Informatics for Educators And Professionals : Journal of Informatics
Vol. 7, No. 2, Juni 2023, 138 - 152
E-ISSN: 2548-3412
Terakreditasi **SINTA Peringkat 5**
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 230/E/KPT/2022
masa berlaku mulai Volume 5 Nomor 1 Tahun 2020 sampai Volume 9 Nomor 2 Tahun 2024

Pengelompokkan Dataset Bus Menggunakan Algoritma K-Means

Anjar Permadi [1,*], Yudhistira Arie Wijaya [2]

[1,2] Manajemen Informatika; STMIK IKMI Cirebon; Jl. Perjuangan No.10B, Karyamulya, Kec.Kesambi, Kota Cirebon, Jawa Barat 45131, 0231-490480; e-mail: anjarpermadi.kkm@gmail.com, yudhistira010471@gmail.com

* Korespondensi: e-mail: anjarpermadi.kkm@gmail.com,

Diterima: 11 Mei 2023; Review: 24 Mei 2023; Disetujui: 07 Juni 2023

Cara sitasi: Permadi A, Wijaya YA. 2023. Pengelompokkan Dataset Bus Menggunakan Algoritma K-Means. Informatics for Educators and Professionals : Journal of Informatics. Vol.7 (2): 138 - 152.

Abstrak: *Data mining* adalah proses eksplorasi data yang bertujuan untuk menemukan informasi baru atau menarik dengan mengidentifikasi pola, hubungan, dan tren yang tersembunyi dalam dataset. Proses pencarian informasi tersebut dapat dilakukan dengan menggunakan metode pengelompokan atau bisa juga disebut clustering dalam data mining. Metode *clustering* digunakan untuk membagi dataset menjadi kelompok-kelompok atau subset berdasarkan kesamaan karakteristik data di setiap kelompok. Metode *Clustering* yang digunakan dalam penelitian ini adalah *K-Means* yang termasuk ke dalam golongan algoritma *Partition Clustering*. Metode ini juga sudah banyak digunakan dalam penyelesaian masalah terkait klasterisasi pejualan, kebakaran hutan, pertanian, transportasi, dan sebagainya. Pada penelitian digunakan algoritma *K-means* untuk mengelompokkan dataset Bus BB berdasarkan data yang diambil selama tahun 2022. Untuk mengubah dataset mentah menjadi informasi yang bermanfaat, sering digunakan proses *Knowledge Discovery in Database (KDD)*. Tahap awal dari proses ini adalah pembersihan data, diikuti oleh seleksi data, transformasi data, dan data mining. Dalam konteks ini, *software Rapidminer* sering digunakan sebagai alat untuk melaksanakan tahapan tersebut. Hasil pemodelan dievaluasi menggunakan *Davies Bouldin Index (DBI)* untuk mengukur kualitas *clustering*. Semakin rendah nilai *DBI*, semakin baik kualitas pemodelan. Penelitian menunjukkan bahwa algoritma *K-Means* efektif dalam mengelompokkan dataset bus BB. Yang nantinya bisa dimanfaatkan oleh perusahaan sebagai gambaran, penelitian juga ini bisa digunakan sebagai masukan bagi perusahaan/penyedia jasa. Penelitian ini juga memberikan kontribusi pada pengembangan pengetahuan dan pemahaman di bidang *data mining* dan pengelompokan data. Dengan menggunakan langkah-langkah *KDD* dan algoritma *K-Means*, penelitian ini dapat menjadi referensi bagi penelitian-penelitian selanjutnya dan membantu memperkaya pemahaman tentang aplikasi *data mining* dalam konteks industri transportasi.

Kata kunci: *Data Mining, Clustering, K-Means, DBI*

Abstract: *Data mining is a data exploration process that aims to discover new or interesting information by identifying patterns, relationships and trends hidden in datasets. The process of searching for this information can be done using the grouping method or it can also be called clustering in data mining. The clustering method is used to divide the dataset into groups or subsets based on the similarity of data characteristics in each group. The clustering method used in this research is K-Means which belongs to the Partition Clustering algorithm group. This method has also been widely used in solving problems related to sales clustering, forest fires, agriculture, transportation, and so on. In this study the K-means algorithm was used to classify*

Copyright@2023. LPPM UNIVERSITAS BINA INSANI

the Bus BB dataset based on data collected during 2022. To turn raw datasets into useful information, the Knowledge Discovery in Database (KDD) process is often used. The initial stage of this process is data cleaning, followed by data selection, data transformation, and data mining. In this context, Rapidminer software is often used as a tool to carry out these stages. Modeling results were evaluated using the Davies Bouldin Index (DBI) to measure clustering quality. The lower the DBI value, the better the modeling quality. Research shows that the K-Means algorithm is effective in classifying BB bus datasets. Which later can be used by companies as an illustration, this research can also be used as input for companies/service providers. This research also contributes to the development of knowledge and understanding in the field of data mining and data grouping. By using KDD steps and the K-Means algorithm, this research can be a reference for further research and help enrich understanding of data mining applications in the context of the transportation industry.

***Keywords**: Data Mining, Clustering, K-Means, DBI*

1. Pendahuluan

Teknik memperoleh atau menambang pengetahuan dari sejumlah besar data dikenal sebagai *data mining* [1]. *Data mining* adalah proses penggunaan metode statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengidentifikasi informasi yang bermanfaat dan menemukan pengetahuan yang tersembunyi dari berbagai database. [2]. Definisi lainnya adalah Pembelajaran berbasis induksi (*induction-based learning*) yaitu proses pembentukan definisi-definisi konsep umum yang dilakukan dengan mengamati contoh-contoh spesifik dari konsep-konsep yang ingin dipelajari. Dalam proses ini, data yang ada digunakan untuk mengidentifikasi pola dan aturan umum yang dapat digunakan untuk mengklasifikasikan data baru [3]. Data mining memainkan peran kunci dalam proses *Knowledge Discovery in Database (KDD)*, di mana algoritma-algoritma tertentu digunakan untuk menggali data, membangun model, dan menemukan pola yang tidak diketahui sebelumnya. Model yang dihasilkan dari proses ini digunakan untuk memahami fenomena data, melakukan analisis, dan bahkan melakukan prediksi terhadap data baru yang ditemui [4]. Terdapat sebuah algoritma dalam data mining yang disebut *K-Means*. *K-Means* berperan sebagai algoritma yang membagi data ke dalam kelompok berdasarkan tingkat kesamaannya [5]. Algoritma ini memiliki implementasi yang sederhana, kinerjanya relatif cepat, mudah untuk disesuaikan, dan banyak digunakan dalam praktik. Metode *K-Means* membagi data menjadi beberapa kelompok berdasarkan kesamaan karakteristiknya. Data dengan karakteristik yang serupa dikelompokkan bersama, sementara data dengan karakteristik yang berbeda ditempatkan dalam kelompok yang berbeda pula [6]. *Clustering* adalah proses dimana titik-titik data dikelompokkan menjadi dua kelompok atau lebih, dengan tujuan agar titik-titik data dalam kelompok yang sama memiliki kesamaan yang lebih tinggi dibandingkan dengan kelompok yang berbeda, berdasarkan informasi yang tersedia dari titik-titik data tersebut [7].

