Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 3 July 2025, pp: 1029-1036
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Analysis of Passenger and Bus Numbers Using Unsupervised Learning

Analisis Hubungan Antara Jumlah Bus dan Jumlah Penumpang Menggunakan *Unsupervised Learning*

Aurell Octaviona Sinaga[1*], Rossi Passarella[2]

[1]Faculty of Computer Science, Sriwijaya University, Indonesia

E-Mail: [1]vionaaurellie@gmail.com, [2]passarella.rossi@unsri.ac.id

Received Mar 26th 2025; Revised Jun 22th 2025; Accepted Jul 02nd 2025; Available Online Jul 31th 2025, Published Jul 31th 2025
Corresponding Author: Shella Norma Windrasari

Abstract

The relationship between the number of buses and the number of passengers is an important aspect in urban transportation analysis. However, the distribution patterns and data clusters that form are often uneven, causing an imbalance in this relationship. This study aims to identify distribution patterns and cluster data on the number of buses and passengers using an unsupervised learning approach. The methods tested consist of four models: K-Means, Fuzzy C-Means (FCM), Gaussian Mixture Model (GMM), and Spectral Clustering. These four models are compared to measure how well the models can cluster data and reveal the relationship between the number of buses and the number of passengers. The four models will be evaluated using the Calinski-Harabasz Index, Silhouette Score, and Davies-Bouldin Index to find the optimal cluster. Based on the testing of the four clustering models using the three evaluation matrices, all models showed that the optimal cluster was 2; however, the K-Means model performed best in grouping the data because it had the best values for each evaluation metric. The K-Means model score on the Silhouette Score was 0.5175, the K-Means model value on the Davies-Bouldin Index was 0.7241, and the score for K-Means on the Calinski-Harabasz Index was 1414.3874. Cluster 1 represents a high number of buses and passengers, while Cluster 2 represents a low number of passengers and buses. Therefore, the results of this study can serve as a reference for the Jakarta city government in redistributing buses from areas with low demand to areas with high demand for operational efficiency.

Keyword: Fuzzy C-Means, Gaussian Mixture Model, K-Means, Spectral Clustering, Unsupervised Learning.

Abstrak

Hubungan antara jumlah bus dan jumlah penumpang merupakan aspek penting dalam analisis transportasi perkotaan. Namun, pola distribusi dan kelompok data yang terbentuk sering kali tidak merata, sehingga menyebabkan ketidakseimbangan dalam hubungan tersebut. Penelitian ini bertujuan untuk mengidentifikasi pola distribusi dan mengelompokkan data jumlah bus dan penumpang menggunakan pendekatan *unsupervised learning*. Metode yang diuji coba terdiri dari empat model yaitu K-Means, Fuzzy C-Means (FCM), Gaussian Mixture Model (GMM), dan Spectral Clustering. Keempat model tersebut dibandingkan untuk mengukur seberapa baik model dapat mengelompokkan data dan mengungkap hubungan antara jumlah bus dan jumlah penumpang. Keempat model akan dievaluasi menggunakan Calinski-Harabasz Index, Silhouette Score, dan Davies-Bouldin Index untuk menemukan klaster optimal. Berdasarkan uji coba keempat model *clustering* menggunakan ketiga matriks evaluasi, semua model menunjukkan klaster optimalnya adalah 2 namun model K-Means memberikan kinerja terbaik dalam mengelompokkan data karena model K-Means memiliki nilai terbaik untuk setiap metrik evaluasi tersebut. Skor model K-Means pada Silhouette Score sebesar 0.5175, nilai model K-Means pada Davies-Bouldin Index sebesar 0.7241, dan skor untuk K-Means terhadap Calinski-Harabasz Index sebesar 1414.3874. Klaster 1 merepresentasikan jumlah bus dan jumlah penumpang yang tinggi sedangkan Klaster 2 merepresentasikan jumlah penumpang dan jumlah bus yang rendah. Sehingga hasil dari penelitian ini dapat menjadi acuan bagi pemerintah kota Jakarta dalam melakukan redistribusi bus dari wilayah dengan kebutuhan rendah ke wilayah dengan permintaan tinggi untuk efisiensi operasional.

Kata Kunci: Fuzzy C-Means, Gaussian Mixture Model, K-Means, Spectral Clustering, *Unsupervised Learning*.

