BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terdahulu digunakan sebagai tinjauan dan acuan yang relevan dengan topik penelitian yang dibahas. Penelitian terdahulu yang dijadikan sebagai acuan landasan teoritis berkaitan dengan topik penerapan pengolahan data dengan metode *Clustering*, secara spesifik dengan menggunakan algoritma *K-Means* dan *DBSCAN*, seperti yang ditampilkan pada tabel 2.1.

Tabel 2. 1 Tabel Perbandingan Penelitian Terdahulu

Artikel Jurnal	Nama Jurnal	Penulis / Tahun	Metode	Hasil	Kesimpulan
"Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm" [20]	Scientific Journal of Informatics	Alamsyah, P. Eko Prasetyo, Danang Dwi Saputro, Siti Harnina Bintari / 2022	*Data Collection, Recency Frequency Monetary Modeling, Data Standardization, Elbow Method, K-Means Cluster Algorithm, Data Visualization, Classification, Comparation Metrics*	Nilai *Cluster* dihasilkan dengan mengimplementasikan *Elbow Method* setelah dievaluasi dengan menggunakan *Calinski Harabasz Index* dengan jumlah *cluster* menjadi 3 bagian. Metode ini menghasilkan 3 segmen pelanggan yang lebih akurat dan optimal yang ditunjukan melalui hasil perhitungan nilai SSE dan CHI yang telah dilakukan	Penerapan integrasi model *RFM* dan *K-Means Clustering* yang dioptimalkan dengan *Elbow Method* dapat digunakan untuk melakukan segmentasi pelanggan dalam perusahaan ritel dalam menentukan nilai k *cluster* yang paling optimal dalam pengelompokan
"Product recommendation for e-commerce business by applying	Innovations in Systems and Software Engineering	Soma Bandyopadhyay, Subro Santiranjan Thakur, Jyotsna	*Principal Component Analysis (pca), K-Means, Clusterin*	Terbentuk jumlah klaster sebanyak K = 4 dari kasih evaluasi *elbow method*, Melakukan	Tujuan utama dari penelitian ini adalah untuk membagi pengguna menjadi

Artikel Jurnal	Nama Jurnal	Penulis / Tahun	Metode	Hasil	Kesimpulan
principal component analysis (PCA) and K-means clustering: benefit for the society" [21]		Kumar Mandal	*g, Mean Squared Error, Scaling Data, Elbow Method*	clustering menggunakan *PCA* berdasarkan jumlah *sales revenue*	kelompok-kelompok kecil, yang dapat dilihat sebagai kelompok pengguna yang menyukai pembelian jenis pakaian yang sama. Segmentasi Produk ditujukan untuk membantu pelaku bisnis untuk mempertahankan stok yang optimal
"Customer Segmentation of Shopping Mall Users Using K-Means Clustering"[22]	Advancing SMEs Toward E-Commerce Policies for Sustainability	Amit Kumar / 2022	*Segmentation Process Design and Business Understanding, Understanding, Preparing and Enriching Data, Cluster Modeling for Segment Identification, Analyzing and Profiling the Revealed Segments*	Jumlah *Cluster Optimum* yang diperlukan untuk pendapatan tahunan menggunakan nilai grafik siku (grafik yang menggambarkan jumlah *cluster* terhadap variasi data). Jumlah *cluster optimum* yang dipilih adalah value 5. Titik biru pada gambar 11 menunjukan sentroid dari masing-masing *cluster.*	Dalam *cluster* berdasarkan usia dan *spending score*, *cluster* berwarna pink dianggap sebagai orang yang lebih muda dan *spending score* yang tinggi dan *cluster* berwarna biru dan hijau mengindikasikan kelompok data yang memiliki *spending score* yang rendah, sehingga pusat perbelanjaan dapat memberikan inovasi berupa penawaran menarik melalui kartu

Artikel Jurnal	Nama Jurnal	Penulis / Tahun	Metode	Hasil	Kesimpulan
					membership terhadap *cluster* yang memiliki *spending score* yang rendah.
"*Customer Segmentation: Transformation from Data to Marketing Strategy* [23]	IAIC International Conference Series Vol. 4, No. 1, 2023, pp. 139 ~ 152	Luciana Abednego, Cecilia Esti Nugraheni, Adelia Salsabina / 2023	*Data Cleaning, Data Scaling, Exploratory Data Analysis (EDA), RFM Model, Principal Component Analysis (PCA), Clustering, Segment Interpretation, Marketing Strategy*	DBSCAN memiliki *Silhouette Index* tertinggi dan *Davies Bouldin Index (DBI)* Terendah sehingga menunjukan hasil *clustering* yang lebih baik dalam hal kohesi dibandingkan dengan algoritma *K-Means Clustering* (0.2996 SHI INDEX, 1.19 DBI INDEX untuk algoritma *K-Means)*, sedangkan untuk DBSCAN, *Silhouette index* sebesar 1.19 dan DBI sebesar 1.00	Algoritma *K-Means Clustering* dan *DBSCAN* cocok untuk digunakan pada studi kasus dimana jumlah *cluster* tidak diketahui dan memiliki bentuk yang tidak teratur dan perlu mengidentifikasi *noise* ataupun *outliers.*
"K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data"[24]	Journal of Sustainability MIS Department	Kayalvily Tabianan , Shubashini Velu, Vinayakumar Ravi / 2022	*K-Means Clustering, Data Mining, Data Visualization, Dashboard, Time Dashboard*	Dari *clustering* yang dilakukan menghasilkan 3 jenis *Cluster* utama, yakni jenis acara, product, dan juga kategori.Kategori produk yang memiliki nilai minat tinggi di masyarakat menghasilkan	Metode penelitian berbasis *deep learning* memiliki hasil akhir kinerja yang lebih baik dibandingkan dengan *machine learning* dengan menerapkan algoritma

