

## BAB III

### METODOLOGI PENELITIAN

#### 3.1 Gambaran Umum Objek Penelitian

Objek penelitian ini adalah data pelanggan sebuah pusat kebugaran (*gym*) yang identitasnya disamarkan untuk menjaga kerahasiaan. Data tersebut berasal dari sistem internal *gym* yang mencatat secara terstruktur informasi keanggotaan, pembelian layanan *personal trainer*, status pembekuan *membership*, dan pola kunjungan pelanggan.

Data yang dianalisis mencakup periode Januari 2022 hingga Maret 2025. Penelitian ini bertujuan untuk melakukan segmentasi pelanggan berdasarkan karakteristik dan pola aktivitas menggunakan algoritma *K-Means Clustering*, serta mengembangkan model prediksi potensi pembelian layanan *personal trainer* dengan algoritma *XGBoost*.

Dalam prosesnya, penelitian mengikuti kerangka kerja CRISP-DM yang meliputi pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi hasil analisis. Hasil dari segmentasi akan digunakan sebagai fitur tambahan dalam proses prediksi untuk meningkatkan akurasi model.

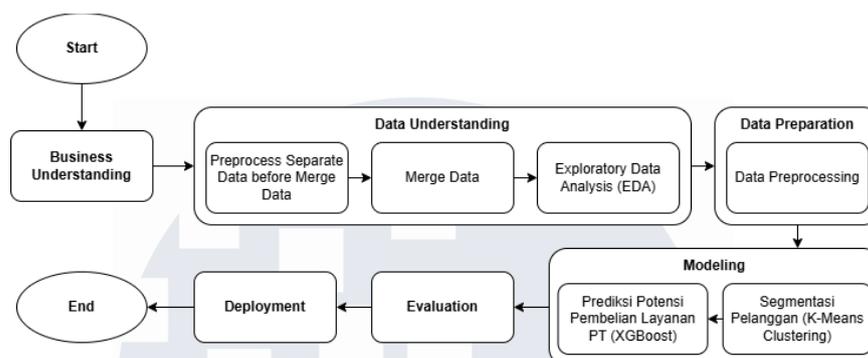
Penelitian ini memiliki batasan bahwa data hanya berasal dari satu *gym* tanpa melibatkan data eksternal, dan analisis terbatas pada variabel yang tersedia dalam sistem internal tanpa memasukkan faktor psikografis atau pengaruh eksternal lainnya.

#### 3.2 Metode Penelitian

Penelitian ini menggunakan alur dan metode penelitian dengan pendekatan metode data *mining* untuk menganalisis dan memodelkan data pelanggan *gym* dengan tujuan segmentasi dan prediksi pembelian layanan *personal trainer*. Metode ini dipilih karena kemampuan data mining dalam menggali pola dan hubungan dalam data besar secara efektif serta aplikatif untuk kebutuhan bisnis.

### 3.2.1 Alur Penelitian

Alur penelitian mengikuti proses standar dalam data *mining* yang diadaptasi dari kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Berikut adalah diagram alur penelitian yang menggambarkan proses tersebut:



Gambar 3. 1 Alur Penelitian

Gambar 3.1 menunjukkan diagram alur penelitian dengan divisualisasikan dalam bentuk *flowchart* yang menjelaskan tentang proses-proses yang dilakukan pada penelitian. Alur penelitian ini menggambarkan tahapan yang dilakukan dalam penelitian dari awal hingga akhir.

Proses dimulai dengan Pemahaman Bisnis, dimana rumusan masalah dan tujuan penelitian ditetapkan sesuai kebutuhan manajemen *gym*. Selanjutnya, pada tahap Pemahaman Data, data pelanggan dikumpulkan dan dieksplorasi untuk memahami karakteristik dan kualitasnya. Tahap Persiapan Data meliputi pembersihan, transformasi, dan penggabungan data agar siap digunakan dalam analisis.