1. PENDAHULUAN

Keberlanjutan mobilitas perkotaan sangat bergantung pada sistem angkutan yang efisien dan ramah lingkungan. Di kota-kota besar, peran angkutan umum menjadi krusial dalam mengurangi tingkat kemacetan [1] serta menekan emisi gas buang yang dihasilkan oleh kendaraan pribadi. Berbagai kota metropolitan, termasuk Jakarta, terus mengembangkan sistem transportasi umum guna meningkatkan aksesibilitas dan kenyamanan bagi masyarakat [2]. Sebagai pusat aktivitas dengan mobilitas tinggi, Jakarta tidak hanya menghadapi tantangan dalam penyediaan angkutan bagi pekerja, tetapi juga bagi pelajar. Salah satu solusi yang diterapkan untuk memenuhi kebutuhan mobilitas siswa adalah penyediaan bus sekolah. Pemerintah daerah telah menginisiasi kebijakan terkait layanan bus sekolah guna mendukung akses pendidikan yang lebih aman dan terjangkau bagi peserta didik. Seiring meningkatnya jumlah siswa di Jakarta, kebutuhan akan layanan transportasi sekolah yang andal juga semakin mendesak. Namun, pelaksanaan sistem ini masih menghadapi berbagai kendala, seperti kemacetan yang memperlambat perjalanan, efisiensi rute yang belum optimal [3], serta keterbatasan jumlah armada yang tersedia. Oleh karena itu, evaluasi dan pengembangan lebih lanjut terhadap sistem transportasi sekolah menjadi suatu kebutuhan agar dapat memberikan layanan yang lebih baik bagi para pelajar.

Ketersediaan armada bus sekolah memiliki keterkaitan erat dengan jumlah penumpang yang menggunakannya, di mana konsep supply and demand [4] juga berlaku dalam sistem transportasi umum, termasuk layanan bus sekolah. Jumlah bus yang memadai diharapkan mampu meningkatkan jumlah siswa yang memanfaatkannya sebagai moda transportasi utama, meskipun faktor lain seperti kualitas layanan, ketepatan waktu, serta cakupan rute turut menentukan tingkat partisipasi penumpang. Oleh karena itu, memahami hubungan antara jumlah bus yang tersedia dengan jumlah penumpang yang menggunakannya menjadi aspek krusial dalam pengelolaan transportasi sekolah [5]. Seiring dengan berkembangnya era digitalisasi, pemanfaatan teknologi analisis data menjadi kebutuhan utama dalam meningkatkan efisiensi sistem transportasi, termasuk dalam perencanaan bus sekolah di Jakarta. Namun, masih terdapat keterbatasan dalam pemanfaatan teknologi untuk memahami keterkaitan antara jumlah bus dan jumlah penumpangnya secara mendalam. Oleh sebab itu, penelitian ini berupaya menerapkan pendekatan berbasis machine learning [6] guna mengidentifikasi pola hubungan yang dapat mendukung pengambilan keputusan yang lebih optimal, sehingga layanan bus sekolah dapat dioptimalkan sesuai dengan kebutuhan siswa dan menciptakan sistem transportasi yang lebih efisien, berkelanjutan, serta mendukung mobilitas pelajar secara aman dan nyaman.

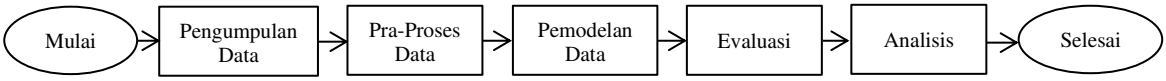
Dalam upaya menganalisis dan mengoptimalkan sistem transportasi sekolah, pendekatan *unsupervised learning* dalam *machine learning* [7] menawarkan metode yang efektif untuk menemukan pola tersembunyi dalam data. *Unsupervised learning* [8] merupakan cabang pembelajaran mesin yang bekerja tanpa label atau variabel target, memungkinkan sistem untuk mengelompokkan data berdasarkan karakteristik tertentu. Keunggulan pendekatan ini terletak pada kemampuannya dalam mengidentifikasi pola kompleks dan struktur tersembunyi dalam data transportasi, sehingga dapat digunakan untuk mengelompokkan rute, menganalisis persebaran penumpang, serta mengevaluasi efisiensi layanan bus sekolah. Berbagai studi telah menerapkan *unsupervised learning* dalam bidang transportasi, seperti analisis pola perjalanan dan optimasi rute, yang menunjukkan potensi besar dalam meningkatkan efektivitas sistem transportasi. Terdapat dua penelitian sebelumnya yang telah mengkaji penerapan teknik *machine learning* dalam menganalisis kendaraan ataupun lalu lintas. Penelitian pertama adalah "Analisis Model Prediksi untuk Layanan Bus Sekolah Jakarta Menggunakan Pendekatan Machine Learning," [9]. Penelitian ini bertujuan memprediksi jenis bus sekolah berdasarkan jumlah penumpang dan jumlah sekolah di Jakarta dari tahun 2017-2019. Ada tujuh model yang diuji coba menggunakan f1-score, dimana model terbaiknya adalah Gradient Boosting . Lalu penelitian kedua yang juga relevan terhadap penelitian ini yaitu "Analisis Model *Clustering* Kecelakaan Lalu Lintas Di Kota Palembang Menggunakan Pendekatan Machine Learning" [10] bertujuan mengelompokkan data kecelakaan lalu lintas di Palembang dari tahun 2020 sampai 2022 berdasarkan karakteristik insiden dan jumlah korban. Penelitian ini menguji coba empat metode *clustering*, dimana metode Spectral Clustering dalam mengelompokkan jenis kecelakaan dan lokasi rawan.

