BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian ini didasarkan pada berbagai studi sebelumnya yang membahas analisis pola perjalanan penumpang dengan pendekatan data mining dan *machine learning* dalam sistem transportasi publik. Terdapat sepuluh jurnal yang menjadi referensi utama dalam penelitian ini, yang mencakup penerapan algoritma K-Means untuk segmentasi dan pengelompokkan data, serta penggunaan algoritma Apriori dalam menemukan hubungan atau pola pada data. Penelitian-penelitian terdahulu ini memberikan landasan teoritis dan metodologis yang memperkuat analisis yang dilakukan dalam studi ini, serta membantu dalam memahami bagaimana teknik analisis data dapat diterapkan untuk meningkatkan efisiensi transportasi publik, khususnya dalam konteks Transjakarta.

Tabel 2.1 Penelitian Terdahulu

No	Judul	Jurnal	Penulis	Metode	Hasil
1.	*Comparison of the DBSCAN and K-MEANS Algorithms in Segmenting Customers Using Public Transportation of Transjakarta Using the RFM Method*	*Indonesian Journal of Machine Learning and Computer Science* (MALCOM), Vol. 4, No. 4, pp. 1346-1361, 2024	Aditiya Saputra, Raka Yusuf	DBSCAN, K-Means	K-Means memiliki *Silhouette score* sebesar 0.714917, dan DBSCAN memiliki nilai 0.699971, mengindikasikan bahwa *clustering* K-Means memiliki keseragaman internal yang lebih baik dan perbedaan antar *cluster* yang lebih jelas.
2.	Penentuan *Cluster* Koridor TransJakarta dengan Metode	Jurnal Rekayasa Sistem dan Teknologi Informasi (Jurnal	Arief Wibowo, Moh Makruf, Inge Virdyna, Farah	K-Means, K-Medoids	Sebelum pandemi, algoritma K-Means menghasilkan 3 *cluster* optimal

No	Judul	Jurnal	Penulis	Metode	Hasil
	Majority Voting pada Algoritma Data Mining	Resti), Vol. 5, No. 3, pp. 565-575, 2021	Chikita Venna		dengan DBI 0,184, sedangkan selama pandemi menghasilkan 2 *cluster* optimal dengan DBI 0,188. Untuk algoritma K-Medoids, sebelum pandemi menghasilkan 3 *cluster* optimal dengan DBI 0,200, dan selama pandemi menghasilkan 4 *cluster* optimal dengan DBI 0,190.
3.	Pengelompokkan Dataset Bus Menggunakan AlgoritmaK-Means	*Informatics for Educators And Professional: Journal of Informatics,* Vol. 7, No. 2, pp. 138, 2023	Anjar Permadi, Yudhistira Arie Wijaya	K-Means	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.
4.	Analisa Penumpang Bus Transjakarta dengan	*Journal Computer Science and Information Systems: J-Cosys,* Vol.	Mohammad Farras Daffauzan	K-Means	Hasil penelitian menunjukkan bahwa algoritma K-Means memiliki tingkat

No	Judul	Jurnal	Penulis	Metode	Hasil
	Metode K-Means	3, No. 2, pp. 61-64, 2023			akurasi sebesar 70% dalam mengklasifikasikan data jenis kendaraan.
5.	Analisa Performa K-Means dan DBSCAN dalam *Clustering* Minat Penggunaan Transportasi Umum	Jurnal Elektronika dan Komputer, Vol. 14, No. 2, pp. 368-372, 2021	Ariel Kristianto	K-Means, DBSCAN	DBSCAN memiliki performa yang baik dalam proses *clustering*, terutama dengan data yang cukup banyak dan berbasis kepada kepadatan data. Hal ini ditunjukkan dengan nilai Silhouette Coefficient yang lebih besar dan mendekati 1.
6.	Pengaruh Faktor Ekonomi Dan Faktor Sosial Pada Pola Pengguna Transportasi Publik Transjakarta di Jakarta Menggunakan Metode *Association Rule*	Program Studi Ekonomi Pembangunan Fakultas Ekonomi Dan Bisnis Universitas Islam Negeri Syarif Hidayatullah, Vol. 13, No. 1, pp. 104-116, 2023	Siti Marhamatul Lathifah	*Association Rule* (Algoritma Apriori)	Pada analisis *Association Rule* menghasilkan pola sebanyak 13 pola asosiasi pengguna transjakarta. Faktor usia dan kondisi fasilitas menjadi faktor utama dan penciri dalam membentuk pola pada setiap kelompok pola asosiasi.
7.	Market Basket Analysis Dengan Metode Algoritma Apriori Untuk Menentukan Pola	Jurnal Teknologi Pintar, Vol. 3, No. 1, pp. 1-16, 2023	Ramot Simangunsong	Apriori	Proses penentuan pola pembelian produk dilakukan dengan menerapkan data mining dengan metode algoritma apriori. Dengan