Penelitian yang dilakukan oleh Erni Dianawati, Putri Previa Yanti, dan Yulia Suryandari pada tahun 2019 berjudul "Klustering Jumlah Penumpang pada Halte Bus Rapid Transit Kota Tangerang". Transit Bus Rapid Transit (BRT) Trans Tangerang menghadapi beberapa kendala, seperti meningkatkan jumlah penumpang untuk mengurangi kemacetan dan menentukan lokasi halte bus yang strategis untuk menarik penumpang. Untuk mengatasi tantangan ini, penting untuk mengumpulkan data dan informasi terkait BRT Trans Tangerang guna mengidentifikasi pola distribusi penumpang. Salah satu pendekatan yang dapat digunakan adalah metode data mining, dengan menggunakan teknik clustering untuk mengelompokkan data ke dalam kelompok yang memiliki kesamaan. Hasil dari penelitian ini dapat memberikan wawasan tentang jumlah lintasan yang dihasilkan setiap bulan, sehingga dengan memahami karakteristik clustering, pihak terkait dapat mengantisipasi kepadatan penumpang pada hari-hari tertentu dan mengatur jadwal dengan lebih efektif [8].

Dalam penelitian ini, pendekatan yang digunakan adalah *Knowledge Discovery In Database (KDD).* Proses KDD melibatkan serangkaian tahapan yang dimulai dari seleksi data hingga interpretasi/evaluasi [9]. Proses *KDD* juga melibatkan interpretasi hasil yang diperoleh dari sekumpulan data dengan mengintegrasikannya dengan pengetahuan dari disiplin ilmu lainnya [10]. Sedangkan untuk pengelompokan datanya menggunakan algortitma *K-Means*. *K-Means* adalah metode pengelompokan data non-hierarkis yang bertujuan untuk mempartisi data menjadi dua kelompok atau lebih [11]. Metode *K-Means* akan membagi data ke dalam

kelompok-kelompok, di mana data dengan karakteristik yang serupa akan ditempatkan dalam kelompok yang sama, sedangkan data dengan karakteristik yang berbeda akan ditempatkan dalam kelompok yang berbeda pula [12]. Pemilihan metode K-Means didasarkan pada beberapa faktor yang membuatnya menjadi pilihan yang baik. Metode ini lebih sederhana dan mudah diimplementasikan, tidak memerlukan waktu yang lama untuk dieksekusi, mudah untuk disesuaikan dengan kebutuhan, dan merupakan salah satu metode yang paling umum digunakan dalam proses data mining [5]. Tujuan dari penelitian ini adalah melakukan proses penambangan data untuk menghasilkan kelompok-kelompok (*cluster*) sesuai dengan pengetahuan atau pola yang terdapat dalam data tersebut.

Penelitian ini dilakukan dengan tujuan untuk menganalisis klasterisasi dataset bus biskita dari setiap koridor. Hal ini diharapkan dapat memberikan manfaat bagi perusahaan atau penyedia jasa angkutan massal bus biskita. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi pengaruh parameter *Euclidean Distance* yang digunakan dalam metode *k-means clustering.*

2. Metode Penelitian

Dalam penelitian ini, penulis menggunakan metode penelitian kuantitatif. Metode penelitian kuantitatif merupakan pendekatan analitis dan matematis yang digunakan untuk mengukur dan menganalisis data kuantitatif, yaitu data yang dinyatakan dalam bentuk angka. Metode penelitian ini sering melibatkan pengumpulan data melalui survei, eksperimen, observasi, atau penggunaan data sekunder, kemudian data tersebut dianalisis menggunakan metode statistik.

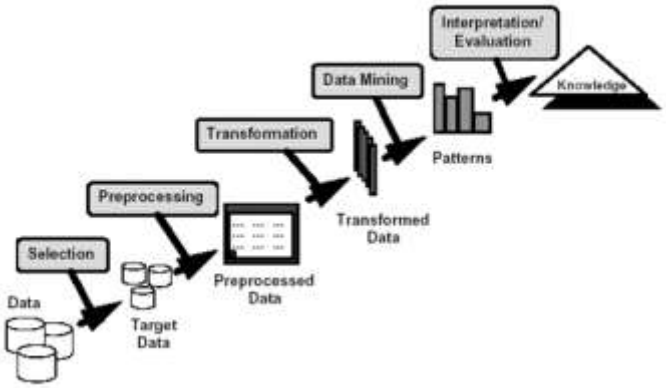
Dalam penelitian ini, digunakan algoritma *K-Means* yang merupakan algoritma pengelompokan yang umumnya berdasarkan perhitungan jarak antara sampel-sampel data. Pada algoritma *K-Means*, semakin kecil jarak antara dua sampel, maka semakin tinggi tingkat kesamaannya [13]. Algoritma ini memiliki efisiensi tinggi dan sering digunakan dalam pengolahan data yang tidak memiliki label untuk melakukan tugas pengelompokan [14].

