BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Gambaran umum objek penelitian ini berfokus pada pembuatan model segmentasi pelanggan berdasarkan pola pembelian pelanggan yang didasari pada data penjualan produk dan info produk fashion yang dimiliki oleh perusahaan XYZ melalui penjualan yang dilakukan melalui *E-Commerce,* khususnya Shopee. Objek penelitian sendiri melibatkan 2 aspek utama, yakni data pembelian produk tersebut dan penerapan dari algoritma *K-Means Clustering* dan *DBSCAN* untuk membuat model segmentasi. Gambaran umum objek penelitian sendiri yang meliputi data transaksi yang memuat jenis-jenis data relevan seperti jenis produk, jumlah pembelian, tanggal transaksi, dan beberapa atribut lainnya yang terkait dalam proses transaksi pembelian barang yang dilakukan di *E-Commerce* Shopee.

Dengan menggunakan algoritma *K-Means* dan *DBSCAN*, fokus dari penelitian ini dilakukan untuk melakukan pembuatan model segmentasi pelanggan berdasarkan pola pembelian yang bertujuan untuk membantu perusahaan untuk melakukan personalisasi strategi pemasaran berdasarkan model segmen pelanggan yang dirancang agar dapat menyusun strategi pemasaran yang lebih efisien dengan tujuan umum untuk mencapai pertumbuhan bisnis yang berkelanjutan. Selain itu, proses utama yang dilakukan dalam penelitian ini yakni untuk menentukan atribut yang menjadi fokus dalam pembagian segmentasi, seperti frekuensi pembelian, jenis produk yang dibeli, sampai dengan total belanja yang dilakukan oleh konsumen.

Data yang digunakan dalam penelitian ini didapatkan dari fitur khusus yang disediakan dalam *marketplace* Shopee. Sesuai dengan objek dan batas penelitian masalah, data yang digunakan berupa data penjualan produk pada perusahaan XYZ yang memuat terdiri dari beberapa produk fashion, seperti tas, dompet, dan sepatu yang dipasarkan dalam *marketplace.* Data penjualan tersebut didapat melalui *scrapping* melalui tampilan *website* Shopee dari sisi penjual, yakni *Shopee Seller Center* untuk melakukan *generate data* yang dimuat dalam berbagai *timeline* yang

bisa diubah sesuai dengan kebutuhan user (periode hari, bulan, sampai dengan tahun).

3.2 Metode Penelitian

Adapun metode yang digunakan dalam penelitian ini yakni metode kuantitatif yang menjelaskan fenomena apa yang terjadi dari koleksi data numerik dan dianalisa kembali menggunakan metode dasar statistika. Metode penelitian kuantitatif dinilai cocok untuk diterapkan dalam topik penelitian ini karena data yang digunakan merupakan data numerikal dalam bentuk pecahan desimal maupun skala kepuasan pelanggan berdasarkan data report penjualan produk *fashion* yang di dapat dari *E-Commerce* perusahaan XYZ [63]. Secara garis besar, metode yang dilakukan dalam penelitian ini didasari oleh tahapan-tahapan yang ada dalam framework CRISP-DM. Berikut merupakan tabel perbandingan antara framework CRISP-DM dengan opsi pilihan framework lainnya:

Tabel 3. 1 Tabel perbandingan framework

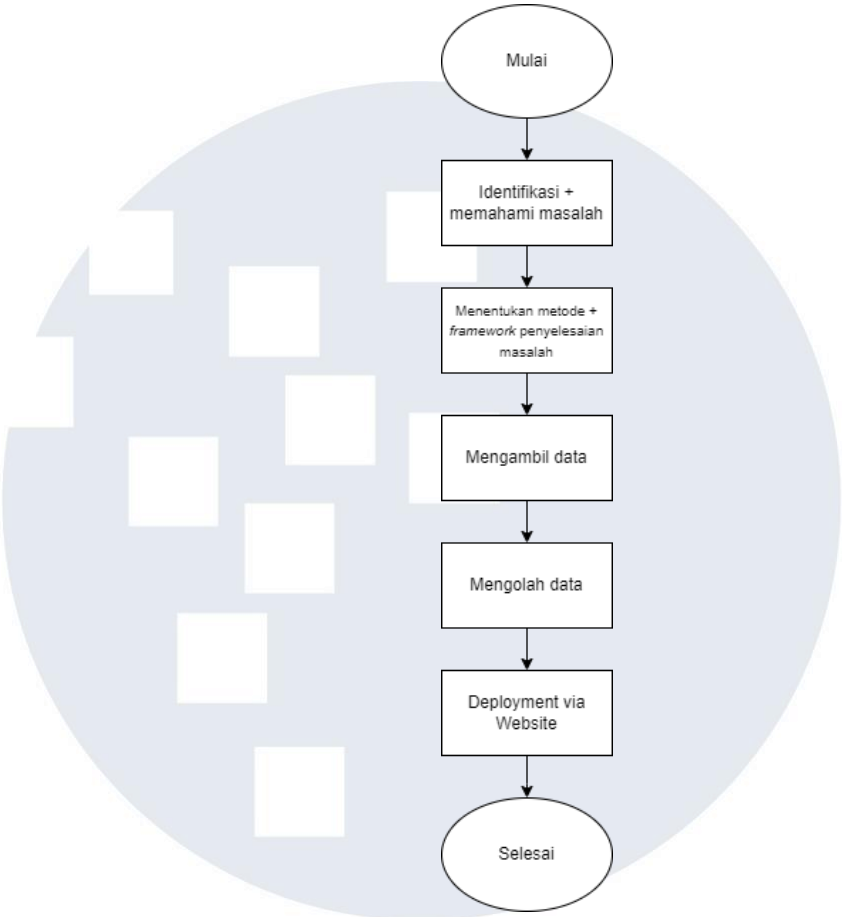
Sumber: [64] [65]

Aspek Pembeda	***CRISP-DM***	**SEMMA**	**KDD**
Tahapan Utama	Terdiri dari 6 tahapan utama (*Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment)*	Terdiri dari 5 tahapan utama (*Sample, Explore, Modify, Model, Assess)*	Terdiri dari 5 tahapan utama (*Selection, Pre Processing, Transformation, Data Mining, Interpretation/Evaluation)*
Kelebihan	*Framework* yang paling sering digunakan dalam data mining, familiar dengan penerapan industri di era saat ini	Mengidentifikasi pola-pola tersembunyi di dalam data yang bisa digunakan untuk kepentingan bisnis (menggunakan tools *SAS*	Mengaplikasikan koteks bisnis dengan sangat erat dalam proses mengidentifikasi tren pasar.

Aspek Pembeda	*CRISP-DM*	SEMMA	KDD
		Enterprise Miner)	
Contoh Penerapan	Mayoritas digunakan dalam Seluruh pengimplementasian proses *Data Mining*	Identifikasi penipuan (*fraud*), analisa portofolio, *forecasting* peluang bangkrut	*Forecasting* trend pasar, *anomaly detection*

3.2.1 Alur Penelitian

Alur penelitian menjelaskan tentang langkah-langkah yang dilakukan dalam penelitian sebagai acuan rencana yang akan dieksekusi agar proses penelitian menjadi terstruktur dan terorganisir. Alur penelitian dijadikan sebagai *roadmap* yang membantu peneliti untuk memahami langkah-langkah yang akan dilakukan secara detail untuk mencapai tujuan dari penelitian itu sendiri. Langkah pertama yang dilakukan dalam penelitian yaitu dimulai dengan memahami masalah apa yang dihadapi oleh perusahaan serta mengidentifikasi permasalahan apa yang muncul. Dalam konteks ini, masalah yang timbul dalam perusahaan yaitu perlunya optimasi strategi pemasaran *online* agar lebih akurat dan tepat sasaran segmentasi, dalam rangka untuk meningkatkan proses bisnis perusahaan dan membantu *sales* perusahaan. Langkah selanjutnya setelah mengetahui permasalahan yang timbul, yaitu mengatasi permasalahan tersebut, dalam penelitian ini memanfaatkan metode *Clustering* dengan memanfaatkan implementasi *Machine Learning* terhadap data-data yang akan diolah untuk mendapatkan i*nsight* bagi perusahaan. Algoritma yang digunakan dalam penelitian ini yaitu algoritma K-Means Clustering dan DBSCAN yang digunakan untuk membantu mengidentifikasi pola-pola yang tersembunyi dalam data sehingga dapat digunakan untuk mengidentifikasi dan merancang strategi yang optimal sesuai dengan target pasar yang sudah ditentukan.