Dari kedua penelitian sebelumnya yang relevan terhadap penelitian ini, memiliki beberapa kekurangan dan keterbatasan. Penelitian pertama hanya berfokus pada prediksi jenis bus, tidak menganalisis hubungan atau pola kelompok dari jumlah penumpang dan bus itu sendiri. Tidak adanya pendekatan unsupervised learning, sehingga tidak menggali struktur data atau pengelompokan alami (*clustering*). Kurang mempertimbangkan dinamika waktu (tahun ke tahun), padahal datanya berupa time-series. Eksplorasi yang kurang mendalam terhadap distribusi bus dan penumpang yang harusnya bisa digunakan untuk perencanaan layanan. Lalu kekurangan pada penelitian kedua yaitu tidak berfokus pada transportasi massal atau pelayanan publik seperti bus sekolah. Tidak mengkaitkan hasil klaster dengan aspek perencanaan kebutuhan layanan (misalnya distribusi bus atau rute). Kekurangan dan perbedaan dari kedua penelitian tersebut menjadi celah bagi penelitian ini dalam menerapkan pendekatan *unsupervised learning* untuk menganalisis hubungan antara jumlah bus dan jumlah penumpang sekolah di Jakarta. Pendekatan ini dipilih karena mampu mengidentifikasi pola tersembunyi dalam data tanpa memerlukan variabel target, sehingga lebih fleksibel dalam memahami keterkaitan antara kedua variabel tersebut.

Sebagai langkah awal dalam penerapan pendekatan baru tersebut, diperlukan perumusan permasalahan yang jelas guna mengidentifikasi pola hubungan antara jumlah bus sekolah dan jumlah penumpang secara sistematis menggunakan metode *unsupervised learning*. Penelitian ini berfokus pada analisis tren jumlah bus dan jumlah penumpang sekolah di Jakarta selama periode 2017–2019 untuk menentukan apakah terdapat pola tertentu yang dapat diidentifikasi. Beberapa faktor utama yang menjadi perhatian meliputi fluktuasi jumlah bus yang beroperasi, distribusi penumpang pada berbagai rute, serta potensi pengelompokan data berdasarkan karakteristik tertentu. Untuk menjawab permasalahan tersebut, penelitian ini mengajukan pertanyaan utama, seperti bagaimana tren jumlah bus sekolah dan jumlah penumpang dari tahun ke tahun, apakah teknik *clustering* dapat mengungkap pola tersembunyi dalam data, serta bagaimana hasil analisis ini dapat berkontribusi dalam perencanaan sistem transportasi sekolah yang lebih efisien di Jakarta.

2. METODOLOGI PENELITIAN

Bagian ini menjelaskan kerangka penelitian yang terdiri atas beberapa tahapan sistematis untuk menganalisis hubungan antara jumlah bus dan jumlah penumpang.Proses analisis divisualisasikan melalui Gambar 1 untuk memberikan pemahaman yang lebih komprehensif terkait pola dan distribusi data.