Artikel Jurnal	Nama Jurnal	Penulis / Tahun	Metode	Hasil	Kesimpulan
				keuntungan yang tinggi dan sebaliknya.	*clustering K-Means* dan pengelompokan SAPK
"Internet Service Provider User Customer Lifetime Segmentation Analysis using RFM and K-Means Algorithm" [25]	Sinkron: Jurnal dan Penelitian Teknik dan Informatika	Muhammad Febrian Rachmadhan Amri , Mohamad Hafidhul Umam , Arief Wibowo, I Made Satrya Ramayu / 2024	*Collecting Data, Transformation & Preprocessing Data, Recency, Frequency, Monetary (RFM) implementation, K-Means Clustering Method*	Berdasarkan variabel dataset, terdapat 104 jenis pelanggan di sektor ritel yang dibagi menjadi 4 segmen, yakni kelas platinum, emas, perak, dan juga perunggu. Dari hasil analisis RFM pada dataset, dapat dilihat bahwa produk yang memiliki jumlah permintaan pasar paling banyak tidak merepresentasikan jumlah pendapatan terbesar.	Segmentasi *Retail & Distribution Services* (RDS) menghasilkan 2 output dengan hasil bahwa *cluster* segmen pelanggan emas dengan atribut moneter (tagihan dengan nilai yang cukup besar) merupakan jumlah pelanggan terbanyak dan pelanggan dalam *cluster* segmentasi tersebut bersedia membayar lebih untuk mendapatkan layanan yang lebih baik.
"IMPLEMENTATION OF THE DBSCAN METHOD FOR CLUSTER MAPPING OF EARTHQUAKE SPREAD LOCATION " [26]	BAREKENG JURNAL ILMU MATEMATIKA DAN TERAPAN	Muhammad Bariklana, Achmad Fauzan / 2023	*Data Processing, Data Exploration, Descriptive Analysis, Clustering, DBSCAN Algorithm, Cluster Evaluation*	Hasil *cluster* terbaik yang diperoleh dari penelitian memiliki data yang cenderung sama dengan peta bahaya gempa dari BPBD provinsi Jawa Barat pada tahun 2020	Hasil penelitian ini membentuk 12 *cluster* yang diperoleh dengan algoritma *DBSCAN* dengan hasil evaluasi sebesar 0.713 (*silhouette coefficient)* yang mengartikan bahwa setiap *cluster*

Artikel Jurnal	Nama Jurnal	Penulis / Tahun	Metode	Hasil	Kesimpulan
					memiliki karakteristik gempa bumi tersendiri. Hal ini bisa dijadikan sebagai bentuk mitigasi bencana gempa bumi agar dapat meminimalkan kerugian yang mungkin terjadi.
"*Customer Segmentation Based on Loyalty Level Using K-Means and LRFM Feature Selection in Retail Online Store"* [27]	Jurnal ELTIKOM : Jurnal Teknik Elektro, Teknologi Informasi dan Komputer	Tiara Lailatul Nikmah, Nur Hazimah Syani Harahap, Gina Cahya Utami, Muhammad Mirza Razzaq	*Preprocessing, LRFM Model, Data Training, Data Test, Elbow Method, K-Means Clustering Algorithm*	Dari hasil penelitian ini terdapat 9 *cluster* yang diuji dengan rentang nilai *possible* dengan menggunakan *Elbow Method.* dari pengaplikasian metode LRFM untuk melakukan pemilihan fitur untuk mengukur tingkat loyalitas pada pelanggan yang dinilai potensial dengan dibagi menjadi 4 segmen, yakni premium, inersia, laten, dan *no-loyality.*	Menghasilkan *Silhouette Score Index* sebesar 0.9438 dimana hasilnya mendekati 1 yang mengindikasikan hasil *clustering* menjadi semakin presisi dan akurat sehingga memungkinkan bagi para pelaku bisnis untuk memberikan prioritas layanan mereka sesuai dengan hasil segmentasi
"The Performance Comparison of DBSCAN and K-Means Clustering for MSMES Grouping Based on Asset Value	Journal of Information Systems Engineering and Business Intelligence	Ni Putu Sutramiani , I Made Teguh Arthana , Pramayota Fane'a Lampung , Shana	*DBSCAN , K-Means Clustering,Scrapping Data, MSME, Davies Bouldin*	Melakukan perbandingan performa algoritma *clustering* aset berdasarkan value MSME dengan menggunakan	Hasil dari penelitian ini disimpulkan bahwa *K-Means* memiliki kemampuan untuk menunjukan

Artikel Jurnal	Nama Jurnal	Penulis / Tahun	Metode	Hasil	Kesimpulan
and Turnover" [28]		Aurelia , Muhammad Fauzi , I Wayan Agus Surya Darma	*Index (DBI)*	algoritma *DBSCAN* dan *K-Means*, dengan hasil skor DBI sebesar 0,39 pada *K-Means* (8 klaster)dan sebesar 1.39 pada *DBSCAN* (5 klaster).	pemisahan data yang lebih baik dengan mengklasifikasikan data menjadi kelompok yang relevan (dengan skor matriks DBI yang lebih rendah).
"Rancang Bangun Model Personalized Learning dengan Algoritma Decision Tree dan Random Forest pada Perusahaan Telekomunikasi " [29]	INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION	Alexander Bryan Wiratman	*CRISP-DM, Data Preparation, Decision Tree, Random Forest*	Model Personalized learning yang dibuat menghasilkan akurasi sebesar 69% untuk algoritma *decision trees* dan 70% untuk algoritma *random forests.* Selain itu juga dibuat dashboard visualisasi data dengan menggunakan tableau untuk merepresentasikan *data* yang diolah.	Penelitian ini berkontribusi untuk membantu perusahaan telekomunikasi untuk memprediksi kebutuhan pelatihan karyawan dengan memanfaatkan penggunaan algoritma *machine learning* dan *dashboard* visualisasi untuk memahami data pelatihan karyawan
"Implementasi Content-based Image Retrieval dalam Pemberian Rekomendasi Produk Fashion XYZ Berbasis Web" [30]	Bachelor Thesis, Universitas Multimedia Nusantara.	Gladys Patricia	*Content-Based Image Retrieval , CRISP-DM*	Hasil dari *Image prediction* yang diimplementasikan dalam bentuk *deployment* melalui *website* sederhana memiliki tingkat akurasi yang lebih baik dibandingkan *image recommendation* di *website* aslinya.	CIBR merupakan metode yang efektif untuk melakukan rekomendasi produk yang lebih akurat berbasis gambar dibandingkan dengan yang diterapkan di *website e-commerce* perusahaan XYZ.