No	Judul	Jurnal	Penulis	Metode	Hasil
	Pembelian Konsumen				metode tersebut penentuan pola pembelian dapat dilakukan dengan melihat hasil dari kecenderungan konsumen membeli produk berdasarkan dari hasil 2 kombinasi itemset.
8.	Integrasi Algoritma Apriori Dan K-Means Dalam Analisis Pola Pembelian Untuk Meningkatkan Strategi Pemasaran	Jurnal Ilmiah Penelitian dan Pembelajaran Informatika, Vol. 10, No. 1, pp. 409-423, 2025	Violita Eka Putri, Hindriyanto Dwi Purnomo	Apriori, K-Means	K-Means berhasil mengelompokkan data dalam 4 *cluster* dengan nilai *Davies Bouldin Index* (DBI) sebesar 0,465. Penggunaan algoritma Apriori dengan dukungan minimum 0,01 dan kepercayaan minimum 0,5, *cluster* 0 menghasilkan 1 aturan dengan kepercayaan tertinggi 75%, sementara *cluster* 3, meskipun dengan dataset lebih kecil (127 transaksi), menghasilkan 16 aturan dengan kepercayaan tertinggi 100%.
9.	Analisis Penjualan Produk Menggunakan Algoritma K-	G-Tech: Jurnal Teknologi Terapan, Vol. 8, No. 2,	Akrim Teguh Suseno	K-Means dan Apriori	Hasil penelitian dengan penggabungan teknik *clustering* (K-

No	Judul	Jurnal	Penulis	Metode	Hasil
	Means dan Apriori	pp. 1288–1296, 2024			Means) dan *association rule* (Apriori) dapat memberikan rekomendasi terhadap transaksi penjualan di Distro Sextors. Selain itu terdapat 5 rule yang dihasilkan dengan memberikan rekomendasi 7 produk yang dapat dijadikan media promosi.
10.	Penerapan Metode K-Means Dan Apriori Untuk Pemilihan Produk Bundling	Journal CERITA: *Creative Education of Research in Information Technology and Artificial informatics*	Syifa Aryanti, Deni Mahdiana, Ade Setiadi	K-Means Dan Apriori	Model Penerapan Metode K-Means Dan Apriori Untuk Pemilihan Produk Bundling menggunakan hasil pengujian dengan pendekatan metode McCall menyatakan bahwa sistem dapat memenuhi kebutuhan user dengan nilai sebesar 83,58% (baik)

Penelitian pertama membandingkan algoritma DBSCAN dan K-Means dalam segmentasi pelanggan menggunakan metode RFM (*Recency*, *Frequency*, *Monetary*). Hasil analisis menunjukkan bahwa K-Means lebih unggul dalam kualitas *clustering* dibandingkan DBSCAN. K-Means memiliki *Silhouette Score* sebesar 0.714917, lebih tinggi dari DBSCAN 0.699971, yang menunjukkan keseragaman internal yang lebih baik. Selain itu, *Davies-Bouldin Index* (DBI) K-Means (0.365776) lebih rendah dibandingkan DBSCAN (0.390784), menandakan

bahwa *cluster* yang dihasilkan lebih kompak dan terpisah dengan baik [9]. Penelitian kedua menerapkan algoritma K-Means dan K-Medoids untuk mengelompokkan data koridor Transjakarta sebelum dan selama pandemi Covid-19. Hasil analisis menunjukkan bahwa dengan K-Means, jumlah *cluster* optimal sebelum pandemi adalah 3 *cluster* (DBI 0,184), sedangkan selama pandemi menjadi 2 *cluster* (DBI 0,188). Sementara itu, dengan K-Medoids, jumlah *cluster* optimal sebelum pandemi adalah 3 *cluster* (DBI 0,200), dan selama pandemi meningkat menjadi 4 *cluster* (DBI 0,190). Hal ini menunjukkan adanya perubahan pola pengelompokan koridor Transjakarta selama pandemi [11].