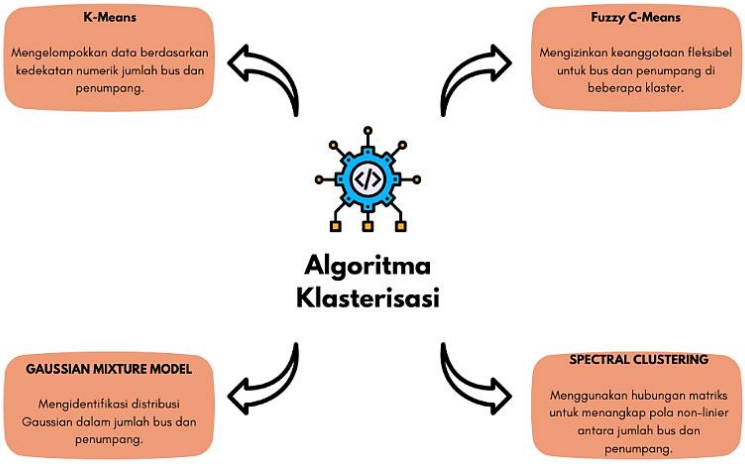


Gambar 1. Kerangka Kerja Penelitian

Penelitian ini menggunakan dataset dari penelitian sebelumnya[9] dimana data tersebut diolah kembali untuk dijadikan penelitian yang berbeda. Data tersebut dipahami menggunakan teknik Analisis Data Eksploratif (EDA) dengan salah satu fungsinya yaitu DataAnalyzer. Selanjutnya proses pengolahan dataset meliputi penghilangan fitur yang tidak terpakai, pengecekkan *missing values*[11], *label encoding*[12], dan matriks korelasi[13]. Setelah didapatkan wawasan baru dan data sepenuhnya siap dipakai, dilanjutkan dengan menguji dataset ke dalam empat algoritma klasterisasi yaitu K-Means, Fuzzy C-Means, Gaussian Mixture Model (GMM), dan Spectral Clustering. Dimana setelah diuji coba ke masing-masing algoritma klasterisasi akan dilakukan analisis lebih lanjut terhadap hasil klaster dari proses tersebut untuk ditarik sebuah kesimpulan terhadap pengujian algoritma klasterisasi dalam menganalisis hubungan jumlah penumpang dengan jumlah bus.

2.1. Pemilihan Algoritma Klasterisasi

Pada tahap ini menjelaskan tujuan dari pemilihan proses klasterisasi yang diuji coba terhadap empat algoritma klaster yang dipilih yaitu K-Means, Fuzzy C-Means, Gaussian Mixture Model (GMM), dan Spectral Clustering. Dimana akan bagian ini menggambarkan perbandingan kinerja atau cara kerja setiap algoritma ke dalam peta konsep seperti Gambar 2.



Gambar 2. Peta Konsep Algoritma Klasterisasi

2.2. K-Means

K-Means adalah algoritma *unsupervised learning* yang digunakan untuk melakukan *clustering* dengan cara membagi data ke dalam *K* kelompok berdasarkan kemiripan karakteristik. Algoritma ini bekerja dengan menentukan *K* centroid secara acak, kemudian menghitung jarak setiap data ke centroid terdekat,

mengelompokkan data berdasarkan kedekatan tersebut, dan memperbarui posisi centroid hingga konvergen. Dalam analisis hubungan antara jumlah bus dan jumlah penumpang sekolah di Jakarta, K-Means[14] dapat digunakan untuk mengelompokkan pola penggunaan bus berdasarkan jumlah penumpang yang diangkut. Misalnya, dataset dapat dikelompokkan ke dalam beberapa klaster yang merepresentasikan tingkat kepadatan penggunaan bus, sehingga dapat diidentifikasi wilayah atau waktu dengan jumlah penumpang tinggi maupun rendah. Persamaan matematis dari model K-Means dijelaskan pada persamaan (1):

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - \mu_i \right\|^2 \quad (1)$$

2.3. Fuzzy C-Means (FCM)

FCM [15] adalah salah satu algoritma *clustering* dalam *unsupervised learning* yang memungkinkan setiap data memiliki derajat keanggotaan dalam lebih dari satu klaster. Berbeda dengan K-Means yang mengelompokkan data secara eksklusif, FCM memberikan nilai keanggotaan berdasarkan tingkat kedekatan data terhadap setiap centroid. Dalam analisis hubungan jumlah bus dan jumlah penumpang sekolah di Jakarta, FCM dapat digunakan untuk mengidentifikasi pola penggunaan bus dengan lebih fleksibel. Misalnya, suatu daerah mungkin tidak secara tegas masuk ke dalam satu klaster tetapi memiliki kemungkinan menjadi bagian dari beberapa klaster dengan tingkat kepastian yang berbeda. Hal ini berguna dalam memahami pola perubahan jumlah penumpang yang mungkin bervariasi tergantung pada hari atau jam tertentu. Dengan sifatnya yang lebih lembut dalam menentukan batas klaster, FCM lebih cocok digunakan jika data memiliki transisi yang tidak jelas antar kelompok dibandingkan metode seperti K-Means yang bersifat lebih kaku dalam pengelompokan. Persamaan matematis dari Fuzzy C-Means dijelaskan pada persamaan (2) dan (3). Derajat Keanggotaan *Fuzzy*:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

