BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini berfokus pada analisis pola perjalanan penumpang Transjakarta dengan menggunakan algoritma K-Means dan Apriori. Data yang digunakan dalam penelitian ini merupakan dataset transaksi perjalanan penumpang Transjakarta, yang diperoleh dari platform Kaggle dan mencakup informasi mengenai koridor yang digunakan, titik awal keberangkatan, titik akhir perjalanan, waktu perjalanan, serta total transaksi yang terjadi [48]. Dengan menerapkan algoritma K-Means, penelitian ini bertujuan untuk mengelompokkan penumpang berdasarkan kesamaan pola perjalanan mereka, sedangkan algoritma Apriori digunakan untuk menemukan asosiasi antara halte yang sering digunakan secara bersamaan. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan studi mengenai analisis perilaku mobilitas penumpang serta penerapan data mining dalam sistem transportasi publik.

3.2 Metode Penelitian

3.2.1 Alur Penelitian

Alur penelitian ini mengikuti metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang terdiri dari enam tahap utama. Langkah pertama adalah *Business Understanding* untuk memahami tujuan penelitian dalam menganalisis pola perjalanan penumpang Transjakarta. Selanjutnya, pada tahap *Data Understanding* data transaksi perjalanan akan dieksplorasi untuk memahami karakteristik datanya. Setelah itu, dilakukan *Data Preparation*, yaitu proses membersihkan data agar data siap digunakan dalam analisis. Tahap berikutnya adalah *Modeling*, yaitu penerapan algoritma K-Means untuk *clusterisasi* pola perjalanan dan algoritma Apriori untuk menemukan asosiasi antar halte. Setelah model terbentuk, dilakukan *Evaluation* untuk menilai keakuratan dan relevansi hasil analisis. Terakhir,

pada tahap *Deployment*, hasil penelitian akan dikaji untuk memberikan rekomendasi bagi peningkatan layanan Transjakarta.

3.2.2 Metode Data Mining

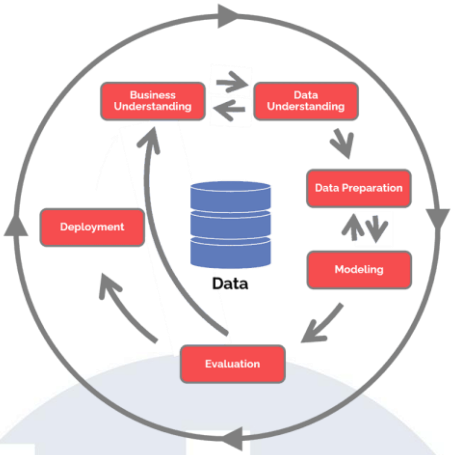
Dalam proses *data mining*, terdapat tiga *framework* utama yang digunakan untuk menggali pola dan informasi berharga dari data. Setiap framework memiliki tahapan, kelebihan, dan kekurangan yang berbeda dalam menganalisis data, tergantung pada karakteristik dataset serta tujuan penelitian. Berikut tabel perbandingan ketiga *framework* tersebut.

Tabel 3.1 Perbandingan Metode Data Mining

Indikator	CRISP DM	SEMMA	KDD
Tahapan	Terdiri dari 6 tahapan, yaitu *Business understanding, Data Understanding, Data preparation, Modeling, Evaluation,* dan *Deployment.*	Tediri dari 5 tahapan, yaitu *Sample, Explore, Modify, Model,* dan *Assessment.*	Terdiri dari 6 tahapan, yaitu *Selection, Preprocessing, Transformation, Data mining, Interpretation/Evaluation,* dan *Knowledge Representation.*
Fokus Tujuan	Aplikasi lintas industri dengan penekanan pada proses bisnis.	Fokus pada analisis dan pemodelan data.	Penemuan pengetahuan baru atau pola menarik dalam basis data.
Kelebihan	Kelebihan CRISP DM adalah memiliki siklus yang dapat	Kelebihan SEMMA adalah terfokus pada analisis dan	Kelebihan KDD adalah melibatkan seluruh proses