Berdasarkan Hasil dari perbandingan penelitian jurnal ilmiah terdahulu yang berkaitan dengan topik penelitian yang akan dilakukan, dapat disimpulkan terdapat beberapa poin mendukung untuk melakukan penelitian ilmiah yang dengan Menggunakan Algoritma K-Means Clustering dan DBSCAN, salah satu faktor utama yang memungkinkan rancang bangun model ini dapat dilakukan yakni dengan pemilihan Algoritma dalam *machine learning* untuk melakukan rancang bangun model segmentasi tersebut, yakni dengan menggunakan algoritma *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. Kedua algoritma ini juga ditujukan untuk menemukan pola / struktur data yang tidak bisa dilihat dalam tampilan data secara langsung tanpa melalui proses pengolahan data. Perbedaan utama yang membedakan algoritma *clustering K-Means* dengan algoritma *DBSCAN* yakni berdasarkan prinsip pengelompokan dimana *K-Means* melakukan pengelompokan berdasarkan jarak terhadap pusat *cluster* atau *centroid*, sedangkan *DBSCAN* melakukan pengelompokan berdasarkan kerapatan dalam data dan *noise* dalam bentuk *cluster* yang tidak teratur.

Berdasarkan penelitian ilmiah terdahulu seperti yang ditampilkan dalam tabel 2.1, sebagian besar objek penelitian segmentasi yang dilakukan dalam artikel jurnal penelitian tersebut menggunakan bantuan algoritma *clustering* seperti *K-Means Clustering* dan juga *DBSCAN*. Pengelompokan segmentasi tersebut dimasukan ke dalam beberapa kategori berdasarkan perilaku yang terjadi, seperti RFM (*Recency, Frequency, and Monetary)*. Hasil segmentasi tersebut memberikan informasi berupa *insight*, dan *dashboard* yang bisa digunakan oleh para pelaku bisnis dan tujuan penelitian untuk mengoptimasi hal yang bisa dibenahi dalam bentuk *deployment* algoritma dengan tampilan *website* sederhana.

Secara keseluruhan, keterhubungan antar artikel jurnal terdahulu berhubungan erat dengan topik tugas akhir dalam penelitian ini, yakni secara garis besar menerapkah penggunaan algoritma *machine learning* yang sama untuk melakukan *clustering*, yakni menggukaan algoritma *K-Means Clustering* dan juga

DBSCAN. Selain mengadopsi pemilihan algoritma dari artikel jurnal terdahulu, penelitian ini juga mengadopsi penggunaan metode pengolahan *data mining* yang sama dengan referensi artikel jurnal terdahulu, yakni penggunaan *framework CRISP-DM* dari tahap *Business Understanding* sampai dengan proses *Deployment.*

Dalam segi aspek kebaruan penelitian, beberapa jurnal yang dijadikan sebagai referensi penelitian terdahulu, belum ada penelitian yang melakukan spesifikasi segmentasi pelanggan berdasarkan pola pembelian dalam industri *Fashion* dengan spesifik yang memanfaatkan data transaksi produk yang didapatkan dari *marketplace* shopee, khususnya dengan metode scraping primer yang disediakan pada halaman Shopee Seller Center yang menggunakan pemanfaatan metode alur *framework CRISP-DM* dalam mengolah data transaksi *marketplace* shopee. Selain itu, penelitian ini juga menerapkan beberapa metode tambahan dalam melakukan pengolahan data, seperti proses *encoding*, lebih tepatnya *label encoding* yang bertujuan untuk mengubah variabel kolom yang mengandung data kategorikal (dalam bentuk objek) menjadi data numerik (dalam bentuk int/float). Berbeda dengan acuan artikel jurnal terdahulu yang menggunakan dataset secara final yang mengandung kolom numerik dalam setiap variabel kolom dari dataset yang digunakan. Tambahan metode lainnya yakni standarisasi data dengan menggunakan *standard scaler* agar hasil dari *clustering* menjadi lebih konsisten dengan membuat sifat statistik pada data menjadi konsisten.

2.2 Tinjauan Teori

2.2.1 *E-Commerce*

E-Commerce merupakan bentuk kegiatan bisnis yang melibatkan aktivitas penjualan / pembelian produk maupun jasa secara *online* yang saling terkoneksi melalui jaringan internet [31]. *E-Commerce* sendiri mencakup beberapa jenis transaksi, seperti pembelian barang/jasa secara online, pembayaran tagihan dan kebutuhan rumah, sampai dengan proses *transfer* dana *e-wallet* dan pertukaran informasi bisnis. Munculnya *E-Commerce* ini tidak hanya menciptakan pergeseran dalam paradigma konsumen, tetapi juga semakin membuka pintu perdagangan menjadi

semakin luas bagi para pelaku bisnis untuk memasarkan produk dan jasa mereka secara global[32]. *E-Commerce* memiliki beberapa karakteristik utama yang membedakan dari bisnis konvensional, seperti *global reach*¸ dan mengutamakan efisiensi operasional yang memungkinkan para pelaku bisnis untuk mengoptimalkan interaksi jual-beli secara *online* tanpa perlu menyediakan media / tempat secara offline sehingga dapat meningkatkan efisiensi dana operasional [33].

Di negara Indonesia sendiri, perkembangan *E-Commerce* terus mengalami pertumbuhan yang cukup pesat. Seperti yang dilampirkan oleh google, sektor *E-Commerce* merupakan penyumbang nilai transaksi kotor dalam perkembangan ekonomi digital di Indonesia (USD$ 82 Miliar). Beberapa sektor industri terpengaruh secara signifikan akibat berkembangnya *platform E-Commerce* di Indonesia, mengingat meningkatnya aksesibilitas pelanggan terhadap berbagai produk dalam kategori-kategori tertentu tanpa adanya batasan geografis [34].

2.2.2 *Marketplace*

Marketplace merupakan platform digital yang merupakan bagian dari *E-Commerce* yang menyediakan berbagai fasilitas bagi pihak penjual yang dapat diakses kepada calon pembeli secara *online*. *Marketplace* memungkinkan kedua belah pihak baik dari segi penjual maupun pembeli untuk terkoneksi secara langsung tanpa bertemu secara fisik. *Marketplace* yang beredar di seluruh dunia umumnya tidak hanya menyediakan platform jual beli antara penjual dan calon pembeli saja, melainkan menyediakan beberapa fitur pembayaran seperti pembayaran tagihan, pengelolaan inventaris, sampai dengan layanan pembelian tiket transportasi hingga voucher game online[35].