Pusat Klaster:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (3)$$

2.4. Gaussian Mixture Model (GMM)

GMM [16] adalah algoritma *clustering* berbasis probabilistik yang mengasumsikan bahwa data berasal dari kombinasi beberapa distribusi Gaussian. Berbeda dengan metode seperti K-Means dan Fuzzy C-Means, GMM tidak hanya mempertimbangkan jarak ke centroid tetapi juga distribusi probabilitas dari setiap klaster, sehingga memungkinkan pengelompokan yang lebih fleksibel. Dalam konteks analisis hubungan jumlah bus dan jumlah penumpang sekolah di Jakarta, GMM dapat digunakan untuk mengidentifikasi pola kelompok dengan mempertimbangkan variasi dan sebaran data yang lebih kompleks. Misalnya, data yang memiliki klaster dengan bentuk lonjong atau tumpang tindih dapat dikelompokkan dengan lebih akurat dibandingkan metode berbasis jarak semata. Keunggulan utama GMM dibandingkan K-Means dan FCM adalah kemampuannya dalam menangani klaster dengan bentuk yang tidak hanya bulat serta memberikan probabilitas keanggotaan yang lebih realistis untuk setiap titik data, sehingga cocok digunakan pada dataset yang memiliki distribusi yang tidak homogen. Persamaan matematis dari Gaussian ditunjukkan pada persamaan (4):

$$P(x) = \sum_{k=1}^{K} \pi_k \cdot N(x \mid \mu_k, \sum_k) \quad (4)$$

2.5. Spectral Clustering

Spectral Clustering [17] adalah algoritma pengelompokan yang memanfaatkan teknik aljabar linier untuk membentuk klaster berdasarkan struktur data dalam ruang berdimensi tinggi. Berbeda dengan metode seperti K-Means yang bergantung pada asumsi bentuk klaster yang cenderung sferis, Spectral Clustering bekerja dengan merepresentasikan data dalam bentuk graf dan memanfaatkan eigenvektor dari matriks ketetanggaan untuk menentukan pemisahan klaster. Dalam analisis hubungan antara jumlah bus dan jumlah penumpang sekolah di Jakarta, metode ini dapat digunakan untuk mengidentifikasi pola perjalanan berdasarkan hubungan antar titik data yang tidak selalu berbentuk linear. Misalnya, wilayah tertentu mungkin memiliki karakteristik penggunaan bus yang serupa dengan beberapa area lain meskipun secara geografis tidak berdekatan, sehingga pendekatan berbasis graf dapat lebih efektif dalam menangkap pola tersebut. Keunggulan utama Spectral Clustering dibandingkan dengan K-Means, Fuzzy C-Means (FCM), dan Gaussian Mixture Model (GMM) adalah kemampuannya dalam mengelompokkan data dengan struktur yang kompleks dan tidak

berbentuk tegas, terutama ketika klaster tidak dapat dipisahkan secara eksplisit dalam ruang berdimensi rendah. Persamaan matematis dari Spectral akan dijelaskan dalam persamaan (5):

$$L = D - W \tag{5}$$

2.6. Evaluasi Algoritma Klasterisasi

Setelah proses klasterisasi dilakukan, akan didapatkan wawasan tentang klaster optimal untuk setiap algoritma klaster yang diuji coba. Kualitas dari klaster optimal pada keempat algoritma ini akan dinilai menggunakan tiga metrik pengukuran yang sama terdiri dari Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI).

1. Silhouette Score[18] mengukur seberapa baik suatu data berada dalam klaster yang benar dibandingkan dengan klaster lain. Metrik ini dihitung menggunakan rumus pada persamaan (6):

$$S(i) = \frac{b(i) - a(i)}{max(a(i)b(i))} \tag{6}$$

   Metrik ini cocok digunakan karena menilai keseimbangan antara kepadatan klaster dan pemisahannya, menjadikannya efektif untuk menentukan jumlah klaster optimal di berbagai algoritma seperti K-Means, Fuzzy C-Means, GMM, dan Spectral Clustering.

2. DBI[19] mengukur kualitas klaster berdasarkan rasio antara penyebaran dalam klaster dan jarak antar-klaster. Rumus perhitungannya dijelaskan pada persamaan (7):

$$DBI = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i}\left(\frac{s_i + s_j}{d_{ij}}\right) \tag{7}$$

   Metrik ini dinilai cocok karena mempertimbangkan kepadatan dan pemisahan antar-klaster, sehingga dapat mengevaluasi apakah klaster terlalu menyebar atau terlalu dekat dalam berbagai metode klasterisasi.