Indikator	CRISP DM	SEMMA	KDD
	diulang, menekankan pada pemahaman bisnis awal, dan dapat diterapkan pada berbagai industri.	pemodelan data, sehingga cocok untuk proyek-proyek dengan fokus pada aspek analisis. Serta strukturnya sederhana dan jelas.	dari seleksi data hingga representasi pengetahuan dan fokus pada penemuan pengetahuan baru dalam data.
Kekurangan	Kekurangan CRISP DM adalah terlalu kompleks untuk proyek-proyek kecil atau sederhana, kurang cocok untuk proyek riset murni tanpa tujuan bisnis yang jelas.	Kekurangan SEMMA adalah kurang fleksibel dalam hal iterasi atau penyesuaian tahapan dan kurang menekankan pemahaman bisnis awal dibandingkan dengan CRISP-DM.	Kekurangan KDD adalah proses yang terlalu luas dan kompleks, cocok untuk proyek-proyek besar tetapi mungkin terlalu berlebihan untuk proyek-proyek kecil.

Berdasarkan tabel perbandingan di atas, penelitian ini menggunakan metode *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai kerangka kerja dalam menganalisis pola perjalanan penumpang Transjakarta. CRISP-DM terdiri dari enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*.



Gambar 3.1 Alur Penelitian CRISP DM

Berikut adalah penjelasan dari alur penelitian yang akan diikuti:

1. *Business Understanding*

   Tahap pertama adalah memahami tujuan dan kebutuhan penelitian dalam konteks analisis pola perjalanan penumpang Transjakarta. Penelitian ini bertujuan untuk mengidentifikasi pola perjalanan berdasarkan transaksi penumpang dan menemukan hubungan antar halte yang sering digunakan secara bersamaan. Hasil analisis diharapkan dapat memberikan wawasan bagi pengelola Transjakarta dalam meningkatkan efisiensi layanan transportasi dan strategi operasional.

2. *Data Understanding*

   Pada tahap ini, dilakukan eksplorasi terhadap dataset transaksi perjalanan Transjakarta untuk memperoleh pemahaman yang lebih dalam mengenai struktur dan karakteristik data. Data yang digunakan terdiri dari 13 kolom, yaitu **transID**, **payCardID**, **corridorID**, **corridorName**, **tapInStops**, **tapInStopsName**, **stopStartSeq**, **tapInTime**, **tapOutStops**, **tapOutStopsName**, **stopEndSeq**, **tapOutTime**, **payAmount**.

   Langkah awal pada proses ini dilakukan dengan mengekspor dan membaca dataset dengan menggunakan pustaka pandas untuk melihat gambaran data secara umum. Kemudian, informasi mengenai

tipe data setiap kolom juga ditampilkan untuk melihat jumlah baris data yang tersedia dan melihat adanya nilai yang hilang atau tidak sesuai.

Selain itu, dilakukan juga beberapa eksplorasi data visual (*Exploratory Data Analysis* / EDA) untuk memahami pola dan persebaran data dalam dataset. *Exploratory Data Analysis* (EDA) adalah proses awal dalam analisis data yang bertujuan untuk memahami struktur, karakteristik, dan pola dalam dataset sebelum dilakukan pemodelan lebih lanjut [49]. EDA dilakukan melalui teknik statistik deskriptif dan visualisasi data, seperti grafik batang dan heatmap.

3. *Data Preparation*

Tahap *data preparation* merupakan langkah penting yang dilakukan sebelum proses pemodelan data. Tahapan ini bertujuan untuk membersihkan dan mempersiapkan data agar siap digunakan dalam pemodelan. Langkah-langkah yang dilakukan dalam tahap ini meliputi penanganan data yang hilang (*missing values*), konversi format data, pembuatan fitur-fitur baru (*feature engineering*), serta proses *data encoding*.