Penelitian ketiga menggunakan metode K-Means *clustering* pada aplikasi RapidMiner untuk mengelompokkan dataset bus. Jumlah *cluster* optimal ditentukan menggunakan parameter default, dan berdasarkan *Davies Bouldin Index* (DBI), k=3 adalah jumlah *cluster* terbaik dengan nilai 0,530. Pengelompokan menghasilkan tiga *cluster*, yaitu *Cluster* 0 (1972 item), *Cluster* 1 (1612 item), dan *Cluster* 2 (140 item). Hasil analisis ini memberikan wawasan tentang karakteristik atribut dalam dataset bus [12]. Penelitian keempat mengungkapkan bahwa algoritma K-Means mampu mengklasifikasikan data jenis kendaraan dengan tingkat akurasi 70%. Oleh karena itu, dapat disimpulkan bahwa algoritma K-Means dapat diterapkan untuk menganalisis data transaksi bus Transjakarta serta mengelompokkan pelanggan berdasarkan jenis kendaraan yang paling sering mereka gunakan berdasarkan data yang tersedia [13].

Penelitian kelima menunjukkan bahwa algoritma DBSCAN menunjukkan kinerja yang optimal dalam proses *clustering*, terutama ketika digunakan pada data berukuran besar dan berbasis kepadatan. Hal ini dibuktikan dengan nilai *Silhouette Coefficient* yang lebih tinggi dan mendekati 1, menandakan kualitas *clustering* yang baik [14]. Penelitian keenam menggunakan metode *Association Rule* dan berhasil membentuk 13 pola asosiasi, dengan usia (15-64 tahun) dan kondisi fasilitas sebagai faktor utama dalam membentuk pola. Faktor lain yang berpengaruh adalah frekuensi penggunaan lebih dari 3 kali, biaya perjalanan kurang dari Rp. 10.000, serta kondisi sarana dan prasarana yang baik [15].

Penelitian ketujuh menunjukkan penerapan algoritma Apriori pada data laporan penjualan dapat membantu menemukan pola kombinasi itemset yang berharga untuk pengambilan keputusan terkait stok barang. Dengan metode *data mining*, pola pembelian produk dapat dianalisis berdasarkan kecenderungan konsumen dalam membeli produk dari kombinasi dua itemset, sehingga membantu optimalisasi strategi penjualan [16]. Penelitian kedepalan menggunakan metode *clustering* yang kemudian dilanjutkan dengan *association rules.* Penelitian ini menghasilkan 4 *cluster* ideal dengan nilai validitas *Davies Bouldin Index* (DBI) sebesar 0,465. Analisis aturan asosiasi difokuskan pada tiga *cluster* yang solusi, yaitu *cluster* 0, 1, dan 3 yang memerlukan peningkatan penjualan produk [17].

Penelitian kesembilan memanfaatkan kombinasi teknik *clustering* (K-Means) dan *association rule* (Apriori) untuk memberikan rekomendasi terkait transaksi penjualan di Distro Sextors. Hasil penelitian ini menghasilkan 5 aturan (*rule*) yang merekomendasikan 7 produk sebagai media promosi, yang ditandai dengan kode produk SE130, SE111, SE128, SE126, SE04, SE40, dan SE11 [18]. Penelitian kesepuluh menerapkan metode K-Means dan Apriori untuk pemilihan produk bundling. Hasilnya menunjukkan bahwa metode K-Means menghasilkan 2 *cluster*, yaitu 58 item *slow-moving* (C0) dan 4 item *fast-moving* (C1), dengan DBI sebesar 0,106, menunjukkan tingkat keakuratan yang optimal. Analisis asosiasi menggunakan RapidMiner mengidentifikasi *Chopper*, *Fork*, dan *Spoon* sebagai *frequent itemset* dengan *support* 12,76%, *confidence* 97,80%, dan *lift ratio* 1,59, menunjukkan bahwa produk tersebut sering dibeli bersamaan, sehingga dapat dijadikan strategi bundling oleh PT. MSD. Pengujian sistem dengan metode *black-box* dan *white-box testing* menggunakan teknik *Equivalence Partitioning* menunjukkan bahwa sistem bekerja sesuai kebutuhan dan valid. Selain itu, berdasarkan evaluasi dengan metode *McCall*, sistem memenuhi kebutuhan pengguna dengan skor 83,58% (baik) [19].