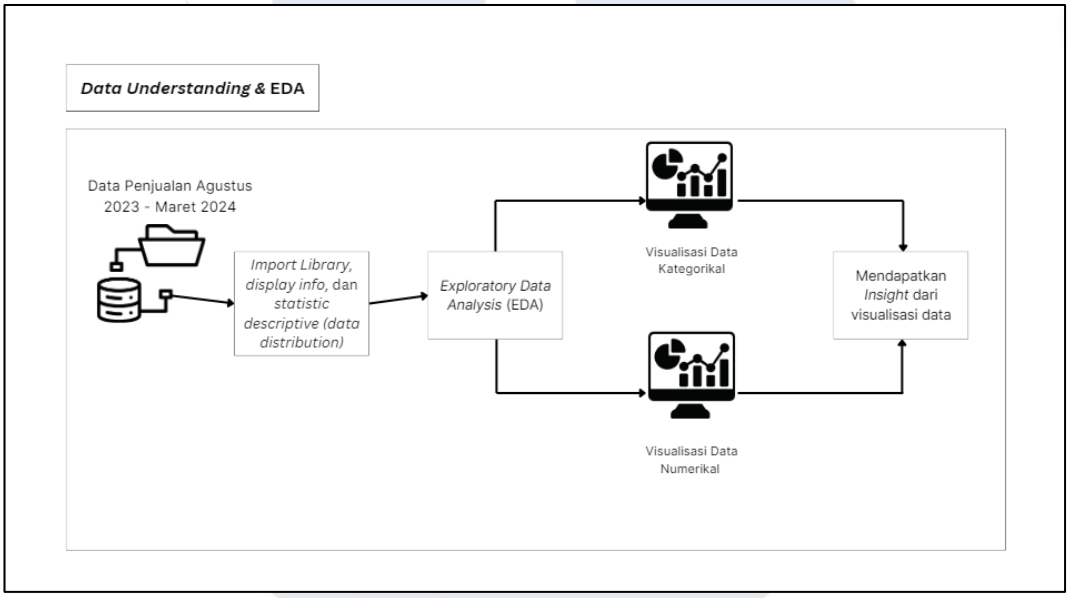


Gambar 3. 1 Alur Penelitian

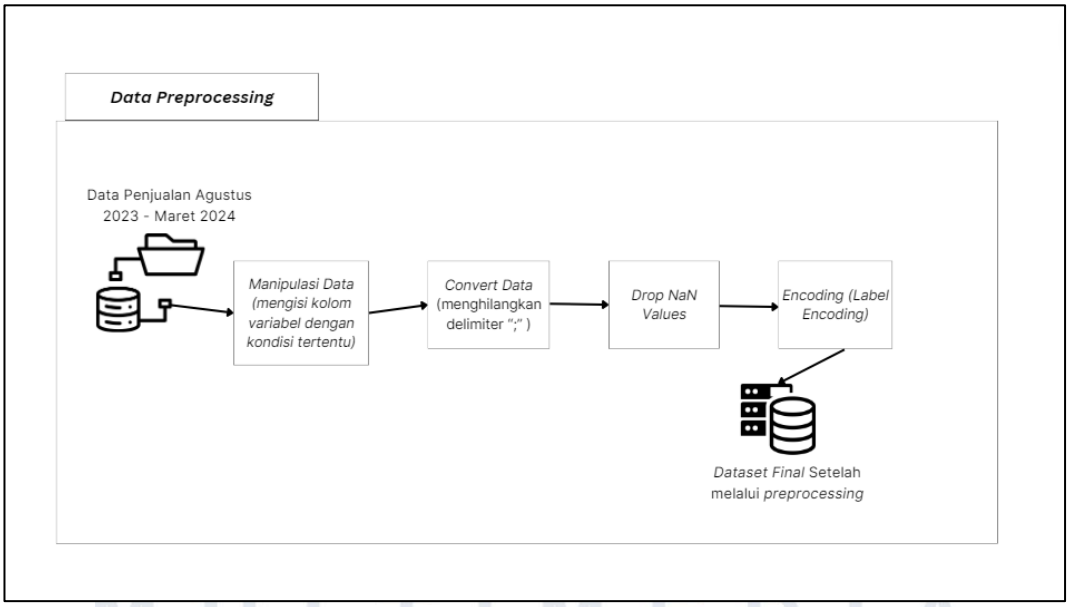
Referensi: [30]