Langkah pertama dalam tahap ini adalah menangani data yang hilang (*missing values*). *Missing value* adalah kondisi di mana suatu entri dalam dataset tidak memiliki nilai atau datanya hilang. Jika tidak ditangani dengan baik, *missing value* dapat memengaruhi kualitas analisis dan hasil model [50]. Oleh karena itu, dilakukan proses *data cleaning*, yaitu proses pembersihan data dari nilai-nilai yang tidak lengkap, tidak konsisten, atau tidak relevan agar data yang digunakan dalam analisis menjadi lebih akurat dan dapat diandalkan [51]. Dalam penelitian ini, proses *data cleaning* dilakukan dengan cara menghapus baris-baris data yang memiliki nilai kosong menggunakan fungsi df.dropna(). Metode penghapusan baris dipilih karena jumlah data yang tersedia masih cukup besar untuk dilakukan

analisis dan pemodelan, sehingga penghapusan beberapa baris data tidak memberikan dampak signifikan terhadap hasil analisis.

Selanjutnya, dilakukan konversi format data pada kolom **tapInTime** dan **tapOutTime** yang semula bertipe *object* (*string*) menjadi tipe *datetime*. Konversi ini penting dilakukan karena data waktu dalam format *object* tidak dapat digunakan secara langsung untuk analisis berbasis waktu, seperti menghitung durasi perjalanan.

Tahap ini juga mencakup *feature engineering* atau pembuatan fitur baru yang bertujuan untuk memperkaya informasi dalam data. Terakhir, dilakukan proses *data encoding* untuk mengubah data kategorikal dan data bertipe boolean menjadi format numerik agar dapat diproses oleh algoritma K-Means. Teknik yang digunakan meliputi teknik *label encoding* dan *one-hot encoding*. Selama proses ini, dilakukan juga penyimpanan *mapping* antara nilai kategorikal asli dan bentuk numeriknya ke dalam sebuah struktur data (label_mappings). *Mapping* ini disimpan agar nilai-nilai numerik tersebut dapat dikembalikan ke bentuk kategorikal aslinya, terutama saat akan digunakan dalam algoritma Apriori yang membutuhkan interpretasi halte secara eksplisit.

Data yang telah melalui proses *preparation* kemudian disesuaikan dan dibagi sesuai dengan kebutuhan dua algoritma dalam penelitian, yaitu algoritma K-Means untuk segmentasi pola perjalanan, dan Apriori untuk analisis asosiasi antar halte yang sering digunakan bersama.

4. *Modeling*

Tahap Modeling merupakan proses penerapan metode data mining untuk menemukan pola dari data yang telah dipersiapkan. Dalam penelitian ini, digunakan dua algoritma, yaitu K-Means serta Apriori.

Algoritma K-Means digunakan untuk melakukan segmentasi pola perjalanan penumpang. Tujuan utama dari penggunaan

algoritma ini adalah untuk mengelompokkan penumpang ke dalam beberapa *cluster* berdasarkan karakteristik perjalanan mereka, seperti durasi perjalanan (**tripDuration**), jarak perjalanan (**tripDistance**), frekuensi perjalanan (**countTrips**), nama halte awal dan akhir (**tapInStopsName** dan **tapOutStopsName**), waktu perjalanan (**timeOfDay**), serta rute yang digunakan (**corridorID**). Dengan segmentasi ini, diharapkan dapat ditemukan kelompok penumpang dengan pola mobilitas yang serupa, yang nantinya dapat menjadi dasar dalam menyusun strategi layanan yang lebih sesuai dengan karakteristik tiap kelompok.