Pada tahap Pemodelan, algoritma *K-Means Clustering* digunakan untuk mengelompokkan pelanggan berdasarkan pola dan karakteristik mereka, kemudian hasil segmentasi ini menjadi fitur tambahan dalam model prediksi menggunakan algoritma *XGBoost*. Tahap Evaluasi bertujuan untuk mengukur performa model dengan metrik yang relevan guna memastikan hasil yang valid dan dapat diandalkan. Terakhir, pada tahap Implementasi, hasil analisis disajikan sebagai rekomendasi strategis yang dapat digunakan manajemen *gym* dalam pengambilan keputusan.

## 2.2.2 Metode Data Mining

### 3.2.2.1 CRISP-DM

Dalam penelitian ini, metode *data mining* digunakan untuk mengolah data pelanggan guna menghasilkan segmentasi dan prediksi potensi pembelian layanan PT. Terdapat beberapa metode umum dalam *data mining*, yaitu CRISP-DM, KDD, dan SEMMA. Tabel berikut menunjukkan perbandingan ketiga metode tersebut:

Tabel 3 1 Tabel Perbandingan Metode *Data Mining*

Sumber: Schröer et al. (2020), Omari (2023), Noor et al. (2023) dan SAS (2023) [28] [50] [51]

Metode	Tahapan	Kelebihan	Kekurangan
CRISP-DM	<ol style="list-style-type: none"> <li>1. <i>Business Understanding</i></li> <li>2. <i>Data Understanding</i></li> <li>3. <i>Data Preparation</i></li> <li>4. <i>Modeling</i></li> <li>5. <i>Evaluation</i></li> <li>6. <i>Deployment</i></li> </ol>	CRISP-DM memiliki struktur yang sistematis dan terstandarisasi, memudahkan proses data mining secara iteratif dari pemahaman bisnis hingga <i>deployment</i> . Fleksibel dan dapat diterapkan di berbagai domain industri.	Kompleksitas dokumentasi dan penerapannya dapat menyulitkan bagi pengguna pemula, serta kurang detail pada teknik <i>modeling</i> .
KDD	<ol style="list-style-type: none"> <li>1. <i>Selection</i></li> <li>2. <i>Preprocessing</i></li> <li>3. <i>Transformation</i></li> <li>4. <i>Data Mining</i></li> <li>5. <i>Interpretation/Evaluation</i></li> </ol>	Fokus pada proses penemuan pengetahuan dan validasi hasil, cocok untuk eksplorasi data dan riset akademik.	Kurang menekankan aspek implementasi bisnis dan tahapan <i>deployment</i> .
SEMMA	<ol style="list-style-type: none"> <li>1. <i>Sample</i></li> <li>2. <i>Explore</i></li> <li>3. <i>Modify</i></li> <li>4. <i>Model</i></li> <li>5. <i>Assess</i></li> </ol>	Pendekatan teknis yang terstruktur dan mudah diterapkan, dengan fokus kuat pada modeling dan evaluasi performa model. fokus pada <i>modeling</i> dan <i>assessment</i> .	Tidak secara eksplisit mencakup tahapan pemahaman bisnis dan <i>deployment</i> , serta lebih terbatas pada alat statistik SAS.

Tabel 3.1 memperlihatkan perbandingan antara tiga metodologi utama dalam proses *data mining*, yaitu CRISP-DM, KDD, dan SEMMA. Metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan pendekatan yang paling banyak digunakan dalam industri karena bersifat fleksibel dan sistematis. Dengan enam tahapan utama, mulai dari pemahaman bisnis hingga *deployment*, CRISP-DM mampu menjembatani kebutuhan bisnis dan analisis data secara iteratif.

Keunggulan utama metode ini terletak pada kerangka kerjanya yang terdokumentasi dan mudah diaplikasikan dalam berbagai domain *industry*.