Langkah selanjutnya yaitu dengan menentukan *framework data mining* yang akan dipilih untuk menerapkan proses pengolahan data. Penelitian ini menggunakan framework CRISP-DM (*Cross Industry Standard Process for Data Mining)* sebagai kerangka alur utama penelitian untuk melakukan implementasi model segmentasi pelanggan berdasarkan pola pembelian dengan menggunakan algoritma *K-Means Clustering* dan *DBSCAN*. Penerapan *framework* ini terbagi sampai dengan proses *deployment* sehingga hasil dari penelitian ini dibuat dalam bentuk *website* yang berisi tampilan antar muka yang diintegrasikan dengan kode *Machine Learning* untuk membentuk klaster dengan menggunakan beberapa bantuan

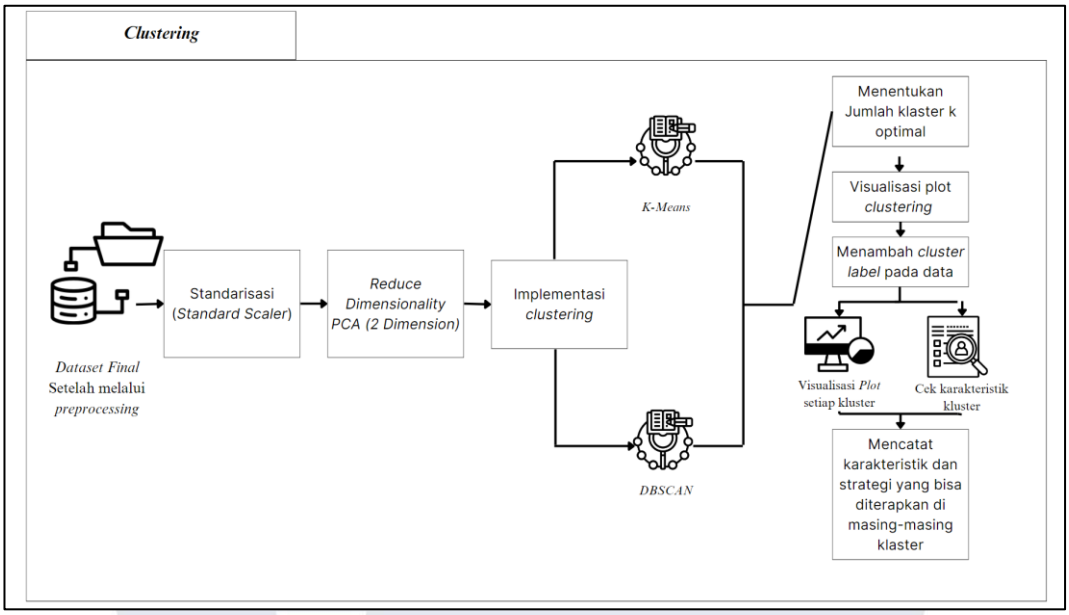
framework. Hasil dari pembentukan klaster dengan menggunakan algoritma *machine learning* tersebut akan dilakukan evaluasi dengan beberapa matriks yang sudah ditentukan dan dibandingkan dengan beberapa jurnal penelitian terdahulu.