Berdasarkan penelitian-penelitian terdahulu tersebut, penelitian ini memiliki perbedaan signifikan karena secara khusus berfokus pada analisis pola perjalanan dan identifikasi hubungan antar halte yang sering digunakan penumpang Transjakarta. Penelitian ini menggunakan kombinasi algoritma K-Means dan

Apriori yang dipilih karena sesuai dengan tujuan yang ingin dicapai. Algoritma K-Means digunakan untuk mengelompokkan penumpang berdasarkan karakteristik pola perjalanan mereka, seperti durasi, waktu, frekuensi, dan jarak perjalanan, sehingga dapat mengungkap segmentasi perilaku mobilitas. Sementara itu, algoritma Apriori digunakan untuk mengeksplorasi hubungan antar halte yang sering digunakan secara bersamaan oleh penumpang, sehingga dapat mengidentifikasi pola asosiasi antar titik perjalanan. Meskipun memiliki tujuan analisis yang berbeda, kedua algoritma ini saling melengkapi dalam memberikan gambaran yang lebih utuh mengenai perilaku dan pergerakan penumpang Transjakarta.

2.2 Teori tentang Topik Skripsi

2.2.1 Transportasi Publik

Transportasi publik adalah sistem angkutan yang dirancang untuk memenuhi kebutuhan mobilitas masyarakat luas dengan menyediakan layanan transportasi yang dapat diakses oleh banyak orang [20]. Layanan ini umumnya menggunakan kendaraan besar seperti bus, kereta, atau angkutan massal lainnya yang memiliki kapasitas angkut yang lebih tinggi dibandingkan dengan kendaraan pribadi. Transportasi publik berfungsi untuk memudahkan perpindahan orang dalam jarak jauh atau dekat dengan biaya yang lebih rendah dan waktu tempuh yang lebih efisien. Di banyak kota besar, termasuk Jakarta, transportasi publik menjadi salah satu solusi utama dalam mengatasi masalah kemacetan, mengurangi polusi, serta memfasilitasi mobilitas penduduk yang terus berkembang [21].

Tujuan utama dari transportasi publik adalah untuk menyediakan sistem transportasi yang efisien, terjangkau, dan ramah lingkungan bagi masyarakat umum. Dalam hal ini, efisiensi tercapai dengan meningkatkan kapasitas angkut kendaraan, memperpendek waktu perjalanan, serta memastikan bahwa armada dan rute yang digunakan dapat memenuhi permintaan masyarakat secara optimal. Selain itu, tujuan lainnya adalah untuk mengurangi ketergantungan pada kendaraan pribadi, yang dapat membantu mengurangi tingkat kemacetan di jalan raya, serta mengurangi emisi karbon dan polusi

udara yang sering kali dihasilkan oleh kendaraan pribadi [22]. Transportasi publik yang baik juga bertujuan untuk mendukung kegiatan ekonomi dan sosial dengan memfasilitasi mobilitas pekerja, pelajar, dan masyarakat umum ke tempat-tempat yang mereka butuhkan untuk beraktivitas.

Sebagai contoh nyata dari sistem transportasi publik yang efisien di Jakarta adalah Transjakarta, yang menggunakan model *Bus Rapid Transit* (BRT). Transjakarta telah menjadi solusi utama untuk mengatasi kemacetan di ibu kota dengan menyediakan angkutan umum yang terjangkau dan dapat diandalkan. Transjakarta juga merupakan model transportasi yang mengutamakan efisiensi dan kapasitas tinggi dengan biaya yang relatif lebih rendah daripada sistem angkutan massal lainnya [23].

2.2.2 Transjakarta

Transjakarta adalah sistem transportasi publik berbasis bus yang beroperasi di Jakarta dan sekitarnya. Diluncurkan pada tahun 2004, Transjakarta bertujuan untuk mengurangi kemacetan dan meningkatkan mobilitas masyarakat di ibu kota yang padat. Sistem ini menggunakan jalur khusus yang disebut Bus Rapid Transit (BRT), yang memisahkan bus dari kendaraan pribadi untuk menghindari kemacetan di jalan raya [24]. Keberadaan jalur khusus ini memungkinkan Transjakarta untuk beroperasi dengan lebih efisien, memperpendek waktu tempuh, dan memberikan kenyamanan lebih bagi para penggunanya. Selain itu, Transjakarta juga menawarkan harga tiket yang terjangkau, menjadikannya pilihan utama bagi masyarakat yang membutuhkan transportasi cepat dan ekonomis di Jakarta [25].