3. CHI[20] mengukur rasio variasi antar-klaster terhadap variasi dalam klaster. Rumusnya tertera pada persamaan (8):

$$CHI = \frac{\sum_{i=1}^{K} n_i \|c_i - c\|^2}{\sum_{i=1}^{K} \sum_{x \in C_i} \|x - c_i\|^2} \times \frac{N - K}{K - 1} \tag{8}$$

   CHI efektif dalam menilai keseimbangan antara kepadatan klaster dan pemisahannya, menjadikannya metrik yang berguna dalam membandingkan performa berbagai algoritma klasterisasi.

3. HASIL DAN PEMBAHASAN

Pada bagian ini berisikan tentang hasil yang diperoleh dari proses klasterisasi terhadap dataset yang telah diproses. Hasil tesebut meliputi distribusi klaster dalam dataset, kinerja dari masing-masing algoritma klasterisasi.

3.1. Evaluasi Kinerja untuk Keempat Model *Clustering*

Dalam menilai kinerja dari keempat algoritma klasterisasi, digunakan tiga metrik evaluasi yang terdiri dari Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI) dalam membantu untuk mencari klaster optimalnya. Ternyata setelah dievaluasi keempat model *clustering* tersebut, semuanya menghasilkan klaster optimalnya 2. Hasil evaluasi akan direpresentasikan ke dalam Tabel 1.

Tabel 1. Hasil Evaluasi Kineeja Keempat Model *Clustering*

Model *Clustering*	Silhouette Score	DBI	CHI
K-Means	0.5175	0.7241	1414.3874
Fuzzy C-Means	0.5125	0.7266	1413.2348
Gaussian Mixture Model	0.4433	0.8724	954.5642
Spectral Clustering	0.4446	0.7707	1123.0370

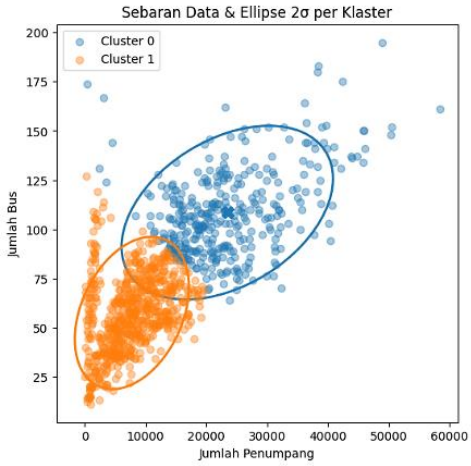
Berdasarkan tabel 1 menunjukkan bahwa metode K-Means adalah model paling unggul dari keempat model *clustering* lainnya. Hal itu dibuktikan oleh skor kinerja K-Means untuk setiap metrik evaluasi yang diujikan terbukti paling baik diantara model *clustering* lainnya. Dari tahap pencarian klaster optimal akan menghasilkan dataset baru yang berisikan pembagian klaster sesuai jumlah klaster optimal. Jumlah pembagian klaster optimal dari setiap model akan dijabarkan ke dalam bentuk Tabel 2.

Tabel 2. Jumlah Klaster setiap Algoritma Klasterisasi

Klaster	K-Means	Fuzzy C-Means	GMM	Spectral Clustering
0	292	302	366	543
1	716	706	642	465

3.2. Analisis Hasil dari Klaster Optimal

Setelah didapatkan klaster optimalnya adalah 2 dan dipilihnya satu model *clustering* terbaik yaitu K-Means, maka dilakukan analisis lebih mendalam terkait hasil klaster berupa interpretasi klaster dan menganalisis hubungan klaster terhadap jumlah bus dan jumlah penumpang berdasarkan tahun yaitu dari tahun 2017-2019 dimana hal itu dilakukan untuk menjawab fokus dari penelitian ini. Dimana interpretasi klaster akan direpresntasikan ke dalam Gambar 3.



Gambar 3. Sebaran Data K-Means

Pada gambar 3 menunjukkan data dapat dibagi menjadi dua kelompok utama dengan karakteristik berbeda. Dimana klaster 1 berwarna biru sedangkan klaster 2 berwarna oranye. Klaster 1 memiliki rentang nilai jumlah penumpang sekitar 10.000 – 55.000 dan jumlah bus sekitar 60 – 200. Dimana klaster 1 menunjukkan jumlah penumpang dan jumlah bus yang tinggi. Sementara klaster 2 mempunyai rentang nilai jumlah penumpang berada di 0 sampai sekitar 15.000 dan jumlah bus berada di 10 sampai sekitar 100. Hal ini membuktikan bahwa klaster 2 memiliki jumlah penumpang relatif rendah dan jumlah bus yang relatif sedikit.

Selanjutnya melakukan analisis hubungan klaster terhadap jumlah bus serta jumlah penumpang berdasarkan periode tahun. Dimana analisis ini akan direpresentasikan ke dalam Tabel 3.