Clustering atau pengelompokan merupakan proses pembentukan kelompok data dari kumpulan data yang tidak memiliki kelompok atau kelas yang diketahui sebelumnya. Tujuannya adalah menentukan kelompok mana setiap data akan termasuk. Metode pengelompokan memisahkan data ke dalam kelompok berdasarkan kesamaan karakteristiknya. Tujuan utamanya adalah mengurangi variasi dalam kelompok yang sama dan meningkatkan variasi antara kelompok-kelompok [15]. Proses ini bekerja dengan mengelompokkan kumpulan data ke dalam kelompok-kelompok atau *cluster-cluster*. Dalam kelompok yang sama, objek-objek memiliki tingkat kesamaan yang tinggi dibandingkan dengan objek-objek di kelompok lain. Namun, objek-objek tersebut memiliki tingkat kesamaan yang rendah dengan objek-objek di *cluster* lain [16]. *Clustering* merupakan salah satu metode dalam Data Mining yang termasuk dalam kategori tanpa arahan (*unsupervised*). Artinya, metode ini tidak memerlukan latihan atau guru serta tidak memerlukan target *output*. Metode clustering ini berfokus pada pengelompokan data berdasarkan karakteristik dan kesamaan antar data tersebut, tanpa memperhatikan hasil yang diharapkan [17].

Davies-Bouldin Index (DBI) adalah salah satu metode yang digunakan untuk mengevaluasi kevalidan atau jumlah *cluster* yang paling optimal dalam suatu metode pengelompokan. Metode ini menggunakan konsep kohesi, yang mengacu pada tingkat kedekatan data dengan titik pusat *cluster* dari *cluster* yang mereka masuki. *DBI* mengukur sejauh mana *cluster-cluster* tersebut saling berbeda dan sejauh mana setiap cluster memiliki kohesi yang tinggi [18]. Metode *DBI* dipilih karena dapat diimplementasikan dalam semua data baik ukuran besar dan kecil sehingga sangat cocok dan dapat diterima untuk perhitungan akurasi *clustering* [19].

Sumber data dalam penelitian ini bersumber dari Website Executive Dashboard KP. Data memiliki Data memiliki 6 field atau atribut yaitu : Nomor Kendaraan, Kode Kendaraan, Jenis Bus, Total Penumpang, Total Ritase, dan KM Tempuh Koridor. Data yang digunakan yaitu data selama tahun 2022, dengan total jumlah data sebanyak 3724 *records.*

Metode yang digunakan yaitu *Knowledge Discovery in Database (KDD)*. *KDD* yaitu proses yang melibatkan pengumpulan data, penggunaan data historis, dan eksplorasi terhadap set data yang besar untuk menemukan pola, keteraturan, atau hubungan yang terdapat di dalamnya [20]. Proses *Knowledge Discovery in Database* diawali dengan menetapkan tujuan dan diakhiri dengan evaluasi [10].



Sumber: Hasil Penelitian (2023)

Gambar 1. Tahapan proses *KDD*

Data Selection

Pada tahap *Data Selection* ini, dilakukan proses seleksi atribut mana yang akan digunakan dalam proses data mining. Dari 6 atribut yaitu Nomor Kendaraan, Kode Kendaraan, Jenis Bus, Total Penumpang, Total Ritase, dan KM Tempuh Koridor. Dataset tersebut akan diseleksi menjadi 4 atribut yaitu KM Tempuh Koridor, Kode Kendaraan, Total Penumpang, dan Total Ritase. Kemudian ditambahkan operator *generate id*, dikarenakan data mentahnya belum memiliki *id*.

Preprocessing

Preprocessing adalah tahap awal dalam analisis data di mana data yang diperoleh dipersiapkan dengan melakukan pembersihan terhadap missing value atau noise. Pada tahap ini, data yang tidak relevan atau tidak konsisten akan diidentifikasi dan dihilangkan untuk memastikan kualitas data yang digunakan dalam analisis.

Data Transformation

Data Transformation atau tranformasi data merujuk pada tahap di mana tipe data dari satu bentuk diubah menjadi bentuk lain agar sesuai dengan persyaratan atau kebutuhan analisis yang akan dilakukan. Pada penelitian ini, tipe data diubah menjadi tipe data numerik.

Data Mining

Tahapan data mining dilakukan dengan menerapkan algoritma atau metode untuk mencari pengetahuan atau informasi yang berharga dari dataset yang ada. Pada penelitian ini, algoritma yang digunakan adalah algoritma *k-means clustering.*

Interpretation/Evaluation

Pada tahap interpretasi atau evaluasi dilakukan dengan menganalisis hasil eksperimen yang telah dilakukan untuk memahami dan mengevaluasi kesesuaian serta relevansi penemuan yang telah ditemukan.

3. Hasil dan Pembahasan

Penerapan KDD dalam penelitian ini yaitu sebagai berikut :

Data Selection :

Read Excel, operator ini digunakan untuk mengimpor atau memasukkan data dari file Excel yang terletak di komputer pengguna ke dalam proses yang sedang berjalan di RapidMiner. Karena dataset yang digunakan adalah dataset bus BB yang terdiri dari 4 file excel berbeda. Maka dari itu dibutuhkan juga 4 operator *read excel* di dalam proses Rapidminer. Untuk parameter yang digunakan adalah parameter *default*.

Sumber: Hasil Penelitian (2023)

Gambar 2. *Read Excel*

Gambar diatas adalah operator *Read Excel* yang tersedia pada aplikasi *Rapidminer*. Parameter pada operator *Read Excel* menggunakan nilai default yang telah ditentukan sebelumnya. Setelah operator *Read Excel* dieksekusi, didapatkan informasi sebagai berikut.

Tabel 1. Statistik dataset file pertama

No	Uraian	Isi
1	*Record*	538
2	*Special Attribute*	0
3	*Regular Attribute*	6
4	*Attributes :*	
	Nomor Kendaraan	*Polynominal, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Jenis Bus	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Tabel 2. Statistik dataset file kedua

No	Uraian	Isi
1	*Record*	1102
2	*Special Attribute*	0
3	*Regular Attribute*	6
4	*Attributes :*	
	Nomor Kendaraan	*Polynominal, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Jenis Bus	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Tabel 3. Statistik dataset file ketiga

No	Uraian	Isi
1	*Record*	1362
2	*Special Attribute*	0
3	*Regular Attribute*	6
4	*Attributes :*	
	Nomor Kendaraan	*Polynominal, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Jenis Bus	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

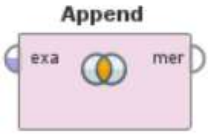
Sumber: Hasil Penelitian (2023)

Tabel 4. Statistik dataset file keempat

No	Uraian	Isi
1	*Record*	722
2	*Special Attribute*	0
3	*Regular Attribute*	6
4	*Attributes :*	
	Nomor Kendaraan	*Polynominal, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Jenis Bus	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Append, operator ini berfungsi menggabungkan beberapa dataset menjadi satu. Dataset yang digunakan terdiri dari 4 file excel yang berisi data koridor bus yang berbeda kemudian dijadikan satu, dengan catatan atribut yang digunakan harus sama.