Salah satu keunggulan Transjakarta adalah kapasitas angkutnya yang tinggi. Dengan armada bus yang besar dan adanya sistem BRT yang mengutamakan kepadatan penumpang, Transjakarta mampu melayani ratusan ribu penumpang setiap harinya. Hal ini sangat penting mengingat Jakarta adalah kota dengan tingkat kemacetan yang tinggi, dan transportasi publik yang efisien menjadi kebutuhan mendesak. Transjakarta tidak hanya menghubungkan berbagai kawasan di Jakarta, tetapi juga menjangkau daerah-

daerah penyangga kota, seperti Depok, Bogor, dan Bekasi, yang semakin memperluas jaringan mobilitas masyarakat. Dengan integrasi berbagai moda transportasi, seperti KRL, MRT, dan LRT, Transjakarta juga berperan penting dalam menciptakan sistem transportasi yang terkoordinasi dan memudahkan perpindahan antar moda [26].

2.3 Teori tentang Algoritma yang digunakan

4.3.1 Data Mining

Data mining adalah proses eksplorasi dan analisis data besar untuk menemukan pola, hubungan, atau informasi tersembunyi yang dapat digunakan untuk pengambilan keputusan yang lebih baik [27]. Data mining sering diterapkan dalam berbagai domain, termasuk pemasaran, kesehatan, keuangan, dan sistem transportasi, untuk memperoleh pengetahuan yang berguna dari data yang ada. Beberapa teknik utama dalam data mining mencakup klasifikasi, *clustering*, asosiasi, dan regresi, yang semuanya memiliki tujuan dan aplikasi yang berbeda-beda tergantung pada jenis data dan tujuan analisis [28].

Salah satu tujuan utama dari data mining adalah untuk mengidentifikasi pola yang dapat memprediksi kejadian di masa depan atau memberikan wawasan yang berguna untuk pengambilan keputusan [27]. Misalnya, dalam sistem transportasi publik seperti Transjakarta, data mining dapat digunakan untuk menganalisis pola perjalanan penumpang. Dengan memanfaatkan data yang terkumpul dari data historis penumpang, perusahaan Transjakarta dapat merancang strategi operasional yang lebih efisien, meningkatkan pelayanan, dan mengurangi waktu tunggu bagi pengguna.

Proses data mining dimulai dengan tahap pembersihan data, di mana data yang tidak relevan atau cacat dihapus atau diperbaiki. Setelah itu, data yang sudah bersih dapat dianalisis menggunakan algoritma yang sesuai untuk menemukan pola atau hubungan yang tersembunyi. Dalam tahap ini, teknik seperti *clustering* dapat digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik, sedangkan klasifikasi dapat digunakan untuk memprediksi kategori atau status dari data baru. Hasil dari proses ini biasanya

berupa model prediktif atau aturan asosiasi yang dapat digunakan untuk meningkatkan kinerja sistem atau membuat keputusan yang lebih informasional [28].

Penggunaan algoritma pembelajaran mesin memungkinkan sistem untuk belajar secara otomatis dari data yang ada dan meningkatkan akurasi prediksi atau klasifikasi seiring waktu [29]. Dalam konteks transportasi, hal ini dapat menyesuaikan layanan berdasarkan pola perjalanan penumpang, memperkirakan lonjakan penumpang pada jam sibuk, dan mengoptimalkan halte yang sering digunakan dalam perjalanan untuk mengurangi kemacetan.

4.3.2 K-Means

K-Means adalah algoritma *clustering* yang digunakan untuk membagi data ke dalam kelompok-kelompok atau *cluster* berdasarkan kemiripan antar data. Algoritma ini termasuk dalam kategori *unsupervised learning*, di mana data yang digunakan tidak memiliki label atau kategori yang sudah ditentukan sebelumnya [30]. K-Means bertujuan untuk mengelompokkan data sehingga setiap data dalam satu *cluster* memiliki jarak yang lebih dekat dengan pusat *cluster* (*centroid*) dibandingkan dengan data di *cluster* lain [31].