Tabel 3. Total Klaster dari tahun 2017-2019

Tahun	Klaster	Total Penumpang	Total bus	Total dalam Klaster
2017	1	1970699.00	9426	88
2017	2	1810906.00	14382	248
2018	1	1974080.00	10325	92
2018	2	1675352.18	13473	244
2019	1	4371062.00	18724	175
2019	2	1567182.00	9578	161

Berdasarkan tabel 3, menunjukkan adanya pola hubungan jumlah bus dan jumlah penumpang dimana klaster 1selalu memiliki rasio penumpang dan bus yang lebih tinggi dibanding klaster 2. Dimana tren pada tahun 2019 memperlihatkan pergeseran kekuatan ke Klaster 1, mungkin karena optimalisasi atau pertumbuhan wilayah tertentu.

4. KESIMPULAN

Berdasarkan hasil analisis visualisasi sebaran data dan tabel-tabel evaluasi, dapat disimpulkan bahwa pemodelan klaster terhadap data jumlah penumpang dan jumlah bus selama tahun 2017–2019 menghasilkan dua kelompok utama dengan karakteristik yang sangat berbeda: Klaster 1 merepresentasikan wilayah atau waktu dengan intensitas penggunaan layanan tinggi (jumlah penumpang dan rasio penumpang/bus yang tinggi), sementara Klaster 2 menggambarkan layanan dengan permintaan lebih rendah dan distribusi data yang

lebih menyebar. Perbedaan ini muncul sebagai hasil dari distribusi alami data yang menunjukkan adanya segmentasi dalam kebutuhan transportasi sekolah, kemungkinan dipengaruhi oleh ukuran sekolah, waktu operasional, atau lokasi geografis. Hasil ini sejalan dengan teori segmentasi dalam analisis data spasial dan transportasi, di mana kelompok padat dan kelompok sebar dapat muncul secara signifikan dalam sistem layanan publik. Analisis tren tahunan juga menunjukkan adanya pergeseran dominasi klister. Tahun 2019 menunjukkan lonjakan drastis pada Klaster 1 baik dari jumlah data, penumpang, dan bus menandakan pertumbuhan atau konsolidasi layanan. Implikasi dari temuan ini sangat penting dalam perencanaan kebijakan transportasi, karena memungkinkan optimalisasi alokasi armada berdasarkan profil klaster. Kelebihannya adalah model ini mampu secara efektif menangkap pola segmentasi dengan visualisasi dan evaluasi kuantitatif, meskipun keterbatasan tetap ada seperti belum mempertimbangkan variabel temporal lebih dalam atau faktor lokasi spasial eksplisit. Untuk pengembangan ke depan, disarankan integrasi dengan data spasial, demografis, serta pemodelan temporal agar segmentasi menjadi lebih dinamis dan akurat. Dalam evaluasi performa empat model *clustering* menggunakan tiga metrik utama Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI). Model K-Means menunjukkan hasil terbaik dengan nilai Silhouette Score tertinggi (0.5175), nilai DBI terendah kedua (0.7241), dan nilai CHI tertinggi (1414.3874), yang mengindikasikan pemisahan klaster yang baik, kompak, dan terstruktur secara jelas. Model Fuzzy C-Means memiliki performa mendekati K-Means namun sedikit lebih buruk dalam ketiga metrik. Sementara itu, GMM dan Spectral Clustering menunjukkan performa lebih rendah, terutama GMM dengan skor CHI dan Silhouette paling rendah. Oleh karena itu, pemilihan model K-Means dalam analisis ini didasarkan atas keunggulannya dalam memisahkan klaster secara jelas dan stabil, kesederhanaannya dalam interpretasi, serta efisiensinya untuk data dengan dimensi terbatas seperti jumlah penumpang dan jumlah bus menjadikannya paling sesuai untuk tujuan segmentasi dan visualisasi dalam konteks pelayanan transportasi sekolah.

REFERENSI

[1] H. N. V. Do, A. Fujiwara, T. A. H. Nguyen, and C. Do, "Investigating influential factors on pick-up/drop-off location choices for school bus services from parents' perspective: A case study in Hanoi, Vietnam," *Asian Transport Studies*, vol. 11, Jan. 2025, doi: 10.1016/j.eastsj.2025.100157.

[2] Y. Hakiminejad, E. Pantesco, and A. Tavakoli, "Public transit of the future: Enhancing well-being through designing human-centered public transportation spaces," *Transp Res Interdiscip Perspect*, vol. 30, Mar. 2025, doi: 10.1016/j.trip.2025.101365.

