BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

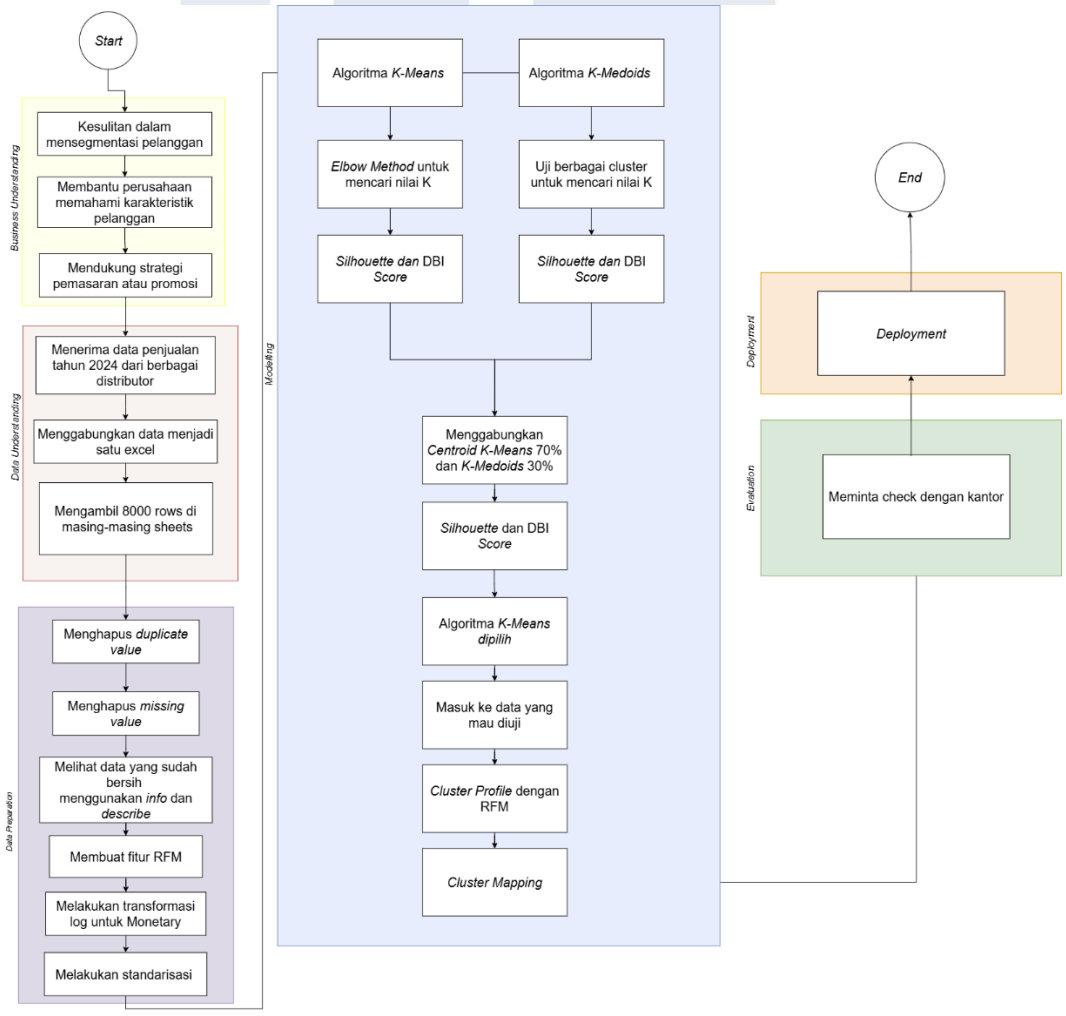
Penelitian ini dilakukan pada PT. X, sebuah perusahaan di Tangerang yang bergerak di bidang distribusi produk makanan ringan di Indonesia. Perusahaan ini menghadapi tantangan dalam mengelola data penjualan yang sangat besar dan kompleks, terutama dalam mengidentifikasi pola pembelian pelanggan. Meskipun telah mengandalkan pencatatan data penjualan, masih terdapat kendala seperti kesalahan pencatatan, ketidakakuratan informasi, serta proses pengambilan keputusan yang lambat. Fokus utama dari penelitian ini adalah melakukan segmentasi pelanggan pada PT X menggunakan model RFM (*Recency*, *Frequency*, *Monetary*) yang diintegrasikan dengan algoritma *clustering K-Means* dan *K-Medoids* untuk menganalisis perilaku serta nilai pelanggan berdasarkan data penjualan. Data yang digunakan dalam penelitian ini merupakan data penjualan produk PT X selama tahun 2024**,** yang mencakup berbagai kategori produk makanan ringan yang dipasarkan oleh perusahaan melalui jaringan distributor di beberapa wilayah Indonesia. Data tersebut berisi informasi mengenai nama pelanggan, wilayah distribusi, bulan dan tahun transaksi, serta jumlah produk yang terjual**.**