Penerapan algoritma K-means dalam analisis data terdiri dari beberapa tahapan [32]. Berikut adalah tahapan-tahapan dalam penerapan K-means:

1. Menentukan sejumlah *cluster* (k) yang diinginkan.
2. Menentukan titik tengah atau *centroid* untuk setiap *cluster*.
3. Mengelompokkan data berdasarkan kedekatannya dengan pusat *cluster* menggunakan pengukuran jarak (*Euclidean Distance*).

   Berikut adalah rumus jarak *Euclidean* antara dua titik $x = (x_1,\ x_2)$ dan $y = (y_1,\ y_2)$:

   $$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

   Rumus 2.1 Jarak *Euclidean* antara dua titik dalam ruang dua dimensi

   Sementara itu, berikut adalah rumus jarak *Euclidean* antara titik $x = (x_1,\ x_2,\ \ldots,\ x_n)$ dan $y = (y_1,\ y_2,\ \ldots,\ y_n)$ dalam ruang berdimensi n:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Rumus 2.2 Jarak *Euclidean* antara dua titik dalam ruang berdimensi *n*

Keterangan:

- x_i dan y_i adalah komponen koordinat dari titik x dan y dalam dimensi ke-i.
- n adalah jumlah dimensi.

4. Menghitung ulang posisi *centroid* untuk setiap *cluster* dengan menghitung rata-rata dari semua titik data yang ada dalam *cluster*. Berikut adalah rumus untuk menghitung *centroid* baru:

$$C_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_i$$

Rumus 2.3 Menghitung *centroid* suatu *cluster*

Keterangan:

- N_k adalah jumlah titik dalam *cluster* k.
- x_I adalah posisi setiap titik di dalam *cluster*.

5. Melakukan iterasi langkah 3 dan 4 sampai perubahan posisi pusat *cluster* atau *centroid* menjadi sangat kecil atau tidak ada perubahan sama sekali.

K-Means memiliki keunggulan dalam hal kesederhanaannya dan efisiensi dalam menangani dataset besar karena cepat dan mudah digunakan, sehingga sering digunakan dalam berbagai aplikasi seperti segmentasi data dan pengelompokan data [31]. Namun, K-Means memiliki beberapa kekurangan, seperti ketergantungannya pada pemilihan jumlah *cluster* (k) yang tepat, dan kerentanannya terhadap data yang tidak terdistribusi dengan baik atau outlier.

4.3.3 Elbow Method

Elbow Method adalah salah satu teknik yang digunakan untuk menentukan jumlah *cluster* optimal dalam algoritma *clustering* seperti K-Means. Metode ini dilakukan dengan menghitung nilai *Within-Cluster Sum of*

Squares (WCSS) untuk berbagai jumlah *cluster* (K), lalu memplot nilai WCSS terhadap K. WCSS mengukur seberapa dekat data dalam suatu *cluster* terhadap *centroid*-nya [33]. Semakin kecil nilai WCSS, maka semakin baik data dikelompokkan. Namun, penambahan jumlah *cluster* secara terus-menerus akan menyebabkan penurunan WCSS yang semakin kecil atau tidak signifikan. Berikut adalah rumus untuk menghitung WCSS [34]:

$$WCSS = \sum_{k=1}^{K} \sum_{i \epsilon C_k} \|x_i - \mu_k\|^2$$

Rumus 2.4 Mengitung nilai *Within-Cluster Sum of Squares* (WCSS)

Dengan C_k adalah *cluster* ke-k, x_i adalah data pada *cluster* tersebut, dan μ_k adalah *centroid cluster* ke-k. Dalam grafik *Elbow Method*, titik "siku" atau *elbow* ditandai sebagai titik di mana penurunan nilai WCSS mulai melambat secara signifikan. Titik ini dianggap sebagai jumlah *cluster* optimal karena setelah titik tersebut, penambahan *cluster* tidak memberikan peningkatan yang signifikan dalam kualitas *cluster*.

