta yang diproses melalui evaluasi kualitas data, transformasi waktu, filtering, feature engineering, imputasi, pembersihan data, dan normalisasi variabel. Segmentasi dilakukan menggunakan Gaussian Mixture Model dengan pemilihan model berdasarkan Bayesian Information Criterion, sedangkan pola keterkaitan perjalanan dianalisis menggunakan Association Rule Mining. Hasil penelitian menunjukkan bahwa dataset akhir berjumlah 168.132 observasi dan model GMM terbaik membentuk lima kluster mobilitas. Hasil ARM menunjukkan aturan asosiasi paling banyak dan stabil muncul pada segmen commuter, terutama Commuter Sore dan Commuter Pagi Dini. Temuan ini memperlihatkan bahwa kombinasi GMM dan ARM dapat memberikan pemahaman yang lebih terstruktur terhadap segmentasi perilaku penumpang dan keterkaitan perjalanan antar koridor.

Keywords / Kata Kunci — *TransJakarta, Gaussian Mixture Model, Association Rule Mining, segmentasi mobilitas, pola perjalanan.*

CORRESPONDING AUTHOR

Arya Ade Wiguna
Fakultas Teknik Industri, Universitas Pembangunan Nasional “Veteran” Yogyakarta, DIY, Indonesia
Email: 123220058[at]student.upnyk.ac.id

1. PENDAHULUAN

Sistem transportasi perkotaan membutuhkan pembacaan data yang mampu menangkap variasi perilaku perjalanan penumpang. Pada layanan TransJakarta, data transaksi tap-in dan tap-out menyimpan informasi waktu, lokasi asal dan tujuan, durasi perjalanan, serta intensitas penggunaan layanan. Informasi tersebut penting karena pola mobilitas penumpang tidak selalu bersifat seragam; sebagian perjalanan menunjukkan rutinitas commuter, sedangkan sebagian lain lebih kasual dan tersebar.

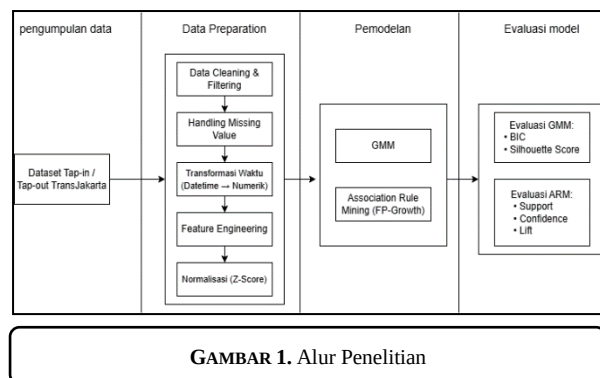
Penelitian terdahulu mengenai TransJakarta banyak menggunakan pendekatan clustering deterministik seperti K-Means untuk membaca pola penyebaran pengguna (Daffauzan, 2023; Wibowo et al., 2021). Pendekatan tersebut bermanfaat sebagai gambaran awal, tetapi setiap observasi tetap ditempatkan pada satu cluster secara tegas. Pada data mobilitas yang memiliki karakter waktu dan ruang, pendekatan clustering spatio-temporal juga menunjukkan bahwa pola perjalanan dapat dibaca melalui kedekatan lokasi dan waktu secara bersamaan (Birant & Kut, 2007).

Artikel ini menawarkan integrasi Gaussian Mixture Model (GMM) dan Association Rule Mining (ARM) pada data mobilitas TransJakarta. GMM dipilih karena mampu membentuk cluster probabilistik pada data yang tidak selalu terpisah secara tegas, sedangkan ARM digunakan untuk membaca keterkaitan pola perjalanan antar koridor. Dasar pemodelan GMM juga sejalan dengan kajian model campuran Gaussian yang menekankan representasi data melalui beberapa komponen probabilistik (Chassagnol et al., 2023; Scrucca et al., 2016).