Sumber: Hasil Penelitian (2023)

Gambar 3. *Append*

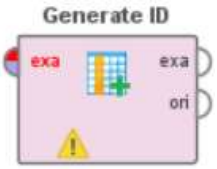
Gambar diatas adalah operator *Append* yang tersedia pada aplikasi *Rapidminer*. Operator *Append* digunakan dengan parameter default. Setelah operator *Append* dieksekusi, informasi yang diperoleh adalah sebagai berikut.

Tabel 5. Statistik dataset setelah menggunakan *append*

No	Uraian	Isi
1	*Record*	3724
2	*Special Attribute*	0
3	*Regular Attribute*	6
4	*Attributes :*	
	Nomor Kendaraan	*Polynominal, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Jenis Bus	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Generate ID, operator ini menambahkan atribut baru dengan peran id di input dataset. Karena pada dataset bus BB tidak ada atribut yang dapat dijadikan *id* maka ditambahkan operator *Generate ID.* Untuk parameter yang digunakan adalah parameter default.



Sumber: Hasil Penelitian (2023)

Gambar 4. *Generate ID*

Gambar diatas adalah operator *Generate ID* yang tersedia pada aplikasi *Rapidminer*. Operator *Generate ID* menggunakan parameter default. Setelah operator *Generate ID* dieksekusi, informasi yang diperoleh adalah sebagai berikut.

Tabel 6. Statik dataset

No	Uraian	Isi
1	*Record*	3724
2	*Special Attribute*	1
3	*Regular Attribute*	6
4	*Attributes :*	
	ID	*Integer, missing 0*
	Nomor Kendaraan	*Polynominal, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Jenis Bus	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*

Sumber: Hasil Penelitian (2023)

Select Attributes, operator ini digunakan untuk memilih atribut atau kolom data yang akan digunakan dalam proses pemrosesan data. Pada dataset bus BB, atribut yang awalnya berjumlah 7 telah difilter menjadi 5 atribut yang akan digunakan. Atribut yang digunakan yaitu : id, KM Tempuh Koridor, Kode Kendaraan, Total Penumpang, dan Total Ritase.

Sumber: Hasil Penelitian (2023)

Gambar 5. *Select Attributes*

Di atas terdapat gambar dari operator *Select Attributes* yang tersedia dalam aplikasi *Rapidminer*. Parameter pada operator *Select Attributes* yang digunakan tampak pada tabel dibawah ini.

Tabel 7. Parameter dan atribut yang dipilih pada operator *Select Attributes*

No	Parameter	Isi
1	*Attribute Filter Type*	*Subset*
2	*Selected Attributes*	ID, Kode Kendaraan, Total Penumpang, Total Ritase, dan KM Tempuh Koridor

Sumber: Hasil Penelitian (2023)

Dari hasil pembacaan operator *Select Attributes* didapat informasi sebagai berikut.

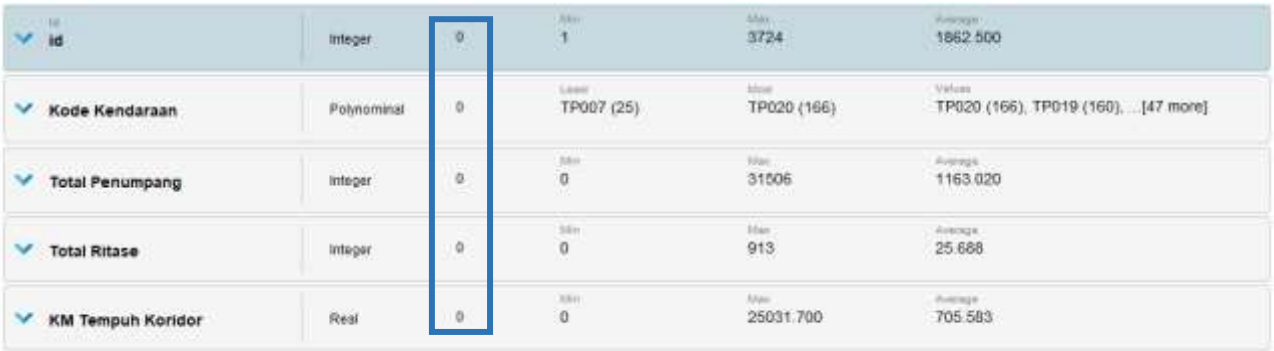
Tabel 8. Statik dataset

No	Uraian	Isi
1	*Record*	3724
2	*Special Attributes*	1
3	*Regular Attributes*	4
4	*Attributes :*	
	ID	*Integer, missing 0*
	Kode Kendaraan	*Polynominal, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Preprocessing

Karena tidak ada data yang hilang atau tidak memiliki nilai pada dataset bus BB, maka tidak perlu dilakukan tahap *preprocessing*.



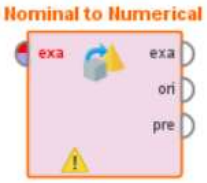
Sumber: Hasil Penelitian (2023)

Gambar 6. Statik dataset

Setelah melihat hasil statistik pada gambar di atas, tidak ada missing value yang ditemukan. Oleh karena itu, proses dapat lanjut ke tahap berikutnya.

Data Transformation

Nominal to Numerical, digunakan untuk mengubah tipe atribut non-numerik menjadi tipe numerik. Pada dataset bus BB, atribut "Kode Kendaraan" akan diubah menjadi tipe numerik.



Sumber: Hasil Penelitian (2023)

Gambar 7. *Nominal to Numerical*

Di atas terdapat gambar dari operator *Nominal to Numerical* yang tersedia dalam aplikasi *Rapidminer*. Parameter pada operator *Nominal to Numerical* yang digunakan tampak pada tabel dibawah ini.

Tabel 9. Parameter pada operator *Nominal to Numerical*

No	Parameter	Isi
1	*Attribute Filter Type*	*Subset*
2	*Selected Attributes*	Kode Kendaraan
3	*Coding Type*	Unique Integer

Sumber: Hasil Penelitian (2023)

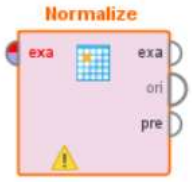
Dari hasil pembacaan operator *Nominal to Numerical* didapat informasi sebagai berikut.