4.3.4 Davies-Bouldin Index

Blablala *Davies-Bouldin Index* (DBI) adalah salah satu metode evaluasi untuk mengukur kualitas hasil *clustering*. Indeks ini mengevaluasi sejauh mana *cluster-cluster* yang terbentuk saling terpisah dan seberapa kompak data dalam masing-masing *cluster*. DBI mempertimbangkan rasio antara jarak antar *cluster* dengan ukuran dispersi dalam *cluster* itu sendiri. Semakin kecil nilai DBI, maka semakin baik kualitas suatu *cluster*, karena hal tersebut menunjukkan *cluster* yang lebih terpisah dan kompak [35]. Berikut adalah rumus dari *Davies-Bouldin Index* [36]:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max i \neq j \left(\frac{S_i + S_j}{M_{ij}} \right)$$

Rumus 2.5 Menghitung nilai *Davies-Bouldin Index* (DBI)

Dengan K adalah jumlah *cluster*, S_i dan S_j adalah ukuran dispersi dari *cluster* ke-i dan ke-j, dan M_{ij} adalah jarak antara *centroid cluster* ke-i dan ke-

j. Nilai maksimum dari rasio tersebut dihitung untuk setiap *cluster*, lalu dirata-ratakan untuk memperoleh nilai DBI secara keseluruhan.

4.3.5 Calinski-Harabasz Index

Calinski-Harabasz Index adalah metrik evaluasi yang digunakan untuk menilai kualitas *clustering* berdasarkan rasio antara dispersi antar *cluster* dan dispersi dalam *cluster*. Semakin besar nilai CH indeks, maka semakin baik kualitas *cluster* yang dihasilkan, karena menunjukkan bahwa *cluster* antar kelompok lebih terpisah [37]. Berikut adalah rumus dari *Calinski-Harabasz Index* [38]:

$$CH = \frac{Trace(SSB)}{Trace(SSW)} \times \frac{n-k}{k-1}$$

Rumus 2.6 Menghitung nilai *Calinski-Harabasz Index*

Dengan $Trace(SSB)$ adalah jumlah kuadrat antar *cluster* (*Sum of Square Between*), $Trace(SSW)$ adalah jumlah kuadrat dalam *cluster* (*Sum of Square Within*), n adalah jumlah total data, dan k adalah jumlah *cluster*. Nilai CH yang tinggi mengindikasikan bahwa *cluster* memiliki pemisahan yang jelas dan struktur yang baik.

4.3.6 Apriori

Apriori adalah salah satu algoritma dalam data mining yang digunakan untuk menemukan aturan asosiasi dalam suatu dataset. Algoritma ini dikembangkan untuk menggali pola yang sering muncul secara bersamaan dalam transaksi atau kejadian yang tercatat dalam data. Prinsip dasar dari algoritma Apriori adalah *frequent itemset mining*, di mana algoritma ini mencari itemset yang sering muncul dalam dataset dan kemudian menghasilkan aturan asosiasi berdasarkan itemset tersebut [39]. Sebagai contoh, dalam analisis transaksi, Apriori dapat digunakan untuk mencari tahu kombinasi produk apa saja yang sering dibeli bersama oleh konsumen.

Proses kerja algoritma Apriori dimulai dengan menentukan frekuensi minimum itemset yang dianggap sering muncul dalam data. Kemudian, algoritma ini menggunakan pendekatan iteratif untuk menghasilkan kombinasi item yang lebih besar berdasarkan itemset yang telah ditemukan

pada iterasi sebelumnya [40]. Hasil akhir dari Apriori adalah aturan asosiasi yang berbentuk *If-Then*, yang menunjukkan hubungan antara item-item dalam dataset, misalnya: "Jika pelanggan membeli produk A, mereka cenderung membeli produk B."

Dalam menggunakan algoritma Apriori, penting untuk menetapkan nilai minimum untuk *support* dan *confidence* yang akan digunakan sebagai ambang batas dalam menentukan pola. Nilai *support* digunakan untuk mengukur seberapa sering suatu kombinasi item muncul dalam seluruh dataset [41]. Berikut adalah rumus untuk menghitung nilai *support* sebuah item [42]:

$$Support(A) = \frac{Jumlah\ transaksi\ yang\ mengandung\ A}{Total\ jumlah\ transaksi}$$

Rumus 2.7 Menghitung nilai *support* sebuah item

Sementara itu, berikut adalah rumus untuk menghitung nilai *support* dua item [42]:

$$Support(A \cup B) = \frac{Jumlah\ transaksi\ yang\ mengandung\ A\ dan\ B}{Total\ jumlah\ transaksi}$$

Rumus 2.8 Menghitung nilai *support* dua item

Sedangkan, nilai *confidence* digunakan untuk mengukur seberapa besar kemungkinan item B muncul jika item A sudah muncul [41]. Berikut adalah rumus untuk menghitung nilai *confidence* [42]:

$$Confidence\ (A \rightarrow B) = \frac{Jumlah\ transaksi\ yang\ mengandung\ A\ dan\ B}{Jumlah\ transaksi\ yang\ mengandung\ A}$$