Sifat probabilistik pada GMM menjadi penting karena perilaku mobilitas penumpang tidak selalu memiliki batas klaster yang tegas. Seorang penumpang dapat menunjukkan karakteristik yang beririsan, misalnya memiliki jam perjalanan seperti commuter tetapi frekuensi perjalanan yang tidak selalu rutin, atau memiliki durasi perjalanan panjang tetapi pola harian yang lebih kasual. Pada kondisi seperti ini, pendekatan deterministik seperti K-Means hanya memberikan satu label klaster, sedangkan GMM menghasilkan probabilitas keanggotaan terhadap beberapa klaster. Probabilitas tersebut membantu membaca tingkat kedekatan suatu observasi terhadap pola mobilitas tertentu, sehingga hasil segmentasi lebih sesuai untuk data perjalanan perkotaan yang heterogen dan saling tumpang tindih

2. METODOLOGI PENELITIAN

Penelitian menggunakan pendekatan eksploratif berbasis data mining dengan alur kerja yang merujuk pada tahapan CRISP-DM, mulai dari pemahaman data, data preparation, pemodelan, evaluasi, hingga interpretasi hasil (Fadillah, 2015). Alur penelitian pada Gambar 1 mencakup pengumpulan data transaksi tap-in dan tap-out, data preparation, pemodelan GMM, evaluasi cluster, penerapan ARM, dan interpretasi pola mobilitas.



GAMBAR 1. Alur Penelitian

TABEL 1. Perubahan jumlah data

Tahap	Jumlah	Keterangan
Data awal	189.500	Transaksi tap-in dan tap-out sebelum pembersihan
Data preparation	168.132	Hasil filtering, imputasi, outlier handling, dan normalisasi
Pemodelan GMM	168.132	Observasi akhir dengan enam variabel pemodelan

Dataset awal terdiri atas 189.500 transaksi yang bersumber dari Kaggle <https://www.kaggle.com/datasets/dikisahkan/transjakarta-transportation-transaction>. Tahap data preparation meliputi penilaian kualitas data, transformasi variabel waktu, filter durasi perjalanan, filter jam operasional, feature engineering, imputasi nilai hilang, pembersihan outlier, dan normalisasi variabel kontinu. Setelah proses tersebut, dataset akhir yang digunakan untuk pemodelan berjumlah 168.132 observasi.

Enam fitur digunakan dalam pemodelan GMM, yaitu jam tap-in, durasi perjalanan, jumlah perjalanan, jumlah hari aktif bulanan, indikator akhir pekan, dan indikator commuter. Variabel kontinu dinormalisasi menggunakan z-score agar berada pada skala yang sebanding, sedangkan variabel biner dipertahankan dalam bentuk 0 dan 1.

Seleksi model Gaussian Mixture Model dilakukan menggunakan Bayesian Information Criterion (BIC). Perhitungan BIC pada Persamaan (1) digunakan untuk membandingkan kandidat jumlah cluster dan tipe

kovarians, sehingga model yang dipilih mempertimbangkan kecocokan model terhadap data dan kompleksitas parameter.

TABEL 2. Variabel pemodelan Gaussian Mixture Model

Variabel	Makna
Jam tap-in	Waktu awal perjalanan penumpang
Durasi perjalanan	Selisih waktu tap-in dan tap-out
Jumlah perjalanan	Intensitas penggunaan layanan
Jumlah hari aktif	Rutinitas penggunaan dalam periode data
Akhir pekan	Indikator perjalanan pada akhir pekan
Commuter	Indikator pola perjalanan rutin

Konfigurasi ini lebih sesuai untuk data mobilitas karena pola perjalanan tiap segmen dapat memiliki variasi dan arah sebaran yang berbeda.

$$BIC = 2\ln(L) - k\ln(n) \quad (1)$$

Keterangan :

BIC = nilai Bayesian Information Criterion.

$\ln(L)$ = nilai log-likelihood model.

k = jumlah parameter model.

n = jumlah observasi yang digunakan dalam pemodelan.

Evaluasi aturan asosiasi dilakukan dengan tiga metrik utama. Support pada Persamaan (2) menunjukkan proporsi transaksi yang memuat antecedent dan consequent secara bersamaan, confidence pada Persamaan (3) menunjukkan peluang consequent muncul ketika antecedent terjadi, sedangkan lift pada Persamaan (4) membandingkan kekuatan aturan dengan kejadian independen antecedent dan consequent.

$$support(A \rightarrow B) = support(A \cup B)/|T| \quad (2)$$

$$confidence(A \rightarrow B) = support(A \cup B)/support(A) \quad (3)$$

$$lift(A \rightarrow B) = confidence(A \rightarrow B)/support(B) \quad (4)$$

Keterangan:

A = antecedent atau item pada sisi kiri aturan asosiasi.

B = consequent atau item pada sisi kanan aturan asosiasi.

T = himpunan seluruh transaksi yang dianalisis.