Penelitian ini akan menggunakan data transaksi penjualan yang diperoleh dari tim internal PT. X, yang mencakup informasi produk yang dibeli, jumlah pembelian, serta informasi lokasi pelanggan. Algoritma *K-Means* dan *K-Medoids* akan diterapkan untuk mengelompokkan pelanggan ke dalam beberapa segmen berdasarkan model RFM (Recency, Frequency, Monetary). Melalui penerapan kedua algoritma ini, pelanggan dengan pola perilaku pembelian yang serupa akan tergabung dalam satu kelompok (*cluster*), sehingga perusahaan dapat lebih mudah menganalisis tingkat loyalitas pelanggan, nilai transaksi, serta merancang strategi pemasaran yang lebih tepat sasaran. Output dari analisis ini akan divisualisasikan menggunakan Streamlit, yang akan memudahkan pengambilan keputusan bisnis berbasis data *real-time*. Dengan visualisasi ini,

perusahaan diharapkan dapat lebih mudah memahami pelanggan agar dapat merancang strategi pemasaran yang lebih efisien dan terarah.

3.2 Metode Penelitian

Metode penelitian ini mengacu pada kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*), yang terdiri dari enam tahap utama: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Tahapan ini dirancang untuk memastikan penerapan algoritma clustering K-Means dan K-Medoids yang efektif dalam mengelompokkan pelanggan di data penjualan dan memberikan wawasan yang dapat digunakan dalam pengambilan keputusan bisnis yang lebih akurat di PT. X.



Gambar 3.1 Alur Penelitian

1. Business Understanding

Tahap ini berfokus pada pemahaman permasalahan bisnis yang ingin diselesaikan. Permasalahan utama yang dihadapi perusahaan adalah kesulitan dalam melakukan segmentasi pelanggan berdasarkan perilaku pembelian. Segmentasi pelanggan diperlukan agar perusahaan dapat memahami karakteristik setiap pelanggan dan menyesuaikan strategi bisnis yang tepat. Tujuan utama dari tahap ini adalah membantu perusahaan dalam memahami perilaku pelanggan berdasarkan data penjualan, serta mengelompokkan pelanggan ke dalam segmen tertentu untuk mendukung strategi pemasaran dan promosi yang lebih efektif. Hasil dari tahap ini adalah penetapan tujuan analisis, yaitu melakukan segmentasi pelanggan menggunakan model RFM dan algoritma K-Means, K-Medoids untuk mengidentifikasi pola perilaku pelanggan.

2. Data Understanding

Pada tahap ini dilakukan proses pengumpulan dan eksplorasi terhadap data yang digunakan dalam penelitian. Data yang digunakan merupakan data penjualan produk tahun 2024 yang dikumpulkan dari berbagai wilayah distribusi seperti SUMBAGUT, SUMBANGSEL, BALI-NUSRA, JAWA BARAT, JABODETABEK, dan KALIMANTAN. Langkah awal yang dilakukan adalah membaca seluruh sheet dalam file tersebut, kemudian menggabungkannya menjadi satu kesatuan data menggunakan fungsi pd.concat() sehingga membentuk satu DataFrame yang disebut combined_data. Selanjutnya, penelitian ini mengambil hingga 8.000 baris data pada setiap sheet untuk memastikan jumlah data yang cukup besar dan representatif. Setelah data digabungkan, dilakukan pengecekan struktur data, jumlah baris, kolom, dan tipe data menggunakan fungsi info() dan describe() untuk memahami karakteristik data yang akan digunakan pada tahap berikutnya.