Korelasi pertumbuhan marketplace di Indonesia sendiri menggambarkan kondisi dari perkembangan E-Commerce yang terjadi. Seperti yang dijelaskan dalam bagian E-Commerce menyumbang peranan penting dalam perekonomian di Indonesia. Berdasarkan data yang

ditampilkan dalam similiarweb.com yang merupakan perusahaan traffic tracker di amerika serikat, marketplace yang paling sering dikunjungi pada tahun 2023 di Indonesia adalah Shopee pada peringkat pertama, disusul dengan Marketplace lokal asal Indonesia, yakni Tokopedia[36].

Marketplace sendiri dapat dimanfaatkan oleh para pelaku bisnis dalam bidang fashion untuk mendorong inovasi model bisnis dengan memanfaatkan akses pasar yang sangat luas dibandingkan sebelumnya. Dengan memanfaatkan marketplace, para pelaku bisnis fashion dapat menjangkau konsumen di seluruh Indonesia tanpa perlu menginvestasikan dalam bentuk model bisnis berupa toko fisik. Pemanfaatan strategi dalam munculnya marketplace juga mengubah tren pasar sehingga pelaku bisnis harus beradaptasi dengan strategi pemasaran yang akan diterapkan[37].

2.2.3 *Fashion*

Fashion diartikan sebagai cerminan ekspresi dari budaya, serta identitas pribadi secara *personal* yang direpresentasikan dalam bentuk pemilihan pakaian, aksesoris, dan barang yang menunjang *style* dalam berpakaian [38]. *Fashion* digunakan sebagai sarana untuk mengekspresikan seseorang mengenai wujud dari perubahan budaya dan sosial yang ada di masyarakat. *Medium fashion* digunakan sebagai bentuk ekspresi diri dalam mengikuti perkembangan *tren*d yang sedang terjadi[39]. Masing-masing pilihan dari pakaian atau aksesoris yang dikenakan menciptakan narasi unik yang membangun citra diri masing-masing di mata orang lain. Sebagai refleksi dari perubahan sosial dan budaya, *fashion* dianggap sebagai simbol yang membentuk suatu narasi yang direpresentasikan oleh pemakai, hal ini tidak terbatas pada pakaian saja, termasuk ke dalam aksesoris sampai dengan gaya rambut yang dikenakan untuk menyampaikan pesan visual dan juga artistik[40].

Industri *fashion* di Indonesia sendiri mengalami transformasi signifikan seiring dengan perkembangan teknologi. Digitalisasi dan perkembangan *E-Commerce* di Indonesia memperluas peluang para pelaku

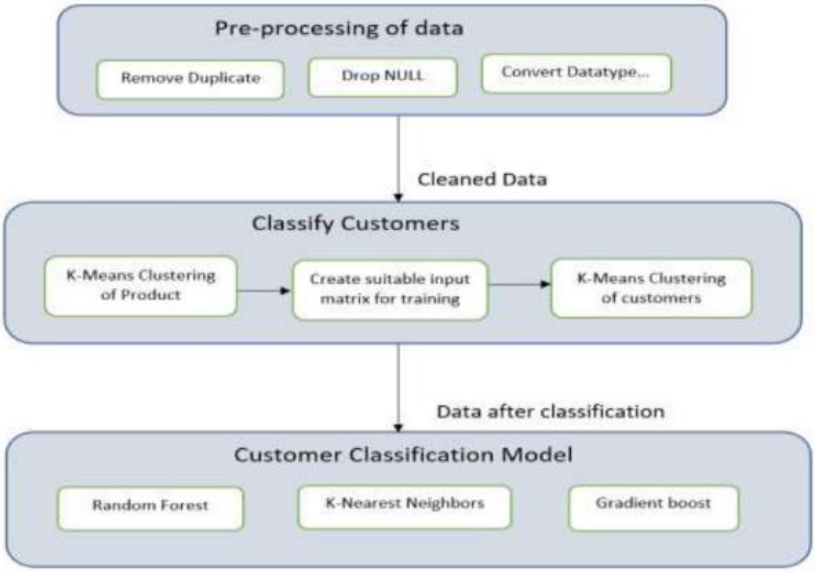
bisnis *fashion* untuk melakukan interaksi dengan produk-produk yang dijual secara *online* [41]. Menurut laporan yang dilampirkan oleh *Business of Fashion (*BOF) pada tahun 2021, perkembangan *E-Commerce* memperpendek siklus produksi dan distribusi sehingga memungkinkan para pelaku bisnis dalam industri *fashion* untuk menjadi lebih responsif terhadap permintaan selera perubahan yang terjadi dalam pasar untuk mengikuti perubahan selera konsumen [42]. Melalui perkembangan teknologi dalam bentuk analisis data dan *artificial intelligence*, platform *E-Commerce* dapat membantu memberikan rekomendasi produk yang disesuaikan dengan preferensi pelanggan sehingga membentuk strategi pemasaran yang lebih personal antara suatu *brand* dengan pelanggan [43].

2.2.4 Segmentasi Pelanggan

Segmentasi Pelanggan merupakan salah satu metode pendekatan untuk mengenal secara kompleks dalam rangka untuk memahami kebutuhan konsumen serta memahami karakteristik dari sekelompok konsumen dengan tujuan utama untuk mengembangkan skala bisnis dengan memberikan strategi pemasaran dan penawaran produk yang lebih spesifik dan relevan [44] . Dalam penerapan nya di *E-Commerce*, Segmentasi pelanggan memiliki peranan penting untuk menyediakan pengalaman belanja yang lebih bersifat *personal* sesuai dengan kebutuhan konsumen sehingga dapat meningkatkan retensi pelanggan. Segmentasi pelanggan merupakan salah satu strategi yang efektif dalam menghadapi persaingan pasar di *e-commerce* yang kompleks. Dalam perspektif *marketing*, segmentasi pelanggan meliputi proses untuk membagi target pasar menjadi kelompok kecil yang dikategorikan berdasarkan karakteristik, perilaku, maupun preferensi pelanggan [45].

Segmentasi pelanggan sendiri bisa dilakukan dengan melakukan analisis perilaku pembelian, dalam studi kasus *E-Commerce* agar dapat meningkatkan kepuasan belanja dan menyajikan produk serta menargetkan iklan sesuai dengan preferensi dan karakteristik setiap segmen [46]. Fungsi

segmentasi mencakup personalisasi pengalaman pelanggan dalam melakukan transaksi belanja, optimasi strategi pemasaran sesuai dengan *audience* yang sesuai (dalam studi kasus ini yaitu industri *fashion)*, dan peningkatan retensi pelanggan yang memainkan peran krusial dalam mencapai kesuksesan dalam proses bisnis yang dilakukan di *E-Commerce* [47].