$support(A \rightarrow B)$ = proporsi transaksi yang memuat A dan B secara bersamaan.

$confidence(A \rightarrow B)$ = peluang B muncul pada transaksi yang memuat A.

$support(B)$ = proporsi transaksi yang memuat consequent B.

$lift(A \rightarrow B)$ = rasio kekuatan asosiasi A terhadap B dibandingkan kemunculan B secara independent

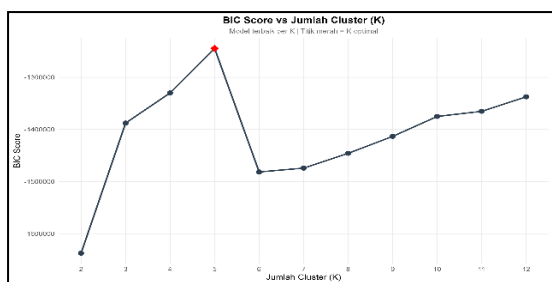
3. HASIL DAN PEMBAHASAN

Hasil data preparation menunjukkan bahwa data akhir berjumlah 168.132 observasi dari 189.500 transaksi awal. Pengurangan terjadi karena proses filtering, imputasi, pembersihan data, dan penanganan outlier. Dataset akhir tetap mempertahankan informasi utama perjalanan, sehingga memadai untuk membaca pola mobilitas berbasis rute dan waktu.

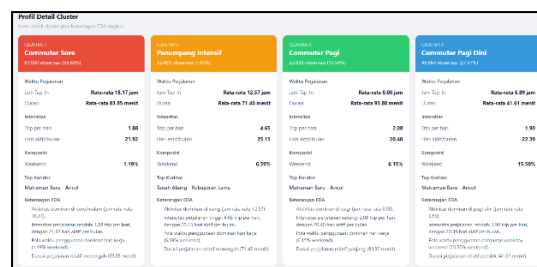
Pada tahap pemodelan, konfigurasi GMM optimal diperoleh pada K=5 dengan tipe kovarians VVV dan nilai BIC -1.244.451,49. Kurva BIC pada Gambar 2 memperlihatkan bahwa penambahan jumlah kluster sampai lima komponen memberikan peningkatan kecocokan model yang paling relevan untuk data penelitian. Nilai BIC terbaik pada konfigurasi K=5 dengan kovarians VVV mengindikasikan bahwa peningkatan fleksibilitas bentuk kluster masih memberikan keseimbangan terbaik antara kecocokan model dan kompleksitas parameter.

Evaluasi kualitas clustering menunjukkan bahwa nilai silhouette tertinggi diperoleh pada K=5, yaitu 0,3956. Walaupun nilai tersebut menunjukkan adanya irisan antarsegmen, kondisi ini sejalan dengan karakteristik mobilitas perkotaan yang tidak selalu membentuk kelompok yang benar-benar terpisah. Dalam konteks tersebut, GMM lebih sesuai dibanding pemisahan deterministik karena tetap mempertahankan informasi probabilitas keanggotaan.

Interpretasi kluster pada Gambar 3 menunjukkan lima segmen utama, yaitu Commuter Pagi Dini, Commuter Pagi, Commuter Sore, Penumpang Intensif, dan Penumpang Kasual. Commuter Sore menjadi segmen terbesar dengan 61.930 observasi atau 36,83% dari dataset final. Commuter Pagi Dini berjumlah 46.863 observasi atau 27,87%, sedangkan segmen lain menunjukkan pola intensitas, durasi, dan akhir pekan yang berbeda.



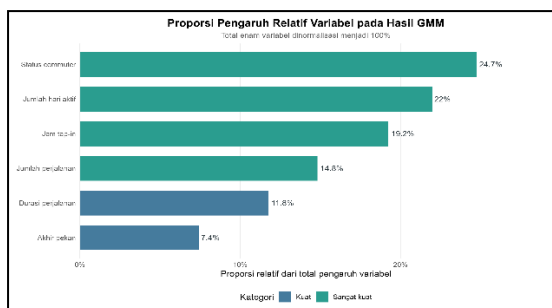
GAMBAR 2. Kurva BIC terhadap jumlah cluster