[3] F. Farzadnia and J. Lysgaard, "Solving the service-oriented single-route school bus routing problem: Exact and heuristic solutions," *EURO Journal on Transportation and Logistics*, vol. 10, Jan. 2021, doi: 10.1016/j.ejtl.2021.100054.

[4] T. Gu, W. Xu, H. Liang, Q. He, and N. Zheng, "School bus transport service strategies' policy-making mechanism – An evolutionary game approach," *Transp Res Part A Policy Pract*, vol. 182, Apr. 2024, doi: 10.1016/j.tra.2024.104014.

[5] R. Kaewkluengklom, F. Kurauchi, and T. Iwamoto, "Application of passenger flow estimation based on network centrality indicators to reorganise the Shizuoka bus network," in *Transportation Research Procedia*, Elsevier B.V., 2025, pp. 1185–1204. doi: 10.1016/j.trpro.2024.12.120.

[6] S. Tao, F. Rowe, and H. Shan, "Nonlinearities and threshold points in the effect of contextual features on the spatial and temporal variability of bus use in Beijing using explainable machine learning: Predictable or uncertain trips?," *J Transp Geogr*, vol. 123, Feb. 2025, doi: 10.1016/j.jtrangeo.2025.104126.

[7] S. S. Turay, C. A. Adams, and A. Ababio-Donkor, "Application of Multi-Class Machine Learning Algorithms To Predicting Commuter Preference & System Benefits of an Integrated Public Transport: the case of a proposed dedicated bus lane system," *Transportation Engineering*, p. 100319, Mar. 2025, doi: 10.1016/j.treng.2025.100319.

[8] M. Rahal, B. S. Ahmed, G. Szabados, T. Fornstedt, and J. Samuelsson, "Enhancing machine learning performance through intelligent data quality assessment: An unsupervised data-centric framework," *Heliyon*, vol. 11, no. 4, Feb. 2025, doi: 10.1016/j.heliyon.2025.e42777.

[9] Sriwahyuni and R. Passarella, "Analisis Model Prediksi Untuk Layanan Bus Sekolah Jakarta Menggunakan Pendekatan Machine Learning," *Indonesian Journal of Computer Science*, 2024.

[10] P. Shobiroh Utami, "Analisis Model Clustering Kecelakaan Lalu Lintas Di Kota Palembang Menggunakan Pendekatan Machine Learning," 2024.

[11] S. Saeed, H. Haron, N. Z. Jhanjhi, M. Naqvi, H. A. Alhumyani, and M. Masud, "Improve correlation matrix of Discrete Fourier Transformation technique for finding the missing values of MRI images," *Mathematical Biosciences and Engineering*, vol. 19, no. 9, pp. 9039–9059, 2022, doi: 10.3934/mbe.2022420.

[12] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[13] D. Makowski, M. Ben-Shachar, I. Patil, and D. Lüdecke, "Methods and Algorithms for Correlation Analysis in R," *J Open Source Softw*, vol. 5, no. 51, p. 2306, Jul. 2020, doi: 10.21105/joss.02306.

[14] J. Homepage, D. Syaputri, P. Herwina Noprita, and S. Romelah, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Implemnetation of K-Means Algorithm for Economic Distribution Clustering Base on Demographics of Population Implementasi Algoritma K-Means untuk Pengelompokan Distribusi Sosial Ekonomi Masyarakat Berdasarkan Demografi Kependudukan," vol. 1, pp. 1–6, 2021.

[15] I. Purnama Sari and I. Hanif Batubara, "Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering for Grouping Students' Abilities in Online Learning Process," *Journal of Computer Science, Information Technology and Telecommunication Engineering (JCoSITTE)*, vol. 2, no. 1, pp. 139–144, 2021, doi: 10.30596/jcositte.v2i1.6504.

[16] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 158–167. doi: 10.1016/j.procs.2020.04.017.

[17] Z. Kang *et al.*, "Multi-graph fusion for multi-view spectral clustering ☆," vol. 189, p. 105102, 2020, doi: 10.1016/j.knosys.

[18] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International journal of online and biomedical engineering*, vol. 19, no. 4, pp. 174–182, 2023, doi: 10.3991/ijoe.v19i04.37059.

[19] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jan. 2020. doi: 10.1088/1757-899X/725/1/012128.

[20] S. P. Lima and M. D. Cruz, "A genetic algorithm using Calinski-Harabasz index for automatic clustering problem," *Revista Brasileira de Computação Aplicada*, vol. 12, no. 3, pp. 97–106, Sep. 2020, doi: 10.5335/rbca.v12i3.11117.