Gambar 3. 2 Alur Proses Pengolahan Data (Data Understanding & EDA)



Gambar 3. 3 Alur Proses Pengolahan Data (Data Preprocessing)

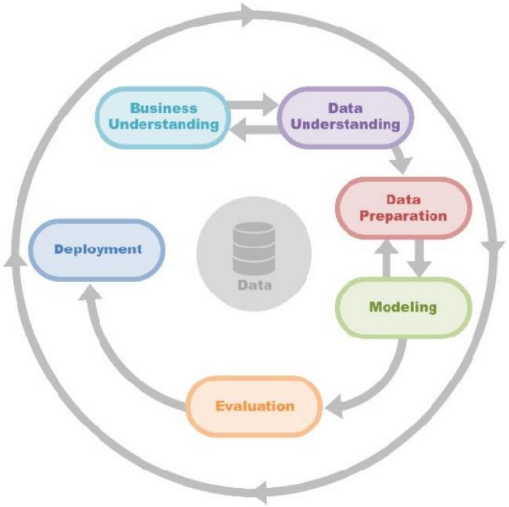


Gambar 3. 4 Alur Proses Pengolahan Data (*Clustering)*

Gambar 3.2 sampai dengan gambar 3.4 menunjukan alur proses pengolahan data sesuai dengan *framework CRISP-DM* dari proses *data understanding* sampai dengan proses *clustering*. Langkah pertama yang dilakukan yaitu dengan mengetahui karakteristik dari data sebelum melakukan pengolahan, dengan melakukan *import library*, sampai dengan menampilkan info karakteristik data secara deskriptif. Langkah selanjutnya yakni dengan melakukan *Exploratory Data Analysis* untuk mendapatkan *insight* karakteristik dari data melalui hasil visualisasi yang dilakukan. Selanjutnya seperti yang ditampilkan pada gambar 3.4, alur proses pengolahan data masuk ke dalam proses *preprocessing* untuk menyiapkan dataset akhir sebelum melakukan pengolahan data *clustering.* Dalam tahapan *preprocessing*, terdapat beberapa teknik pengolahan data seperti *encoding* dan juga melakukan manipulasi kolom variabel data serta melakukan *convert* untuk menghilangkan *delimiter* dalam kolom data numerik agar terbaca sebagai tipe data numerik (int, float) serta mempertahankan nilai asli dari data tersebut. Setelah menghasilkan dataset akhir yang sudah diolah, langkah selanjutnya yaitu dengan melakukan *clustering* dengan bantuan algoritma *K-Means Clustering* dan *DBSCAN* serta beberapa teknik pengolahan data seperti *Standard Scaler* dan *Reduce*

Dimensionality PCA. Setelah melakukan standarisasi dan *PCA*, selanjutnya menentukan jumlah klaster k optimal dengan bantuan matriks evaluasi *Silhouette Score* dan *Davies-Bouldin Index*. Hasil jumlah klaster k optimal kemudian divisualisasikan untuk melihat distribusi penyebaran klaster dan menambahkan kolom "*Cluster Label"* kedalam dataset untuk mengetahui kategori dari setiap kolom data memasuki jenis *Cluster* 0 sampai dengan 7.

3.2.2 Metode Data Mining



Gambar 3. 5 Alur Framework CRISP-DM

Sumber: [65]

CRISP-DM (*Cross Industry Standard Process for Data Mining)* merupakan salah satu *framework* dalam data mining yang digunakan untuk membantu proses *data mining* secara terstruktur dengan membagi tahapan menjadi beberapa bagian. CRISP-DM memiliki 6 langkah utama dalam menerapkan proses data mining, yakni *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [65]. CRISP-DM dimanfaatkan dalam penelitian ini untuk mendefinisikan 6 langkah utama dalam menerapkan proses data mining sebelum membuat permodelan menggunakan kedua algoritma *Clustering* tersebut agar hasil analisa penelitian menjadi lebih jelas dan terstruktur. Berikut Merupakan penjelasan mengenai tahapan utama dalam mengimplementasikan framework CRISP-DM [64]:

A. *Business Understanding*

Dalam tahap *Business Understanding*, peneliti memfokuskan pada proses pemahaman topik yang sedang diteliti dan menentukan *requirements* serta objektif-objektif apa yang ingin diselesaikan. Tujuan utama dalam penelitian ini untuk mengidentifikasi pola pembelian produk *fashion* yang terjadi di perusahaan agar penelitian ini dapat memberikan wawasan yang lebih dalam tentang preferensi dan kebiasaan belanja pelanggan, yang dapat menjadi dasar segmentasi yang lebih akurat.