Dalam proses penerapan algoritma K-Means, dilakukan tahapan *pre-processing* khusus untuk *clustering*, yaitu pemilihan fitur yang akan digunakan dalam pemodelan antara lain **tripDuration**, **tripDistance**, **countTrips**, **tapInStopsName**, **tapOutStopsName**, **timeOfDay**, dan **corridorID**, serta dilakukan proses normalisasi data menggunakan MinMaxScaler. Kemudian, nilai jumlah *cluster* (K) ditentukan dengan menggunakan dua metode, yaitu *Elbow Method* dan *Davies-Bouldin Index*, yang menghasilkan nilai optimal K = 4. Hasil *clustering* divisualisasikan menggunakan *Principal Component Analysis* (PCA) untuk mereduksi dimensi data dan menampilkan hasil pengelompokan dalam grafik dua dimensi.

Kemudian algoritma Apriori digunakan untuk menemukan asosiasi antar halte yang sering digunakan secara bersamaan oleh penumpang dalam satu hari. Peran utama algoritma ini adalah untuk mengeksplorasi pola hubungan antar halte berdasarkan riwayat perjalanan harian pengguna yang direpresentasikan sebagai transaksi.

Untuk mendukung analisis asosiasi, data dikonversi kembali ke bentuk kategorikal agar hasil asosiasi lebih mudah diinterpretasikan. Selanjutnya, dibuat kolom baru bernama **startRoute** yang merupakan salinan dari kolom **tapInStopsName**, serta kolom baru bernama **tapInDate** yang berasal dari waktu *tap in*. Data kemudian

dikelompokkan berdasarkan **payCardID** dan **tapInDate** untuk merepresentasikan transaksi harian. Transaksi yang berisi lebih dari satu halte unik kemudian digunakan dalam proses pencarian *frequent itemsets* dan *association rules*. Dalam analisis Apriori ini, digunakan nilai *minimum support* sebesar 0,005 (0,5%) untuk memastikan bahwa hanya kombinasi halte yang muncul cukup sering yang dianalisis lebih lanjut dan untuk menghindari pola yang terlalu jarang dan tidak signifikan. Selain itu, digunakan nilai *lift* sebesar 1 sebagai ambang minimum untuk menyaring aturan asosiasi yang memiliki korelasi positif. Hal ini bertujuan agar aturan yang dipilih adalah kombinasi halte yang lebih sering terjadi seacra bersamaan dibandingkan secara acak, sehingga pola yang ditemukan lebih relevan untuk dianalisis sebagai kebiasaan perjalanan pengguna Transjakarta.

5. *Evaluation*

Tahap evaluasi bertujuan untuk menganalisis kualitas hasil model K-Means dan Apriori. Untuk mengevaluasi hasil *clustering* K-Means, digunakan *Calinski-Harabasz Index* (CH Index). CH Index dipilih karena metrik ini dapat menghitung seberapa baik data dipisahkan antar cluster dengan mempertimbangkan rasio antara variansi antar *cluster* dan variansi dalam *cluster*. Selain itu, CH Index lebih efisien secara komputasi dibandingkan metrik lain seperti *Silhouette Score*, yang membutuhkan waktu lebih lama terutama pada dataset besar. Berdasarkan hasil perhitungan, diperoleh nilai CH Index sebesar 96.335,90, yang menunjukkan bahwa pemodelan menghasilkan pemisahan *cluster* yang baik dan struktur yang cukup kompak.

Sementara itu, hasil Apriori berhasil mengidentifikasi aturan asosiasi yang kuat antara halte-halte yang sering digunakan secara bersamaan oleh penumpang. Beberapa halte menunjukkan pola perjalanan dengan nilai *support*, *confidence*, dan *lift* yang tinggi, yang

mengindikasikan hubungan yang signifikan antar halte yang sering digunakan secara bersamaan.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, dengan judul dataset "Transjakarta - Public Transportation Transaction" yang dapat diakses melalui tautan https://www.kaggle.com/datasets/dikisahkan/transjakarta-transportation-transaction. Dataset ini dipublikasikan pada tahun 2023 dan berisi data simulasi transaksi perjalanan penumpang Transjakarta yang menggambarkan mobilitas pengguna layanan Transjakarta [48]. Dataset ini terdiri dari 182.520 baris dan 13 kolom, yang mencakup informasi terkait detail transaksi perjalanan penumpang. Kolom-kolom dalam dataset ini antara lain, **transID**, **payCardID**, **corridorID**, **corridorName**, **tapInStops**, **tapInStopsName**, **stopStartSeq**, **tapInTime**, **tapOutStops**, **tapOutStopsName**, **stopEndSeq**, **tapOutTime**, **payAmount**.