Gambar 2. 1 Sistem Arsitektur untuk melakukan Customer Segmentation

Sumber: [48]

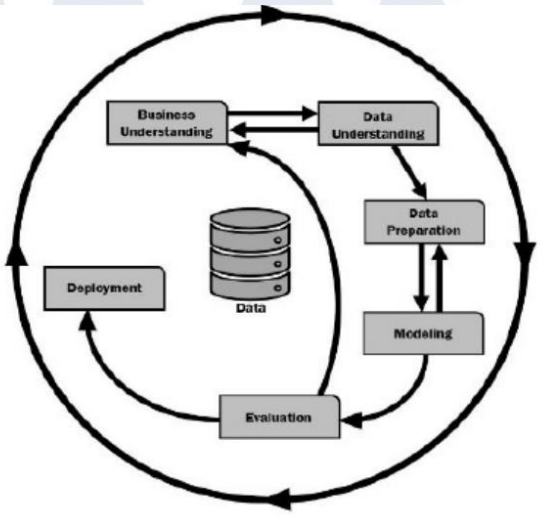
Gambar 2.1 menjelaskan mengenai sistem arsitektur untuk melakukan *Customer Segmentation* dalam proses bisnis di *E-Commerce*. Proses segmentasi dimulai dengan pembersihan data yang melibatkan penghapusan duplikat dan baris yang kosong. Dalam tahap ini, juga dilakukan penghapusan pesanan yang dibatalkan, dan tipe data kolom dikonversi sesuai dengan yang diperlukan untuk pra-pemrosesan. Data yang telah didefinisikan kemudian digunakan untuk mengelompokkan pelanggan ke dalam kategori yang berbeda melalui pembelajaran tanpa pengawasan menggunakan algoritma K-Means (contoh kasus pada jurnal penelitian tersebut)[48].

Sistem arsitektur yang dilampirkan pada gambar 2.1 menggambarkan langkah-langkah yang diambil untuk membersihkan data, mengelompokkan pelanggan menggunakan algoritma K-Means, dan melatih serta menguji model terawasi untuk meningkatkan segmentasi pelanggan. Proses ini memungkinkan identifikasi kelompok pelanggan yang lebih baik untuk pengambilan keputusan yang lebih efektif dalam konteks *e-commerce* [48]

2.3 Framework, Algoritma, dan Metode Evaluasi

2.3.1 Framework

CRISP – DM (Cross Industry Standard Process – Data Mining) merupakan salah satu metodologi data mining yang dipakai ketika perusahaan ingin mengimplementasikan proses data mining untuk menunjang keperluan industri yang dilakukan oleh data mining expert. CRISP-DM pertama kali dikemukakan oleh 3 perusahaan yang bergerak dalam bidang data-mining pada tahun 1996, yaitu Daimier-Benz, Integral Solution Ltd (ISL) yang kemudian mengganti nama menjadi SPSS dan perusahaan terakhir yaitu NCR, salah satu perusahaan konsultan spesialis Data Mining [49].



Gambar 2. 2 Framework CRISP-DM

Sumber: [49]

Berikut merupakan penjelasan mengenai tahapan utama dalam implementasi CRISP-DM [49]:

A. *Business Understanding*

Business Understanding merupakan langkah pertama dalam proses CRISP-DM yang merujuk pada langkah awal yang dilakukan oleh perusahaan untuk menentukan tujuan serta menentukan permasalahan utama apa yang akan di selesaikan.

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. Tahapan ini merupakan tahapan yang cukup memakan waktu dan biaya bagi perusahaan dalam proses implementasi CRISP-DM.

C. *Data Preparation*

Tahap *data preparation* mencakup proses persiapan data sebelum dilakukan *modelling.* Perbedaan utama proses *data understanding* dengan *data preparation* yaitu pada proses *Data Preparation* data sudah fix dan siap untuk dilakukan modelling, manipulasi, convert, serta melihat *Missing Value* pada data.

D. *Modeling*

Tahap *modeling* tidak hanya memanfaatkan algoritma yang ada, tetapi membuat kombinasi menjadi model prediksi yang lebih baik. Penerapan *modeling* di implementasikan dengan membentuk model-model prediksi dengan bantuan algoritma yang dipilih.

E. *Evaluation*

Tahap *Evaluation* mencakup evaluasi hasil *modelling* yang dilakukan dalam proses sebelumnya untuk memastikan kembali apakah model tersebut valid sebelum diimplementasikan ke dalam proses *deployment*.

F. Deployment

Tahap *Deployment* merupakan proses penerapan algoritma dan model yang sudah ditentukan secara final sebagai machine learning yang bisa digunakan oleh pihak ke 3 untuk kepentingan bisnis.

SEMMA adalah salah satu *framework* dari *data mining* yang merupakan kependekan dari *Sample, Explore, Modify, Model, dan Assess*, yang dikembangkan oleh SAS Institute. Proses SEMMA dimulai dengan pengambilan sampel data (*Sample*), yang kemudian dieksplorasi (*Explore*) untuk menemukan pola dan anomali dalam dataset. Selanjutnya, data dimodifikasi untuk memastikan kualitas dan relevansi, diikuti dengan pembangunan model berdasarkan data yang telah diolah dengan tahap akhir untuk menilai (Assess) model yang telah dibangun untuk mengukur seberapa baik kinerja model tersebut. SEMMA dirancang untuk memberikan kerangka kerja yang sistematis dalam pengolahan d*ata mining* yang berfokus pada implementasi teknik statistik dan analisis data [50].

Namun, jika dibandingkan dengan CRISP-DM (*Cross Industry Standard Process for Data Mining*), *SEMMA* memiliki beberapa kekurangan karena CRISP-DM lebih komprehensif dan fleksibel karena mencakup tahapan bisnis sampai dengan proses *deployment*. CRISP-DM dimulai dengan pemahaman bisnis dan data, lalu dilanjutkan dengan persiapan data, pemodelan, evaluasi, dan deployment, yang membuat framework CRISP-DM lebih cocok untuk berbagai jenis proyek data mining yang membutuhkan pendekatan yang lebih mendalam tentang proses bisnis yang terjadi. Secara garis besar, SEMMA lebih teknis dan berfokus pada analisis data statistik, sedangkan CRISP-DM menawarkan

framework yang lebih lengkap untuk menggabungkan hasil pengolahan data dengan kebutuhan bisnis yang terjadi di lapangan.