B. *Data Understanding*

Tahap *data understanding* merupakan proses untuk menentukan data apa yang akan digunakan. Tidak hanya mencari data, tetapi *data-mining expert* dalam tahap ini juga memahami kekuatan dan kelemahan dari data yang akan digunakan, apakah data tersebut layak digunakan dan memuat variabel / informasi-informasi penting yang bermanfaat bagi perusahaan atau tidak. Dalam tahap *Data Understanding,* penelitian ini mengumpulkan data pembelian produk *fashion* di perusahaan serta melakukan *Exploratory Data Analysis (EDA)* untuk melakukan analisa karakteristik data seperti distribusi dan korelasi nya.

C. *Data Preparation*

Tahap *data preparation* mencakup proses persiapan data sebelum dilakukan *modeling.* Perbedaan utama proses *data understanding* dengan *data preparation* yaitu pada proses *Data Preparation* data sudah fix dan siap untuk dilakukan modelling, manipulasi, *convert*, serta melakukan *encoding* untuk mengubah data yang memiliki jenis data ‘object’ menjadi numerikal pada dataset yang didapat dari penjualan produk fashion melalui *marketplace* Shopee di perusahaan XYZ.

D. Modeling

Pada tahap ini peneliti memfokuskan *machine learning-approaches* untuk membuat model yang terbaik. Dalam proses ini peneliti menggunakan algoritma *K-Means Clustering* dan *DBSCAN*. Untuk membuat model segmentasi pelanggan berdasarkan pola pembelian konsumen. Pemilihan algoritma tersebut didasari pada penelitian-penelitian sebelumnya yang memiliki keterkaitan erat dengan topik yang dipilih serta karakteristik dari K-Means Clustering dan *DBSCAN* yang cocok digunakan untuk melakukan *clustering* dan pengelompokan data berdasarkan segmentasi. Dalam proses modeling, beberapa metode pengolahan data diterapkan seperti *standard scaler* untuk melakukan standarisasi data dan juga *PCA Reduce Dimensionality* untuk membagi data menjadi 2 dimensi utama.

E. Evaluation

Pada tahap ini, test data diperlukan untuk menentukan hasil akurasi dari model segmentasi pelanggan yang kemudian dibandingkan satu sama lain. Performa model segmentasi diukur dengan *matrix* acuan yang relevan, seperti dengan menggunakan *Silhouette Score* untuk algoritma *K-Means Clustering* dan *Davies-Bouldin Index* untuk *DBSCAN* berdasarkan dataset penjualan produk fashion pada perusahaan XYZ yang didapat melalui fitur yang tersedia di *marketplace* Shopee. Metode evaluasi yang digunakan menggunakan beberapa matriks dari algoritma *clustering*, yakni *Silhouette Score* dan *Davies-Bouldin Index* untuk mengukur seberapa baik hasil klaster yang terbentuk.

F. Deployment

Tahap *Deployment* dengan menggunakan model yang telah dibuat untuk melakukan segmentasi pelanggan diimplementasikan dalam bentuk *website* sederhana yang menjalankan kode Python sebagai

Back-End website dalam bentuk API. API yang dirancang menggunakan *framework* fastAPI untuk menjalankan kode Python di *website* yang diintegrasikan dengan ReactJS sebagai *front-end website.*

3.3 Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian skripsi ini dimulai dengan melakukan wawancara dengan pihak perusahaan XYZ untuk melakukan diskusi serta melakukan perolehan data untuk memahami kolom-kolom dan jenis variabel yang terdapat dalam dataset penjualan produk *fashion* yang dilakukan di *E-Commerce*. Proses wawancara dilakukan secara online melalui aplikasi *Zoom* pada tanggal 17 Maret 2024 dengan narasumber *Business Manager / Founder* pihak perusahaan XYZ. Sedangkan data yang akan digunakan untuk perancangan model merupakan data penjualan dengan menggunakan timeline perbulan di tahun 2023 dan 2024 (*continues data)* yang didapat dari *scraping* primer melalui fitur yang disediakan dari *marketplace* Shopee melalui halaman website dari sisi penjual perusahaan XYZ sebagai objek penelitian (Shopee Seller Center).