Tabel 10. Statik dataset

No	Uraian	Isi
1	*Record*	3724
2	*Special Attributes*	1
3	*Regular Attributes*	4
4	*Attributes :*	
	ID	*Integer, missing 0*
	Kode Kendaraan	*Numerical, missing 0*
	Total Penumpang	*Integer, missing 0*
	Total Ritase	*Integer, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Normalize*,* digunakan untuk menskalakan nilai agar cocok dalam rentang tertentu. Menyesuaikan rentang nilai sangat penting saat berhadapan dengan atribut unit dan skala yang berbeda. Dalam dataset bus BB atribut yang akan dirubah menggunakan *normalize* yaitu KM Tempuh Koridor, Total Penumpang, dan Total Ritase.



Sumber: Hasil Penelitian (2023)

Gambar 8. Normalize

Gambar diatas adalah operator *Normalize* yang tersedia pada aplikasi *Rapidminer*. Parameter pada operator *Normalize* yang digunakan tampak pada tabel dibawah ini. Operator ini

melakukan perhitungan nilai rata-rata dari atribut dengan tujuan mengurangi jarak antara data-data yang ada.

Tabel 11. Parameter dan atribut yang dipilih pada operator *Normalize*

No	Parameter	Isi
1	*Attribute Filter Type*	*Subset*
2	*Selected Attributes*	KM Tempuh Koridor, Total Penumpang, dan Total Ritase
3	*Method*	*Range Transformation*

Sumber: Hasil Penelitian (2023)

Dari hasil pembacaan operator *Normalize* didapat informasi sebagai berikut.

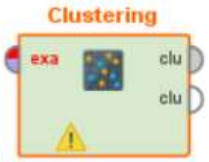
Tabel 12. Statik dataset

No	Uraian	Isi
1	*Record*	3724
2	*Special Attributes*	1
3	*Regular Attributes*	4
4	*Attributes :*	
	ID	*Integer, missing 0*
	Kode Kendaraan	*Numerical, missing 0*
	Total Penumpang	*Real, missing 0*
	Total Ritase	*Real, missing 0*
	KM Tempuh Koridor	*Real, missing 0*

Sumber: Hasil Penelitian (2023)

Data Mining

Dalam proses ini, metode yang diterapkan adalah pengelompokan menggunakan algoritma *k-means clustering*.



Sumber: Hasil Penelitian (2023)

Gambar 9. *K-Means Clustering*

Gambar diatas adalah operator *K-Means Clustering* yang tersedia pada aplikasi *Rapidminer*. Parameter default digunakan pada operator *K-Means Clustering*, dengan nilai k yang diubah secara iteratif mulai dari 2 hingga 20.

Tabel 13. Parameter pada operator K-Means Clustering

No	Parameter	Isi
1	*K*	*2-20*

Sumber: Hasil Penelitian (2023)

Kemudian, ditambahkan operator Performance dengan menggunakan metode *Davies-Bouldin Index (DBI)* untuk mengukur nilai *DBI*. Hal ini bertujuan untuk mendapatkan informasi mengenai nilai *DBI* pada hasil *clustering*.

Sumber: Hasil Penelitian (2023)

Gambar 10. *Cluster Distance Performance*

Gambar di atas adalah tampilan operator *Performance* yang tersedia dalam aplikasi RapidMiner. Operator ini digunakan untuk melihat hasil *Davies-Bouldin Index (DBI)* pada proses

clustering. Parameter pada operator *Performance* yang digunakan tampak pada tabel dibawah ini.

Tabel 14. Parameter pada operator *Performance*

No	Parameter	Isi
1	*Main Criterion*	*Davies Bouldin*

Sumber: Hasil Penelitian (2023)

Dari hasil pembacaan operator *Performance* maka didapat hasil sebagai berikut.

Tabel 15. Hasil operator *Performance (DBI)*

No	K	DBI
1	2	0,607
2	3	0,530
3	4	0,644
4	5	0,692
5	6	0,714
6	7	0,885
7	8	0,931
8	9	0,899
9	10	0,886
10	11	0,871
11	12	0,896
12	13	0,884
13	14	0,916
14	15	0,872
15	16	0,889
16	17	0,866
17	18	0,898
18	19	0,844
19	20	0,863

Sumber: Hasil Penelitian (2023)

Interpretation/Evaluation

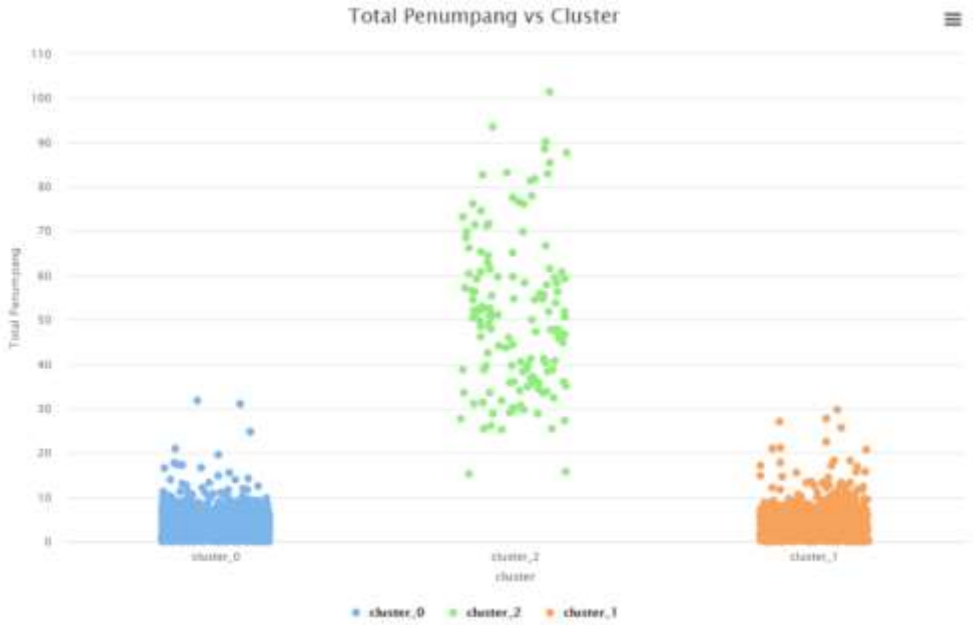
Berdasarkan perbandingan *DBI* dengan metode *K-Means* dari k-2 sampai dengan k-20, ditemukan bahwa cluster terkecil yang mendekati nilai 0 adalah k-3, dengan nilai *DBI* sebesar 0,530. Oleh karena itu, dapat disimpulkan bahwa k-3 dengan nilai DBI 0,530 merupakan hasil cluster terbaik. Pada hasil clustering ini, terdapat 3 cluster dengan jumlah item sebagai berikut: cluster 0 sebanyak 1972 items, cluster 1 sebanyak 1612 items, dan cluster 2 sebanyak 140 items.