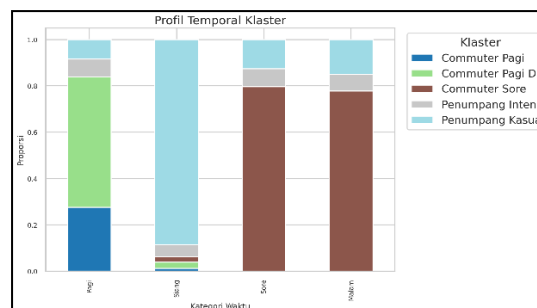
GAMBAR 3. Profil detail cluster hasil Gaussian Mixture Model

Perbedaan tersebut menunjukkan bahwa GMM mampu membentuk segmen yang tidak hanya berbeda secara waktu, tetapi juga berbeda dari sisi intensitas dan rutinitas perjalanan. Variabel status commuter, jumlah hari aktif, dan jam tap-in menjadi pembeda penting, sehingga penamaan klaster dapat dikaitkan dengan pola penggunaan layanan yang terlihat pada data. Gambar 4 menunjukkan proporsi pengaruh relatif setiap variabel terhadap pembentukan klaster GMM. Status commuter memiliki kontribusi relatif terbesar sebesar 24,74%, diikuti jumlah hari aktif sebesar 21,99%, jam tap-in sebesar 19,25%, jumlah perjalanan sebesar 14,82%, durasi perjalanan sebesar 11,77%, dan indikator akhir pekan sebesar 7,42%. Urutan ini menegaskan bahwa pembentukan klaster terutama dipengaruhi oleh rutinitas, intensitas, dan waktu perjalanan penumpang.

Tahap ARM menggunakan transaksi yang telah diberi label klaster untuk melihat pola keterkaitan perjalanan pada masing-masing segmen. Ringkasan evaluasi menunjukkan bahwa aturan paling banyak muncul pada segmen commuter. Commuter Sore menghasilkan 24 aturan terpilih, sedangkan Commuter Pagi Dini menghasilkan 21 aturan. Commuter Pagi menghasilkan 2 aturan, sementara Penumpang Intensif dan Penumpang Kasual tidak menghasilkan aturan terpilih pada parameter yang digunakan.



GAMBAR 4. Proporsi Pengaruh Relatif Variabel pada Hasil GMM



GAMBAR 5. Profil temporal klaster berdasarkan kategori jam

Analisis temporal pada Gambar 5 memperlihatkan bahwa dominasi segmen berubah menurut kategori waktu. Pada pagi hari, Commuter Pagi Dini memiliki proporsi tertinggi sebesar 56,44%. Pada siang hari, Penumpang Kasual mendominasi sebesar 88,55%. Pada sore dan malam hari, Commuter Sore menjadi segmen utama dengan proporsi masing-masing 79,58% dan 77,69%. Pola ini memperlihatkan bahwa hasil segmentasi GMM konsisten dengan interpretasi perilaku perjalanan harian.

TABEL 3. Ringkasan evaluasi Association Rule Mining

Klaster	Aturan	Interpretasi
Commuter Sore	24	Pola asosiasi paling kuat dan stabil
Commuter Pagi Dini	21	Pola rutin awal hari cukup jelas
Commuter Pagi	2	Aturan terpilih terbatas
Penumpang Intensif	0	Pola rute lebih tersebar pada parameter yang digunakan
Penumpang Kasual	0	Perjalanan kasual tidak membentuk aturan kuat

Secara integratif, hasil GMM dan ARM menunjukkan bahwa segmentasi probabilistik dapat menjadi dasar untuk membaca pola perjalanan yang lebih operasional. GMM membedakan karakteristik pengguna berdasarkan waktu, durasi, intensitas, dan indikator perilaku, sedangkan ARM memperjelas koridor atau pasangan origin-destination yang dominan di dalam setiap segmen. Temuan ini dapat digunakan sebagai pertimbangan awal dalam evaluasi layanan, terutama pada segmen commuter yang memiliki pola perjalanan lebih stabil.

Kekuatan utama pendekatan gabungan ini terletak pada urutan analisisnya. Jika ARM diterapkan langsung pada seluruh transaksi, aturan yang terbentuk berpotensi didominasi rute dengan volume tinggi dan mengaburkan

perbedaan perilaku antar penumpang. Dengan menempatkan GMM sebagai tahap segmentasi awal, aturan asosiasi dapat dibaca dalam konteks kelompok yang lebih homogen. Akibatnya, pola pada Commuter Sore, Commuter Pagi Dini, dan segmen lain tidak lagi diperlakukan sebagai satu agregat besar, tetapi sebagai karakter perjalanan yang memiliki waktu, intensitas, dan kecenderungan rute berbeda.