KDD, atau *Knowledge Discovery in Databases*, adalah metodologi / / *framework data mining* yang terdiri dari beberapa tahap utama, yakni *Selection, Preprocessing, Transformation, Data Mining,* dan *Interpretation/Evaluation.* Proses KDD dimulai dengan pemilihan data yang relevan (*Selection*), diikuti dengan *Preprocessing* data (untuk membersihkan dan melakukan *filter* pada data yang akan ditransformasikan ke dalam format yang sesuai untuk analisis. Tahap berikutnya adalah *data mining* yang memanfaatkan algoritma *machine learning* untuk menemukan pola dan *insight* baru dalam data. Tahap terakhir adalah interpretasi dan evaluasi (Interpretation/Evaluation), yang bertujuan untuk menginterpretasikan hasil dan menilai kualitas penemuan *insight* baru tersebut [51].

CRISP-DM menawarkan kerangka kerja yang lebih terstruktur dan lebih mengorientasikan dalam sisi bisnis, dengan tahapan yang mencakup pemahaman bisnis dan deployment, yang dimana tidak dijelaskan secara detail dalam *framework* KDD. Sementara KDD lebih fokus pada aspek teknis dan penemuan pengetahuan dalam data, sedangkan CRISP-DM menawarkan pendekatan yang lebih menyeluruh dan adaptif untuk memastikan bahwa hasil analisis *data mining* memberikan *value* nyata dan *insight* yang dapat diaplikasikan dalam konteks bisnis.

2.3.2 Metode & Algoritma

2.3.2.1 *K-Means Clustering*

K-Means Clustering merupakan salah satu algoritma dalam *machine learning* yang ditujukan untuk melakukan pengelompokan data menjadi *cluster* (kelompok) berdasarkan kemiripan pada pola perilaku tertentu yang dikumpulkan [24]. Hasil dari algoritma *K-Means clustering* biasanya digunakan dalam bentuk pengelompokan seperti segmentasi terhadap struktur-struktur data untuk

memberikan wawasan dan *insight* dalam memahami struktur dan karakteristik dataset tersebut untuk mendukung pengambilan keputusan bisnis maupun perencanaan strategi. Contoh penerapan algoritma *K-Means Clustering* yaitu dalam bentuk segmentasi pelanggan sampai dengan analisis pola dalam dataset yang sulit ditemukan apabila tidak melakukan *clustering* /pengelompokan .[52]

$$J(C_k) = \sum_{k=1}^{K} \sum_{k} \|X_i \quad -\mu_k\|^2$$

Rumus 2. 1 Sum of Squared Errors

Sumber: [22]

Keterangan:

- k = Jumlah klaster
- Xi = Jarak titik data ke i
- μ_k = Pusat klaster k

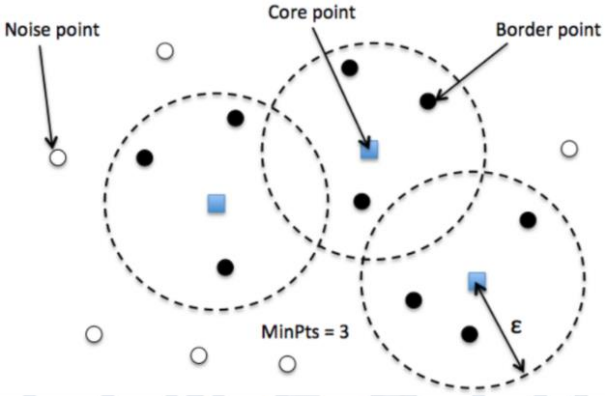
Dalam mengolah data dengan menggunakan algoritma K-Means Clustering, tujuan utama yang ingin dicapai yakni dengan meminimalisir jumlah kesalahan kuadrat yang didapat dari perhitungan perbedaan antara setiap titik data dan pusat kelompoknya, seperti yang ditampilkan dalam Rumus 2.1 yang menjelaskan mengenai *Sum Of Squared Errors (SSE)*. Secara garis besar, algoritma *K-Means Clustering* dijalankan dengan langkah sebagai berikut:

- Menginisialisasi jumlah centroid secara acak dalam domain data
- Mengelompokkan data menjadi variabel “K” kelompok dengan menetapkan setiap titik data ke *centroid* berdasarkan jarak

- Menghitung rata-rata dari semua objek dalam setiap kelompok dan melakukan pemindahan *centroid* ke posisi tersebut

2.3.2.2 *DBSCAN*

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) merupakan salah satu algoritma *density-based clustering* dalam *machine learning* yang digunakan untuk melakukan *clustering* yang difokuskan pada kepadatan jumlah data. Algoritma *DBSCAN* melakukan pengelompokan data-data dalam representasi titik-titik yang dekat 1 sama lain berdasarkan kepadatan dalam daerah. Titik-titik data yang tidak termasuk dalam kelompok (*Cluster)* dianggap sebagai "*Noise*" karena tidak dapat dimasukan ke dalam *cluster* yang didasari dari batasan kepadatan dan jarak oleh parameter epsilon (*ε)* dan minimum *data points* yang dijadikan sebagai 2 input parameter utama. Salah satu kelebihan dari Algoritma *DBSCAN*, peneliti tidak perlu mendefinisikan jumlah dari *clusters* "K" secara spesifik dalam proses pengelolaan data, peneliti hanya perlu untuk melakukan perhitungan jarak antara *value* dalam dataset.



Gambar 2. 3 DBSCAN Algorithm Point

Sumber: [23]

Algoritma *DBSCAN* mengklasifikasi titik-titik sebagai *core point* untuk membentuk *cluster*, *border points* yang berada dalam jarak epsilon dari *core point* atau *noise* maupun *outlier points* seperti yang ditampilkan dalam gambar 2.2.2 [23].