Tabel 16. Hasil *Clustering* k-3 yang merupakan nilai *DBI* terbaik

No	Parameter	Isi
1	Cluster 0	1972
2	Cluster 1	1612
3	Cluster 2	140
	Total Number of Items	3724

Sumber: Hasil Penelitian (2023)

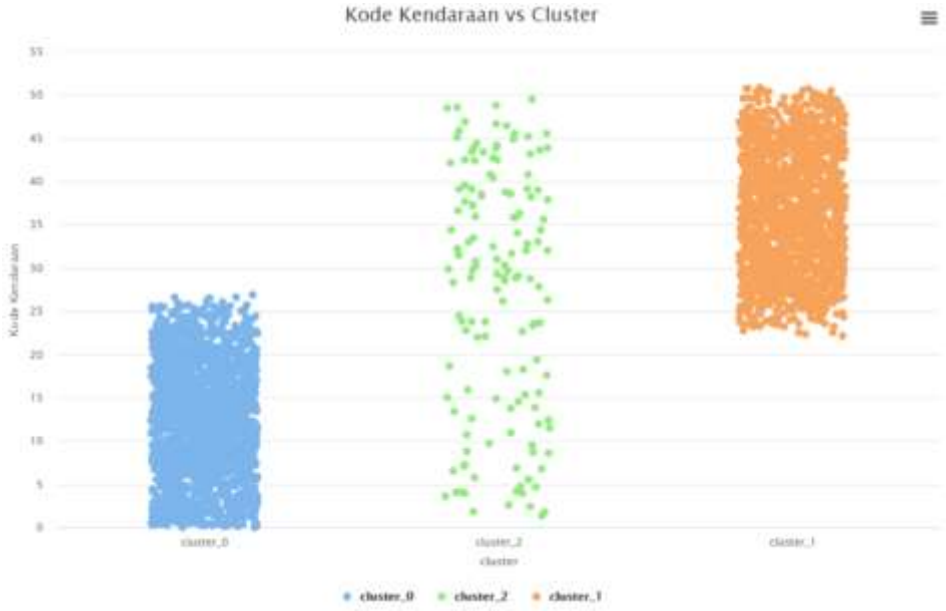
Setelah diketahui k-3 yang merupakan hasil DBI dengan nilai terkecil yaitu pada *numerical measure Euclidean Distance*. Kemudian dilihat pada statistik visualisasi *scatter*. Maka diketahui hasil dari setiap atribut yang ada pada dataset Bus BB sebagai berikut.



Sumber: Hasil Penelitian (2023)

Gambar 11. Total Penumpang *vs Cluster*

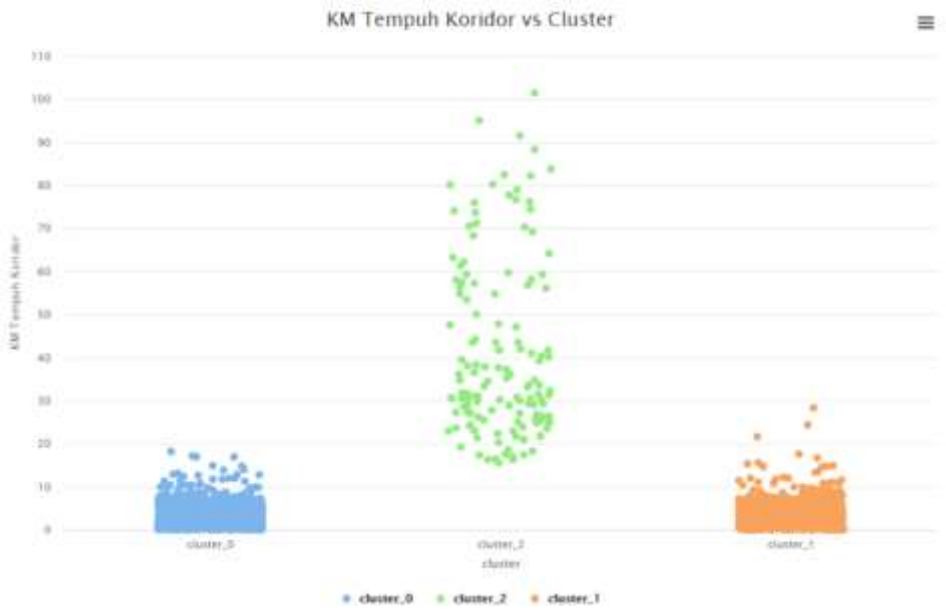
Pada gambar 11 terdapat 3 *cluster*. Dari 3 *cluster* ini ternyata *cluster* 0 dan *cluster* 1 terlihat lebih padat dibandingkan *cluster* 2 yang terlihat terpencar. Pada cluster 0, nilai cluster dimulai dari 0 sampai dengan 28,61. Pada cluster 1, nilai cluster dimulai dari 0 sampai dengan 25,47. Dan pada cluster 2, nilai cluster dimulai dari 18,82 sampai dengan 100. Nilai tertinggi pada cluster 0 dan cluster 1 hanya berbeda tipis, dengan selisih sebesar 3,14.



Sumber: Hasil Penelitian (2023)

Gambar 12. Kode Kendaraan *vs Cluster*

Pada gambar 12 terdapat 3 *cluster*. Dari ketiga *cluster* ini ternyata *cluster* 0 dan *cluster* 1 terlihat lebih banyak dan padat, dibandingkan *cluster* 2 yang terlihat terpencar. Pada cluster 0, nilai cluster nya dimulai dari 0 sampai dengan 24. Pada cluster 1, nilai cluster dimulai dari 25 dan nilai tertinggi nya adalah 48. Dan pada cluster 2, nilai cluster dimulai dari 0 sampai dengan 48.