3. Data Preparation

Tahap ini bertujuan untuk mempersiapkan data agar siap digunakan dalam proses pemodelan. Data yang diperoleh masih memerlukan

pembersihan dan transformasi agar dapat digunakan secara optimal. Langkah pertama yang dilakukan adalah menghapus nilai duplikat agar tidak terjadi pengulangan data pelanggan yang sama. Selanjutnya, dilakukan penghapusan nilai kosong untuk menjaga konsistensi dan keakuratan data. Setelah proses pembersihan, dilakukan pembuatan fitur RFM (*Recency*, *Frequency*, *Monetary*) berdasarkan data transaksi pelanggan. Nilai *Recency* dihitung sebagai selisih antara bulan terakhir dan pertama pelanggan melakukan pembelian, *Frequency* dihitung sebagai jumlah bulan pelanggan aktif bertransaksi, sedangkan *Monetary* merupakan total kuantitas pembelian produk selama periode pengamatan. Agar mengurangi pengaruh *outlier* pada fitur *Monetary*, dilakukan transformasi logaritmik menggunakan fungsi log1p(). Kemudian, semua variabel numerik distandarisasi menggunakan StandardScaler agar memiliki skala yang seragam. Hasil akhir dari tahap ini adalah data yang sudah bersih, tertransformasi, dan terstandarisasi sehingga siap digunakan dalam proses pemodelan.

4. Modeling

   Tahap ini merupakan inti dari proses analisis, di mana dilakukan penerapan tiga algoritma *clustering* yaitu *K-Means, K-Medoids.* Pada algoritma K-Means, dilakukan proses penentuan jumlah cluster optimal menggunakan metode Elbow Method, kemudian hasil cluster dievaluasi menggunakan dua metrik yaitu *Silhouette Score* dan *Davies–Bouldin Index* (DBI) untuk menilai kualitas pengelompokan data. Selanjutnya, algoritma *K-Medoids* diterapkan dengan menggunakan parameter init='k-medoids++' agar pemilihan medoid awal lebih optimal dan hasil cluster lebih stabil. Proses pengujian dilakukan dengan mencoba berbagai jumlah cluster dari dua hingga lima, dan hasilnya dievaluasi menggunakan nilai *Silhouette* dan DBI *Score* untuk menentukan jumlah cluster terbaik. Berdasarkan hasil pengujian, algoritma *K-Means* menghasilkan performa terbaik dengan nilai *Silhouette Score* sebesar 0.895 dan nilai *DBI* sebesar 0.29, yang menunjukkan bahwa hasil segmentasi sudah optimal.

Selanjutnya dilakukan analisis profil klaster dengan menampilkan ringkasan statistik yaitu *Recency*, *Frequency*, *Monetary* per klaster dari *K-Means* yang telah dibuat. Hasil nya memperlihatkan rata-rata dari RFM masing-masing *cluster*. Model RFM tersebut membentuk tiga segmen pelanggan utama, yaitu *Loyal Customer*, *Reguler Customer*, dan *New Customer*. Segmen *Loyal Customer* merupakan kelompok pelanggan dengan frekuensi transaksi tinggi dan nilai pembelian besar, segmen *Reguler Customer* memiliki aktivitas pembelian yang sedang, sedangkan segmen *New Customer* merupakan pelanggan baru dengan aktivitas pembelian yang masih rendah. Setelah segmentasi terbentuk, dilakukan *Cluster Mapping* untuk mengidentifikasi jumlah pelanggan pada setiap segmen. Hasil analisis ini digunakan sebagai dasar dalam memberikan rekomendasi strategi bisnis perusahaan.

5. Evaluation

   Tahap berikutnya adalah evaluasi, yaitu proses validasi hasil segmentasi dengan pihak perusahaan atau kantor terkait. Peneliti melakukan cross-check dengan pihak perusahaan untuk memastikan bahwa hasil segmentasi yang diperoleh memang mencerminkan perilaku pelanggan sebenarnya di lapangan. Melalui proses ini, pihak perusahaan dapat memberikan masukan mengenai kesesuaian tiap segmen terhadap karakter pelanggan yang mereka amati. Apabila hasil model dianggap belum sesuai dengan kondisi bisnis, maka dilakukan penyesuaian kembali pada tahap Data Preparation atau Modeling. Dengan demikian, tahap evaluasi tidak hanya memastikan keakuratan model secara statistik, tetapi juga menjamin bahwa hasil analisis benar-benar relevan dan dapat diterapkan untuk mendukung keputusan strategis perusahaan.

6. Deployment

   Tahap terakhir dari metodologi CRISP-DM adalah *deployment* atau penerapan hasil model ke dalam bentuk nyata agar dapat digunakan oleh pihak perusahaan. Pada tahap ini, model K-Means yang telah terbukti memberikan hasil yang terbaik. Selanjutnya, hasil segmentasi pelanggan dikembangkan dan divisualisasikan di aplikasi berbasis *Streamlit* yang

menampilkan hasil segmentasi secara *real-time*. Dengan adanya implementasi ini, perusahaan dapat menggunakan hasil segmentasi pelanggan sebagai dasar pengambilan keputusan dalam strategi pemasaran, perencanaan promosi, serta pengelolaan hubungan pelanggan yang lebih efektif dan terarah.