2.3.2.3 *Standard Scaler*

Standard Scaler merupakan salah satu teknik standarisasi data yang diterapkan dalam *pre-processing* data sebelum menerapkan algoritma *Machine Learning* ke dalam dataset. *Standard Scaler* berfungsi untuk melakukan standarisasi fitur-fitur numerik yang ada dalam dataset dan menghilangkan efek skala dari variabel-variabel dalam data, sehingga dapat dipastikan bahwa semua variabel memiliki kontribusi yang seimbang dalam memenuhi kebutuhan analisis yang dilakukan oleh algoritma-algoritma *clustering* yang bersifat sensitif terhadap fitur data, seperti *K-Means* dan *DBSCAN*.

$$z = \frac{x - \mu}{\sigma}$$

Rumus 2. 2 Standard Scaler Formula

Sumber: [53]

Keterangan:

- Z = Matriks data yang distandarisasi
- X = Matriks data sebelum distandarisasi (data asli)
- μ = Mean vektor setiap fitur data
- σ = Vektor simpangan baku setiap fitur data

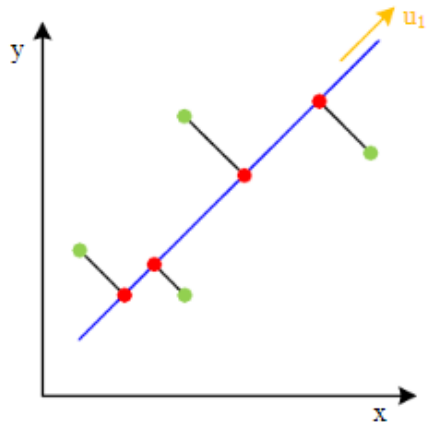
Dengan menerapkan *standard scaler*, proses *clustering* akan menghasilkan klaster yang tidak terpengaruh oleh perbedaan skala antar fitur, sehingga menjadi lebih konsisten dan akurat. Pada umumnya, hasil standarisasi yang dilakukan tidak langsung

diterapkan ke algoritma *clustering* seperti *K-Means* dan *Dbscan* langsung terhadap data yang diproses, jika jumlah datasetnya kompleks dan besar, maka langkah selanjutnya yakni dengan menerapkan metode PCA untuk melakukan *Reduce Dimensionality* pada data sebelum melakukan *clustering*. [53]

2.3.2.4 *PCA Reduce Dimensionality*

PCA (*Principal Component Analysis)* merupakan teknik yang digunakan untuk melakukan reduksi dimensi yang ada dalam dataset sebelum melakukan permodelan dengan menggunakan *machine learning*. Tujuan utama dilakukan teknik PCA ini yaitu menyederhanakan kompleksitas data, dengan cara memproyeksikan data yang diambil dari dimensi yang cenderung tinggi, ke dimensi yang lebih rendah.

Penerapan PCA tetap mempertahankan informasi-informasi bawaan dalam dataset dan membaginya menjadi *N Components* yang bisa ditentukan sendiri oleh user. PCA sangat berguna untuk mengatasi masalah dimensi yang tinggi dalam dataset dengan cara mengurangi variabel-variabel kolom dalam dataset yang tidak relevan / bersifat redundan, sehingga dapat menghasilkan *clustering* yang lebih efektif



Gambar 2. 4 Orthogonal Projection PCA

Sumber: [54]

Sebelum mengimplementasikan metode PCA untuk mengurangi dimensional data, implementasi PCA memerlukan langkah-langkah Standarisasi data, seperti *Standard Scaler* sehingga data diproses agar memiliki nilai simpangan baku satu dan rata-rata nol, karena PCA sendiri bersifat sangat sensitif terhadap variabel data. Setelah itu, penentuan jumlah komponen utama juga diperlukan dan disesuaikan dengan tujuan dan analisa data yang dihasilkan, pada umumnya *N Components* yang biasanya dipilih yakni 2 / 3. Hal ini juga dipertimbangkan sesuai dengan kompleksitas data dan kapasitas memori perangkat yang digunakan. PCA meningkatkan interpretasi pada dataset serta dalam waktu yang sama meminimalisir *information loss* dalam data. [54]

2.3.3 Metode Evaluasi

2.3.3.1 *Silhouette Score*

Silhouette Score merupakan salah satu matriks evaluasi yang digunakan dalam algoritma *clustering*, seperti algoritma *K-Means* dan *DBSCAN*. Matriks evaluasi *Silhouette Score* menghitung seberapa baik setiap *sample* data yang ditempatkan dalam klaster tertentu dan mengukur seberapa jauh jarak titik tersebut dari klaster lainnya.

$$silhouette\ (O_i) = \frac{b(O_i) - a(O_i)}{\max\{a(O_i), b(O_i)\}}$$

Rumus 2. 3 Silhouette Score Formula

Sumber: [55]

Keterangan:

- S(Oi) = Nilai *Silhouette Score* untuk titik data i

- b(Oi) = Mean / rata-rata jarak sampel (yang direpresentasikan sebagai titik) I dengan sampel dalam klaster yang sama
- a(Oi) = Mean / rata-rata jarak sampel (yang direpresentasikan sebagai titik) I dengan sampel dalam klaster lain yang terdekat dengan klaster tersebut.

Silhouette Score menggunakan indeks nilai diantara -1 sampai dengan 1, dimana skor yang tinggi (mendekati 1) menunjukan bahwa hasil sampel data (yang direpresentasikan dalam titik-titik) data ditempatkan dengan baik dalam klaster masing-masing, dan sebaliknya. Jika hasil indeks nilai yang dihasilkan rendah, hal itu menunjukan bahwa titik-titik sampel data ditempatkan di klaster yang salah sehingga hasil *clustering* yang dilakukan kurang baik.

Fungsi dari matriks evaluasi *Silhouette Score* ini yaitu untuk menentukan seberapa banyak jumlah klaster k optimal yang terbaik sebelum melakukan algoritma *clustering*, agar dapat meningkatkan pemahaman dan melakukan visualisasi *clustering* yang akurat. [55]

2.3.3.2 *Davies-Bouldin Index (*DBI*)*

$$\text{DBI} = \left(\frac{1}{\text{n}}\right) x \sum max \frac{(R_i + R_j)}{(d(C_i, C_j))}$$

Rumus 2. 4 Davies Bouldin Index Formula

Sumber: [28]

Keterangan:

- n/k = Jumlah klaster

- Ri = Rata-rata jarak dari titik klaster ke dalam pusat klaster i
- d(Ci, Cj) = Jarak antara dua pusat klaster
- maxj≠i(d(ci,cj)σi+σj) = Hasil indeks jarak yang terbentuk, semakin kecil nilainya maka semakin baik hasil *Clustering* yang dihasilkan.