Sumber: Hasil Penelitian (2023)

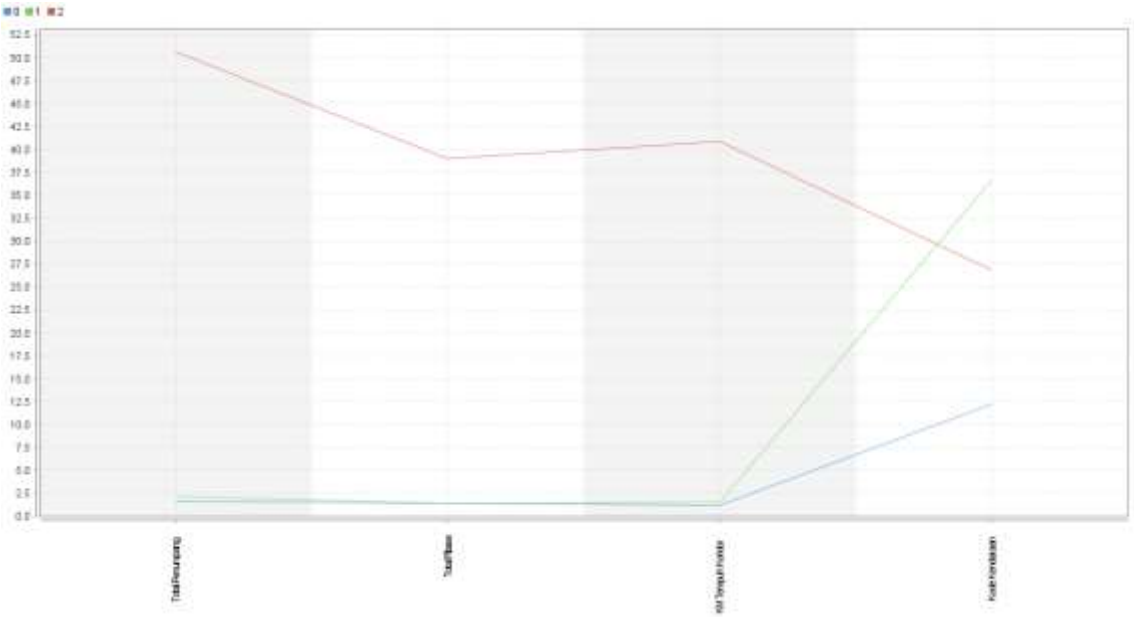
Gambar 13. KM Tempuh Koridor *vs Cluster*

Pada gambar 13 terdapat 3 *cluster*. Dari 3 *cluster* dapat dilihat bahwa *cluster* 0 dan *cluster* 1 terlihat lebih rapat dibandingkan *cluster* 2 yang terlihat terpencar. Pada cluster 0, nilai cluster dimulai dari 0 sampai dengan 16,17. Pada *cluster* 1, nilai cluster nya dimulai dari 0 sampai dengan 24,04. Dan pada *cluster* 2, nilai cluster dimulai dari 18,09 sampai dengan 100. Nilai tertinggi pada cluster 0 dan cluster 1 hanya berbeda tipis, yaitu sebesar 7,87.

Sumber: Hasil Penelitian (2023)

Gambar 14. Total Ritase *vs Cluster*

Pada gambar 14 terdapat 3 *cluster*. Dari 3 *cluster* dapat dilihat bahwa *cluster* 0 dan *cluster* 1 terlihat lebih padat dibandingkan *cluster* 2 yang terlihat terpencar. Pada cluster 0, nilai cluster dimulai dari 0 sampai dengan 21,08. Pada *cluster* 1, nilai cluster dimulai dari 0 sampai dengan 23,65. Dan pada *cluster* 2, nilai cluster dimulai dari 16,31 sampai dengan 100. Nilai tertinggi pada cluster 0 dan cluster 1 hanya berbeda tipis, yaitu sebesar 2,57. Dari gambar-gambar diatas, diketahui bahwa nilai terendah dan tertinggi pada setiap atribut berbeda-beda.



Gambar 15. *Plot* k-3 pada *cluster model*

Berdasarkan gambar di atas, terlihat bahwa cluster 2 memiliki kecenderungan yang sangat berbeda dari cluster 0 dan cluster 1. Cluster 2 menonjol secara dominan, sementara cluster 0 dan cluster 1 memiliki kemiripan yang lebih tinggi di antara keduanya. Terdapat kesamaan dalam bentuk atribut Total Penumpang, Total Ritase, dan KM Tempuh Koridor antara cluster 0 dan cluster 1. Namun, terdapat peningkatan pada atribut Kode Kendaraan di kedua cluster tersebut.

Cluster 0 awalnya memiliki nilai yang berkisar antara 0-2,5 pada atribut Total Penumpang, Total Ritase, dan KM Tempuh Koridor. Namun, terjadi peningkatan yang signifikan pada atribut Kode Kendaraan dengan penambahan nilai sebesar 12,5.

Cluster 1 juga awalnya memiliki nilai yang berkisar antara 0-2,5 pada atribut Total Penumpang, Total Ritase, dan KM Tempuh Koridor. Namun, pada atribut Kode Kendaraan terjadi peningkatan yang lebih besar dengan penambahan nilai sebesar 36.

Sementara itu, Cluster 2 memiliki pola yang berbeda dari cluster 0 dan cluster 1. Pada atribut Total Penumpang, nilai cluster 2 adalah 50, kemudian mengalami penurunan pada atribut Total Ritase dengan nilai 39, dan mengalami peningkatan pada atribut KM Tempuh Koridor dengan nilai 41. Namun, pada atribut Kode Kendaraan terjadi penurunan yang signifikan dengan penurunan nilai sebesar 27.

Dari data tersebut, dapat diamati bahwa setiap atribut, yaitu Total Penumpang, Total Ritase, KM Tempuh Koridor, dan Kode Kendaraan, memiliki karakteristik atau pola yang berbeda-beda. Dengan demikian, dapat disimpulkan bahwa tiap atribut memiliki karakteristik dan pola yang berbeda dalam membentuk cluster-cluster tersebut.