3.4 Variabel Penelitian

3.4.1 Variabel Independen

Variabel independen adalah variabel yang memengaruhi atau menjadi faktor penyebab dalam penelitian ini. Variabel ini merupakan atribut yang digunakan untuk menganalisis pola perjalanan penumpang Transjakarta menggunakan algoritma K-Means dan Apriori. Untuk analisis *clustering* menggunakan algoritma K-Means, variabel independen yang digunakan antara lain adalah durasi perjalanan (**tripDuration**), jarak tempuh perjalanan (**tripDistance**), jumlah perjalanan yang dilakukan (**countTrips**), nama halte tempat penumpang naik (**tapInStopsName**), nama halte tempat penumpang turun (**tapOutStopsName**), waktu perjalanan yang dikategorikan ke dalam beberapa bagian waktu seperti pagi, siang, sore, dan malam (**timeOfDay**), serta identitas koridor yang digunakan (**corridorID**). Sementara itu, dalam analisis asosiasi menggunakan algoritma Apriori, variabel independen yang digunakan meliputi **startRoute** yang merupakan salinan dari **tapInStopsName** untuk menunjukkan titik awal perjalanan, **tapInDate** yang merupakan tanggal perjalanan yang diperoleh dari data waktu tap in (**tapInTime**), serta

payCardID yang digunakan untuk mengelompokkan perjalanan berdasarkan pengguna dan tanggal. Variabel-variabel ini digunakan untuk mengidentifikasi pola perjalanan penumpang serta hubungan antar perjalanan pada halte yang sering digunakan secara bersamaan.

3.4.1 Variabel Dependen

Variabel dependen adalah variabel yang dipengaruhi oleh variabel independen dan menjadi hasil dari analisis yang dilakukan dalam penelitian ini. Pada proses *clustering* dengan K-Means, variabel dependen adalah hasil pengelompokan penumpang ke dalam beberapa *cluster* berdasarkan kesamaan karakteristik perjalanan mereka. Setiap *cluster* yang dihasilkan merepresentasikan kelompok penumpang dengan pola perjalanan tertentu. Sedangkan dalam penerapan algoritma Apriori, variabel dependen adalah aturan asosiasi antar halte yang menunjukkan kombinasi rute perjalanan yang sering terjadi secara bersamaan. Hasil analisis dari variabel dependen ini diharapkan dapat memberikan wawasan bagi pengelola Transjakarta dalam meningkatkan efisiensi operasional serta merancang strategi layanan yang lebih optimal sesuai dengan pola perjalanan penumpang.

3.5 Teknik Analisis Data

Dalam penelitian ini, teknik analisis data dalam penelitian ini dilakukan melalui dua pendekatan, yaitu algoritma K-Means untuk mengelompokkan penumpang berdasarkan karakteristik perjalanan seperti durasi, jarak, frekuensi, dan halte keberangkatan, serta algoritma Apriori untuk menemukan asosiasi antar halte berdasarkan kemunculan bersamaan dalam satu hari perjalanan. Sebelum analisis dilakukan, data terlebih dahulu melalui proses pembersihan dan transformasi agar sesuai dengan format input algoritma. Evaluasi terhadap hasil *clustering* dilakukan menggunakan *Calinski-Harabasz Index*. Sementara itu, evaluasi terhadap hasil Apriori dilakukan menggunakan metrik *support*, *confidence*, dan *lift*. Hasil dari penelitian ini dapat memberikan wawasan mengenai pola mobilitas penumpang yang dapat dimanfaatkan dalam optimalisasi rute dan pengelolaan armada Transjakarta.