Salah satu indeks matriks evaluasi yang digunakan dalam algoritma *clustering K-Means*, yaitu j dems *Davies-Bouldin Index* (*DBI*). Indeks *DBI* sendiri dimanfaatkan untuk melakukan pengukuran terhadap kualitas *partition* dari sebuah klaster yang bersifat *homogen* (*Inter-Cluster Dissimilarity)*.

Berbeda dengan cara kerja matriks evaluasi *Silhouette Score,* indek skor nilai *DBI* yang semakin rendah, maka menunjukan semakin baik partisi dari hasil klaster tersebut sehingga jumlah klaster optimal dapat ditentukan dari indeks *DBI* yang paling rendah. Adapun rumus matematis dari *Davies-Bouldin Index* ditampilkan pada rumus 2.3.

Rumus 2.3 menampilkan rumus dari indek *DBI*, dimana k menunjukan jumlah klaster yang terbentuk, sedangkan d(ci,cj) menunjukan nilai jarak pusat klaster I dan j, yang bisa ditentukan dengan memanfaatkan berbagai matriks pengukur jarak, seperti *euclidean distance* maupun *manhattan distance* [28].

2.4 Tools yang digunakan

2.4.1 Jupyter Notebook

Jupiter Notebook merupakan aplikasi berbasis *server-client* yang memungkinkan pengguna untuk mengedit dan menjalankan code pemrograman via halaman website. Jupyter Notebook support lebih dari 40 bahasa pemrograman. Seperti Python, R, Julia, dan Scala. Selain itu kelebihan utama dari Jupiter Notebook yaitu kita bisa menghasilkan

output dari code pemrograman dengan hasil yang interaktif dan beragam, seperti HTML, gambar, video. LaTex, sampai dengan format MIME custom[56] .

Dalam menampilkan dan menjalankan dokumen notebook, Jupyter Notebook memiliki tampilan Dashboard yang merepresentasikan *Control Panel* yang menunjukan *local* files dari perangkat desktop kita dengan bantuan kernel [57]. *Notebook* Kernel sendiri merupakan bentuk dari komputasi mesin yang bertugas untuk mengeksekusi kode pemrograman yang diketik dalam Notebook Document.

2.4.2 Visual Studio Code

Visual Studio Code merupakan salah satu *platform software development* berupa IDE yang paling terkenal dan banyak digunakan untuk melakukan pengembangan *software* dan melakukan beberapa keperluan pemrograman lainnya. Visual Studio Code diciptakan oleh perusahaan Microsoft dan memungkinkan user untuk melakukan *code writing* dengan banyak sekali ekstensi yang disediakan oleh komunitas untuk mendukung produktivitas dalam mengembangkan *software.*

Kelebihan utama Visual Studio Code sendiri dibandingkan dengan *platform* IDE lainnya yakni fleksibilitas yang sangat tinggi, karena dapat menjalankan banyak sekali bahasa pemrograman dan memanfaatkan ekstensi yang membantu dalam pengembangan *software.* Selain fleksibel, desain *user interface* dari Visual Studio Code sendiri tergolong simpel dan mudah dimengerti dan cenderung ringan dan tidak memakan banyak memori komputer. [58]

2.4.3 Python

Python adalah bahasa pemrograman utama yang digunakan dalam tools *Jupyter Notebook* maupun Google Colaboratory. Python sendiri dapat digunakan dalam proses *development website,* sampai dengan proses data *analysis* dan pengaplikasian metode *machine*

learning, sampai dengan *Artificial Intelligence*. Salah satu kelebihan utama dari bahasa pemrograman ini yaitu menggunakan jumlah *syntax code* yang cenderung singkat dan sedikit dan menggunakan kosa kata yang mudah dipahami dan mirip dengan penggunaan bahasa sehari-hari [59]. Selain itu, dukungan dari komunitas yang besar juga memudahkan *developer* untuk menggunakan syntax bahasa pemrograman Python dalam proyek masing-masing .

Python sering digunakan dalam implementasi *data analysis* maupun algoritma-algoritma dalam *machine learning.* Selain memiliki komunitas yang besar, Python juga memiliki *library external* yang cukup luas yang dapat digunakan untuk membantu implementasi *machine learning* seperti *scikit-learn, tensorflow*. Selain itu Python juga bersifat *open source* sehingga tidak membutuhkan pungutan biaya bagi para *developer* dan bisa dijalankan dalam beberapa platform *operating system* seperti Windows, Mac, sampai dengan linux [60].

2.4.4 JavaScript

JavaScript merupakan bahasa pemrograman yang diterapkan dalam pengembangan *Software*, khususnya dalam pengembangan *website*. Kelebihan utama bahasa JavaScript sendiri memiliki fleksibilitas yang tinggi dan memanfaatkan konsep *reusable component* sehingga tidak perlu berulang-ulang mendefinisikan satu fungsi yang sama. JavaScript banyak digunakan dalam pengembangan *website* khususnya dalam pengembangan *front-end* dengan memanfaatkan beberapa *framework* yang cukup familiar, seperti ReactJS.

JavaScript sendiri menggunakan bantuan *runtime* dari Node.Js untuk membangun *environment website* yang dijalankan dari perangkat komputer masing-masing. JavaScript menggunakan konsep *redux, useEffect* dan pengoptimalan *State* dalam konsep *Syntax code* nya. Selain itu, kelebihan dari JavaScript sendiri didukung oleh beberapa *library* yang cukup lengkap untuk membantu *Software Developer* dalam

mengembangkan tampilan *website,* seperti MaterialUI yang menyediakan beberapa komponen utama yang bersifat *Re-Usable.* [61]

2.4.5 FastAPI

FastAPI sendiri merupakan *framework* yang digunakan untuk merancang pembentukan *Application Programming Interface* dari bahasa pemrograman Python. Dari hasil pembentukan *clustering* model segmentasi pelanggan dengan memanfaatkan *Machine Learning* lewat bahasa pemrograman Python, hasil data yang didapatkan akan ditampilkan dengan bantuan pembentukan API melalui *framework* ini.

Keunggulan utama dari *framework* ini yaitu dapat menangani *request* dari *front-end website* dalam jumlah yang kompleks dan berat serta mampu menghasilkan dokumentasi API berdasarkan skema yang didefinisikan oleh *user* sehingga dapat menggunakan *endpoint* API yang sudah disediakan. Hasil API yang dibentuk dari *framework* ini digunakan untuk menampilkan data berupa *object* JSON ke bagian *website deployment* yang dibuat menggunakan JavaScript, khususnya *framework* React.js.[62]