Rumus 2.9 Menghitung nilai *confidence*

Selain nilai *support* dan *confidence*, nilai *lift* juga penting digunakan untuk mengevaluasi kekuatan suatu aturan asosiasi. Nilai *lift* menunjukkan hubungan antara dua item apakah bersifat mendukung satu sama lain atau tidak. Berikut adalah rumus untuk menghitung nilai *lift* [43]:

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)}$$

Rumus 2.10 Menghitung nilai lift

Apriori memiliki berbagai aplikasi dalam berbagai domain, termasuk di bidang pemasaran, *e-commerce*, dan manajemen rantai pasokan. Dalam konteks sistem transportasi, seperti Transjakarta, algoritma Apriori dapat digunakan untuk menemukan pola atau hubungan antara penggunaan halte tertentu oleh penumpang. Misalnya, Apriori dapat mengidentifikasi halte yang sering digunakan bersamaan pada waktu-waktu tertentu atau mengungkapkan pola perjalanan penumpang berdasarkan hari atau jam tertentu. Dengan temuan ini, operator transportasi dapat mengoptimalkan rute, jadwal, dan kapasitas armada untuk meningkatkan efisiensi layanan dan kenyamanan penumpang.

2.4 Teori tentang tools/software yang digunakan

2.4.1 Jupyter Notebook

Jupyter Notebook merupakan aplikasi sumber terbuka yang digunakan untuk membuat dan berbagi dokumen yang menggabungkan kode yang dapat dijalankan, visualisasi, serta teks naratif [44]. Jupyter Notebook mendukung berbagai bahasa pemrograman, termasuk Python, R, Julia, Java dan Python adalah bahasa yang paling sering digunakan. Dengan Jupyter Notebook, pengguna dapat menulis kode secara langsung, menjalankan kode, dan melihat hasilnya secara langsung di dalam dokumen yang sama. Hal ini Jupyter Notebook efisien untuk eksplorasi data, analisis statistik, dan pengembangan model pembelajaran mesin.

Keunggulan utama Jupyter Notebook terletak pada interaktivitas dan fleksibilitasnya. Pengguna dapat menulis dan menjalankan kode secara bertahap, serta memungkinkan eksperimen yang lebih cepat dan iteratif tanpa perlu menjalankan seluruh program setiap saat [45]. Selain itu, Jupyter Notebook juga mendukung berbagai jenis visualisasi, seperti grafik dan diagram, yang memungkinkan pengguna untuk menampilkan hasil analisis data dalam bentuk yang mudah dipahami. Misalnya, dengan menggunakan pustaka Python seperti Matplotlib, Seaborn, atau Plotly, pengguna dapat membuat visualisasi data langsung dari kode yang telah dijalankan, memudahkan untuk melakukan eksplorasi data secara visual [46].

2.4.2 Python

Python adalah bahasa pemrograman yang populer dan banyak digunakan di berbagai bidang, termasuk pengembangan perangkat lunak, analisis data, kecerdasan buatan, dan automasi. Bahasa ini dikenal karena sintaksisnya yang sederhana dan mudah dipahami, menjadikannya pilihan utama baik untuk pemula maupun profesional berpengalaman. Python mendukung berbagai paradigma pemrograman, seperti pemrograman prosedural, berorientasi objek, dan fungsional, yang memberikan fleksibilitas bagi pengembang untuk memilih pendekatan yang sesuai dengan kebutuhan [47].

Salah satu keunggulan utama Python adalah pustakanya yang sangat luas, seperti NumPy dan Pandas untuk analisis data, Matplotlib dan Seaborn untuk visualisasi, serta TensorFlow dan scikit-learn untuk machine learning [46]. Hal ini membuat Python menjadi bahasa yang sangat kuat dalam ilmu data dan analisis statistik. Python juga dikenal karena kemampuannya untuk mengolah data dalam jumlah besar dengan kecepatan yang optimal. Hal ini sangat penting dalam konteks analisis data dan kecerdasan buatan, di mana pengolahan data yang cepat dan akurat sangat diperlukan. Dengan berbagai pustaka dan alat yang tersedia, Python menjadi bahasa pilihan dalam bidang ilmiah dan teknologi untuk melakukan penelitian, eksperimen, dan pengembangan perangkat lunak berbasis data.