3.3.1 Populasi dan Sampel

Penelitian ini menggunakan teknik pengambilan sampel *Purposive Sampling* dimana setiap komponen populasi dalam data diambil berdasarkan karakteristik dan kriteria tertentu berdasarkan topik yang diteliti (segmentasi pelanggan berdasarkan pola pembelian produk *fashion*). Metode pengambilan sampel *Purposive Sampling* bertujuan untuk mendapatkan kelompok data secara spesifik yang menggambarkan pola dan hubungan entitas yang terjadi untuk merepresentasikan hasil pembuatan model segmentasi [66]. Dalam Penelitian ini memfokuskan objek penelitian berupa data penjualan produk yang diambil dari fitur dalam marketplace yang mengelompokan produknya menjadi beberapa kelompok produk *fashion* seperti beberapa jenis tas, aksesoris, sepatu, dan dompet berdasarkan pola pembelian produk di kategori *fashion* yang di lakukan di perusahaan XYZ melalui platform *marketplace*, khususnya yaitu Shopee.

Dalam proses pengambilan data dengan metode *Scrapping* primer dengan memanfaatkan fitur dalam *Shopee Seller Center,* langkah pertama yang harus dilakukan yakni dengan melakukan registrasi akun untuk mendapatkan akses terhadap fitur *Scrapping* yang disediakan oleh *Shopee Seller Center.* Setelah memiliki akun, selanjutnya user harus melakukan login dan mengakses ke menu halaman utama *Shopee Seller Center,* kemudian klik menu bar "Pesanan Saya" dan pilih *timeline* waktu pesanan dibuat, dalam penelitian ini *dataset* yang di generate menggunakan *timeline* 1 bulan dalam bentuk format excel yang memuat data penjualan selama *timeline* yang dipilih. Adapun jumlah karakteristik dari masing-masing dataset perbulan akan ditampilkan pada rincian bab 4.

3.3.2 Periode Pengambilan Data

Periode Pengambilan data yang dilakukan dalam penelitian ini berdasarkan pengambilan dataset penjualan yang dilakukan sejak bulan Agustus 2023 – Maret 2024. Alasan utama periode dataset yang diambil sampai dengan Maret 2024 yakni data yang digunakan dalam penelitian ini bersifat *Continue* dan terus berlanjut, oleh karena itu menggunakan periode dataset terakhir yang lengkap selama *timeline* 1 bulan, yakni bulan maret 2024 agar penelitian ini bersifat *Up to date* dan menghasilkan *clustering* yang relevan dengan sesuai dengan kondisi terdekat, mengingat model segmentasi selesai dibuat pada bulan April 2024 sesuai dengan periode waktu yang penelitian yang sudah ditentukan.

3.4 Teknik Analisis Data

Penelitian ini menggunakan algoritma *clustering K-Means Clustering* dan algoritma *Density-Based Spatial Clustering of Applications with Noise (DBSCAN).* Kedua algoritma tersebut ditujukan untuk melakukan *clustering* dengan tujuan untuk membagi data menjadi beberapa kelompok berdasarkan persamaan-persamaan tertentu [67]. Kedua algoritma ini juga ditujukan untuk menemukan pola / struktur data yang tidak bisa dilihat dalam tampilan data secara langsung tanpa melalui proses pengolahan data dari kedua model tersebut akan dibandingkan dan peneliti akan menentukan model manakah yang menghasilkan model segmentasi

pelanggan berdasarkan pola pembelian produk *fashion* terbaik. Berikut merupakan tabel perbandingan dari kedua algoritma yang digunakan dalam penelitian ini

Tabel 3. 2 Tabel perbandingan algoritma

Sumber: [67] [68] [69]