Meskipun demikian, hasil penelitian tetap perlu ditafsirkan sebagai pola empiris berbasis data transaksi, bukan rekomendasi operasional langsung. Nilai support, confidence, dan lift menunjukkan kekuatan keterkaitan dalam data, tetapi belum memasukkan faktor layanan seperti kapasitas armada, headway, kepadatan halte, kondisi cuaca, atau perubahan jadwal. Oleh karena itu, keluaran GMM dan ARM lebih tepat diposisikan sebagai dasar awal untuk mengenali segmen dan koridor prioritas sebelum dilakukan analisis operasional yang lebih rinci

4. KESIMPULAN

Penelitian ini menunjukkan bahwa kombinasi Gaussian Mixture Model dan Association Rule Mining mampu menghasilkan pemahaman yang lebih terstruktur terhadap mobilitas penumpang TransJakarta. Dataset akhir berjumlah 168.132 observasi, dan model GMM optimal diperoleh pada K=5 dengan tipe kovarians VVV serta nilai silhouette 0,3956. Lima segmen yang terbentuk memperlihatkan variasi temporal, intensitas penggunaan, dan kecenderungan perjalanan. ARM kemudian memperjelas pola origin-destination dan dominasi koridor pada tiap segmen, terutama pada kelompok commuter yang menghasilkan aturan asosiasi paling stabil.

Penelitian selanjutnya dapat memperluas data ke periode yang lebih panjang, menguji model pada data operasional terbaru, serta mengintegrasikan variabel eksternal seperti cuaca, hari libur, kepadatan wilayah, atau headway layanan. Evaluasi spasial juga dapat diperdalam dengan data jaringan halte dan koridor agar rekomendasi layanan dapat dibuat lebih presisi

REFERENSI

- [1] Badan Pusat Statistik. (2023). Statistik mobilitas penduduk dan tenaga kerja 2023. Badan Pusat Statistik. <https://www.bps.go.id/>
- [2] Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1), 208-221. <https://doi.org/10.1016/j.datak.2006.01.013>
- [3] Chassagnol, B., Bichat, A., Boudjeniba, C., Wuillemin, P., Guedj, M., Gohel, D., Nuel, G., & Becht, E. (2023). Gaussian Mixture Models in R. 15(June), 56-76.
- [4] Daffauzan, M. F. (2023). Analisa penumpang bus TransJakarta dengan metode K-Means. *Journal Computer Science and Information Systems: J-Cosys*, 3(2), 61-64. <https://doi.org/10.53514/jco.v3i2.395>
- [5] Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). 1, 260-270.
- [6] Lathifah, S. M. (2023). Pengaruh faktor ekonomi dan sosial pada pola pengguna transportasi publik TransJakarta menggunakan metode association rule. *Skripsi*.
- [7] Nugrahanti, D. A. (2022). Implementasi text association rules terhadap cyberterrorism di Twitter. *Repository UIN Jakarta*.
- [8] Patel, E., & Kushwaha, D. S. (2020). Clustering cloud workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*, 171, 158-167. <https://doi.org/10.1016/j.procs.2020.04.017>
- [9] Saputra, A., & Yusuf, R. (2024). Perbandingan algoritma DBSCAN dan K-Means dalam segmentasi pelanggan pengguna transportasi publik TransJakarta menggunakan metode RFM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(4), 1346-1361. <https://doi.org/10.57152/malcom.v4i4.1516>
- [10] Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1), 289-317.
- [11] Zamri, N. A., Jaya, M. I., Yaakob, S. S., Amnur, H., & Kasim, S. (2024). Multidimensional indicator for data quality assessment in wireless sensor networks: Challenges and opportunities. *International Journal on Advanced Science, Engineering and Information Technology*, 14(5), 1663-1672. <https://doi.org/10.18517/ijaseit.14.5.11492>
- [12] Wibowo, A., Makruf, M., Virdyna, I., & Venna, F. C. (2021). Penentuan klaster koridor TransJakarta dengan metode majority voting. *Jurnal Teknologi Informasi*, 1(10), 565-575.
- [13] Reynaldi, R., Faiz Djarot, R. J., Wahyudi, M., Sumanto, S., & Budiman, A. S. (2025). Analisa pola penyebaran pengguna layanan TransJakarta dengan metode K-Means Clustering. *Journal Software, Hardware and Information Technology*, 5(2), 128-138. <https://doi.org/10.24252/shift.v5i2.205>