4. Kesimpulan

Dataset bus dapat dikelompokkan dengan menggunakan metode *K-Means clustering* pada aplikasi *Rapidminer*. Dalam pengelompokan ini, parameter *default* digunakan untuk menentukan jumlah *cluster* yang optimal. Berdasarkan hasil pengelompokan dengan *metode K-Means* dan evaluasi menggunakan *Davies-Bouldin Index (DBI)*, ditemukan bahwa *cluster* terbaik adalah k-3 dengan nilai *DBI* sebesar 0,530. Terdapat tiga *cluster* yang terbentuk, yaitu *cluster* 0 dengan 1972 item, *cluster* 1 dengan 1612 item, dan *cluster* 2 dengan 140 item. Selain itu, hasil dari pengelompokan data menggunakan metode K-Means ini memberikan wawasan tentang karakteristik atribut dalam dataset bus. Dan mungkin untuk penelitian selanjutnya, dapat dipertimbangkan penggunaan algoritma lain seperti klasifikasi, algoritma *Random Forest*, *Naïve Bayes*, atau algoritma lainnya. Hal ini bertujuan untuk mendapatkan hasil yang lebih baik dan berbeda, baik dari segi nilai *DBI* maupun pengelompokan data. Hasil penelitian ini diharapkan dapat menjadi referensi bagi peneliti lain yang tertarik untuk mengelompokkan data menggunakan metode K-Means dan menjalankan penelitian dengan pendekatan lainnya.

Referensi

[1] M. Krishnamoorthy and R. Karthikeyan, “Pattern mining algorithms for data streams using itemset,” *Meas. Sensors*, vol. 24, no. June, p. 100421, 2022, doi: 10.1016/j.measen.2022.100421.

[2] D. P. Utomo and M. Mesran, “Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung,” *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[3] Hafizah, Tugiono, and W. R. Maya, “Penerapan Data Mining Dalam Memprediksi Jumlah Penumpang Pada CV . Surya Mandiri Sukses Dengan Menggunakan Metode Regresi Linier,” *J. Teknol. Inf. dan Sist. Komput. TGD*, vol. 2, no. 1, pp. 54–61, 2019.

[4] Y. Goktua Siadari and D. Saripuna, “Data Mining Untuk Mengestimasi Jumlah Penumpang Pada Pt. Pinem Lau Guna Medan Dengan Menggunakan Metodere Gresi Linear Berganda,” *J. CyberTech*, vol. x. No.x, no. x, 2020.

[5] Y. P. Sari, A. Primajaya, and A. S. Y. Irawan, “Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang,” *INOVTEK Polbeng - Seri Inform.*, vol. 5, no. 2, p. 229, 2020, doi: 10.35314/isi.v5i2.1457.

[6] A. Sulistiyawati and E. Supriyanto, “Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan,” *J. Tekno Kompak*, vol. 15, no. 2, p. 25, 2021, doi: 10.33365/jtk.v15i2.1162.

[7] V. Herlinda and D. Darwis, “Analisis Clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means,” *Darwis, Dartono*, vol. 2, no. 2, pp. 94–99, 2021, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[8] E. Dianawati, P. P. Yanti, and Y. Suryandari, “Klustering Jumlah Penumpang pada Halte Bus Rapid Transit Kota Tangerang,” *J. Sist. Cerdas*, vol. 2, no. 3, pp. 163–172, 2019, doi: 10.37396/jsc.v2i3.34.

[9] A. Wibowo, Moh Makruf, Inge Virdyna, and Farah Chikita Venna, “Penentuan Klaster Koridor TransJakarta dengan Metode Majority Voting pada Algoritma Data Mining,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 565–575, 2021, doi: 10.29207/resti.v5i3.3041.

[10] S. Widaningsih, “Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm,” *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[11] J. Hutagalung and F. Sonata, “Penerapan Metode K-Means Untuk Menganalisis Minat Nasabah,” *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 1187, 2021, doi: 10.30865/mib.v5i3.3113.

[12] A. Aditya, I. Jovian, and B. N. Sari, “Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019,” *J. Media Inform. Budidarma*, vol. 4, no. 1, p. 51, 2020, doi: 10.30865/mib.v4i1.1784.

[13] K. Shao, G. Mei, and Y. Wu, “Investigating changes in global distribution of Ozone in 2018 using k -means clustering algorithm,” *J. Comput. Math. Data Sci.*, vol. 3, no. March, p. 100028, 2022, doi: 10.1016/j.jcmds.2022.100028.

[14] W. Liu, P. Zou, D. Jiang, X. Quan, and H. Dai, “Zoning of reservoir water temperature field based on K-means clustering algorithm,” *J. Hydrol. Reg. Stud.*, vol. 44, no. June, p. 101239, 2022, doi: 10.1016/j.ejrh.2022.101239.

[15] H. Priyatman, F. Sajid, and D. Haldivany, “Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa,” *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 1, p. 62, 2019, doi: 10.26418/jp.v5i1.29611.

[16] I. R. Mahartika and A. Wibowo, “Data Mining Klasterisasi dengan Algoritme K-Means untuk Pengelompokkan Provinsi Berdasarkan Konsumsi Bahan Bakar Minyak Nasional,” *Pros. Semin. Nas. SISFOTEK (Sistem Inf. dan Teknol.*, vol. 3, no. 1, pp. 87–91, 2019, [Online]. Available: https://seminar.iaii.or.id/index.php/SISFOTEK/article/view/108

[17] K. Handoko and L. Sabda Lesmana, “Computer Based Information System Journal PENGELOMPOKKAN DATA MINING PADA JUMLAH PENUMPANG DI BANDARA HANG NADIM INFORMASI ARTIKEL KATA KUNCI,” *Cbis J.*, vol. 02, 2018, [Online]. Available: http://ejournal.upbatam.ac.id/index.php/cbis

[18] E. Muningsih, I. Maryani, and V. R. Handayani, “Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi

Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, pp. 95–100, 2021, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/10428/4839

[19] S. I. Murpratiwi, I. G. Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *J. Pendidik. Teknol. dan Kejuru.*, vol. 18, no. 2, p. 152, 2021, doi: 10.23887/jptk-undiksha.v18i2.37426.

[20] N. . Anggraeni, "Teknik Clustering Dengan Algoritma K-Medoids Untuk Menangani Strategi Promosi Di Politeknik Tedc Bandung," *J. Teknol. Inf. dan Pendidik.*, 2019.