Algoritma	Kegunaan	Kelebihan	Kekurangan
K-Means Clustering	*K-Means Clustering* digunakan untuk membagi data ke dalam kelompok berdasarkan jarak ke pusat *cluster* (*centroid)* untuk melakukan analisis pola, serta membagi data-data menjadi segmentasi tertentu berdasarkan jumlah klaster optimal yang ditentukan melalui matriks evaluasi.	Kelebihan utama *K-Means* adalah kemudahannya dalam melakukan implementasi dan skalabilitas tinggi untuk menangani jumlah dataset yang besar.	*K-Means* sensitif terhadap posisi awal *centroid* atau pusat *cluster.* Selain itu algoritma ini juga kurang cocok untuk diterapkan pada *cluster* dalam bentuk yang beragam.
DBSCAN	*DBSCAN* digunakan untuk mengidentifikasi *cluster* dengan bentuk dan ukuran yang beragam, serta berorientasi terhadap kepadatan data untuk mengatasi *noise* dalam data	Kelebihan utama yaitu adalah kemampuannya untuk menangani cluster dengan bentuk dan ukuran yang beragam tanpa memerlukan jumlah cluster sebagai parameter. Selain itu, *DBSCAN* dapat menangani noise dalam data dengan baik.	*DBSCAN* kurang efektif untuk data dengan dimensi yang tinggi dan kompleks. Selain itu, algoritma ini sensitif terhadap parameter jarak yang digunakan untuk mendefinisikan *cluster*, sehingga dapat mempengaruhi hasil *clustering*.
Gaussian Mixture Model (GMM)	*GMM* digunakan untuk membagi data ke dalam kelompok berdasarkan distribusi statistik (dengan asumsi data berasal dari beberapa distribusi gaussian yang berbeda)	Kelebihan utama GMM yaitu kemampuannya untuk menangani *cluster overlap* dam memberikan probabilitas pada setiap titik data	*GMM* kurang cocok untuk diimplementasikan dalam penerapan *clustering* sederhana karena memakan *resource* yang besar jika dibandingkan dengan ke 2 algo diatas, serta sensitif terhadap inisialisasi parameter (kurang cocok untuk melakukan segmentasi)

Tabel 3. 3 Tabel perbandingan tools yang akan digunakan

Sumber: [70] [71]

Tools	Kelebihan	Kekurangan
Jupyter Notebook	Bersifat *Multi-Language Interactive Computing*, karena *support* lebih dari 40 bahasa pemrograman. Jupyter Notebook cenderung tidak memakan memori komputer dalam jumlah yang besar dan memungkinkan user untuk melakukan *export* terhadap hasil pekerjaan mereka dalam beberapa format, seperti PDF, HTML, sampai dengan JSON sehingga bersifat fleksibel.	Tidak ada integrasi dengan *Integrated Development Environment (*IDE), serta tidak ada fitur visualisasi yang optimal.
Rstudio	Bersifat *Open-Source* dan menyediakan *packages* dalam jumlah yang banyak (lebih dari 10.000 packages didalam repository CRAN) serta fitur yang bervariatif untuk melakukan operasi *machine learning*.	Memakan Memori perangkat desktop dengan cukup besar, serta tampilan dari *user interface* dalam Rstudio cenderung bersifat kuno dan tidak menarik.
Google Colaboratory	Menyediakan *development environment* berbasis cloud yang memungkinkan pengguna untuk menjalankan notebook Python tanpa perlu menginstal apapun. Mendukung koneksi langsung ke Google Drive dan integrasi dengan TensorFlow.	Keterbatasan sumber daya komputasi karena menggunakan environment developer lokal serta memiliki ketergantungan pada koneksi internet untuk penggunaan secara optimal.

Berdasarkan tabel perbandingan mengenai pilihan tools yang tersedia untuk diterapkan dalam penelitian ini, peneliti memutuskan menggunakan bantuan tools Jupyter Notebook dengan bahasa pemrograman Python untuk melakukan pengolahan sampai dengan pembuatan model. Kelebihan utama dari Jupyter Notebook yaitu memuat library yang bisa menghasilkan output dari code seperti HTML, gambar, video. LaTex, sampai dengan format MIME custom. Jupyter Notebook dijalankan dalam *environment local* sehingga privasi dan *credential* dari data yag digunakan lebih terjaga jika dibandingkan dengan Google Colab yang memiliki basis *Cloud* untuk menjalankan Python.