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini menggunakan metode dokumentasi, yaitu dengan mengumpulkan data historis penjualan yang tersedia di perusahaan sebanyak 890 ribu baris di mana setiap baris merepresentasikan satu aktivitas pembelian pelanggan. Namun, penelitian ini tidak menggunakan teknik sampling statistik atau pengambilan data secara acak, melainkan menerapkan proses reduksi dan agregasi data berbasis kebutuhan analisis. Seluruh data transaksi terlebih dahulu melalui tahap data cleansing, kemudian diagregasi ke tingkat pelanggan (*customer-level aggregation*) menggunakan model RFM (*Recency*, *Frequency*, *Monetary*), sehingga setiap pelanggan direpresentasikan oleh satu baris data yang mencerminkan perilaku pembeliannya.

Melalui proses ini, jumlah data berkurang menjadi sekitar 21.000 baris, yang selanjutnya setelah dilakukan penghapusan duplikasi dan standarisasi data pelanggan menghasilkan sekitar 3.700 customer unik yang digunakan sebagai objek segmentasi. Pendekatan ini dilakukan agar analisis clustering lebih fokus pada karakteristik dan nilai pelanggan, bukan pada transaksi individual, sehingga hasil segmentasi yang diperoleh lebih relevan untuk mendukung pengambilan keputusan bisnis.

Data yang digunakan merupakan data sekunder yang diperoleh langsung dari pihak perusahaan dalam bentuk file Microsoft Excel. File tersebut berisi data transaksi penjualan produk makanan ringan selama tahun 2024, yang mencakup beberapa wilayah distribusi seperti SUMBAGUT, SUMBANGSEL, BALI-NUSRA, JAWA BARAT, JABODETABEK, dan KALIMANTAN. Data yang dikumpulkan mencakup beberapa atribut penting, yaitu wilayah, distributor, nama pelanggan, bulan transaksi, tahun transaksi, nama produk, dan jumlah

kuantitas penjualan. Seluruh data dari setiap wilayah digabungkan menjadi satu dataset utama menggunakan library Pandas di Python agar dapat diolah dan dianalisis secara keseluruhan.

Sebelum data digunakan dalam proses analisis, dilakukan tahap verifikasi dan validasi data bersama pihak perusahaan untuk memastikan keakuratan, kelengkapan, serta konsistensi data yang diterima. Jika ditemukan adanya duplikasi, nilai kosong atau data yang tidak relevan, maka dilakukan proses pembersihan agar kualitas data tetap terjaga. Dengan demikian, teknik pengumpulan data yang digunakan memastikan bahwa data yang dianalisis adalah data aktual dan valid untuk dijadikan dasar dalam proses segmentasi pelanggan menggunakan pendekatan *data mining*.

3.4 Teknik Analisis Data

Teknik analisis data pada penelitian ini dilakukan dengan menggunakan bahasa pemrograman Python karena memiliki berbagai *library* yang mendukung proses *data mining* dan analisis klaster secara komprehensif. Beberapa pustaka yang digunakan meliputi *Pandas* untuk pengolahan dan pembersihan data, *NumPy* untuk perhitungan numerik, serta *Matplotlib* dan *Seaborn* untuk visualisasi data dan hasil klasterisasi. Pustaka *Scikit-learn* digunakan untuk proses scaling data menggunakan *StandardScaler*, penerapan algoritma K-*Means*, serta evaluasi model menggunakan metrik *Silhouette Score* dan *Davies–Bouldin Index* (DBI). Selain itu, pustaka *PyClustering* dan *sklearn-extra* digunakan untuk penerapan algoritma K-Medoids sebagai pembanding terhadap hasil clustering K-Means. Pustaka *Joblib* dimanfaatkan untuk menyimpan model dan *scaler* yang telah dilatih, sementara *Streamlit* digunakan untuk membuat aplikasi visual interaktif yang menampilkan hasil segmentasi pelanggan secara dinamis. Seluruh proses analisis data dilakukan di lingkungan *Jupyter Notebook* yang memungkinkan eksekusi kode, analisis, serta dokumentasi hasil secara terstruktur. Dengan memanfaatkan berbagai pustaka tersebut, penelitian ini dapat melaksanakan seluruh tahapan analisis mulai dari pembersihan data, pembentukan fitur RFM, penerapan algoritma clustering, hingga visualisasi dan penyajian hasil segmentasi pelanggan secara interaktif.