Namun demikian, CRISP-DM juga memiliki beberapa kekurangan, di antaranya adalah kurangnya detail pada aspek teknis saat *modeling*, serta tingkat kompleksitas yang mungkin menyulitkan pengguna pemula dalam memahami alur penerapannya. Meskipun demikian, struktur yang terstandarisasi membuat metode ini tetap menjadi pilihan yang kuat untuk berbagai jenis proyek *data mining*. Dibandingkan dengan metode KDD, yang lebih menekankan pada proses penemuan pola secara eksploratif, dan SEMMA, yang lebih fokus pada aspek teknis *modeling* dan evaluasi, CRISP-DM dinilai lebih lengkap karena mencakup seluruh siklus *data mining*, termasuk aspek pemahaman bisnis dan implementasi hasil. CRISP-DM terdiri dari enam tahapan utama, sebagai berikut:

a. *Business Understanding*

Tahap ini bertujuan memahami tujuan bisnis, kebutuhan, dan permasalahan yang ingin diselesaikan menggunakan *data mining*. Fokus utama adalah merumuskan pertanyaan penelitian dan tujuan yang jelas. Dalam penelitian ini, tahap *business understanding* dilakukan dengan mengidentifikasi kebutuhan pusat kebugaran untuk segmentasi pelanggan dan prediksi potensi pembelian layanan *personal trainer* sebagai dasar pengambilan keputusan strategi pemasaran.

b. *Data Understanding*

Tahap ini mencakup pengumpulan data, eksplorasi awal, dan penilaian kualitas data untuk memastikan data yang akan digunakan sudah sesuai dan lengkap. Pada penelitian ini, data dikumpulkan dari sistem internal yang terdiri dari beberapa data seperti data keanggotaan, pembelian PT, pembekuan *membership*, dan kunjungan pelanggan dengan periode data Januari 2022 hingga Maret 2025 yang akan digabungkan menjadi satu. Data juga dipelajari pada tahapan ini.

c. *Data Preparation*

Tahap ini meliputi pembersihan data, pengisian nilai hilang, transformasi variabel, dan persiapan *dataset* yang siap untuk pemodelan. Dalam penelitian ini, data diproses dengan melakukan penanganan data duplikat, *missing value*, dan *outlier*, *encoding* variabel kategorikal, konversi tipe data, dan penggabungan hasil segmentasi sebagai fitur tambahan untuk model prediksi.

d. *Modeling*

Pada tahap ini, algoritma *data mining* diterapkan untuk membangun model sesuai tujuan penelitian, seperti segmentasi atau prediksi. Penelitian ini menggunakan *K-Means Clustering* untuk segmentasi pelanggan dan *XGBoost* untuk prediksi potensi pembelian layanan *personal trainer*.

e. *Evaluation*

Pada tahap ini, evaluasi dilakukan untuk menilai kualitas model yang dihasilkan, baik secara statistik maupun dari sisi bisnis, untuk memastikan model dapat digunakan secara efektif. Dalam penelitian ini, model dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*, *ROC Curve* serta diperiksa apakah hasil segmentasi dan prediksi dapat memberikan *insight* yang berguna bagi manajemen *gym*.

f. *Deployment*

Tahap terakhir meliputi penerapan hasil analisis dan model ke dalam konteks bisnis nyata untuk pengambilan keputusan dan perbaikan berkelanjutan. Pada penelitian ini, hasil analisis disajikan dalam bentuk rekomendasi strategis bagi manajemen *gym* untuk mengoptimalkan pemasaran dan layanan *personal trainer*.

Dengan menerapkan metode CRISP-DM, penelitian ini dilakukan secara sistematis dari pemahaman bisnis hingga evaluasi model prediksi. Metode ini memastikan bahwa setiap tahap penelitian dilakukan dengan pendekatan berbasis data, sehingga hasil yang diperoleh dapat digunakan untuk memberikan rekomendasi strategis bagi pengelola *gym* dalam meningkatkan penjualan layanan *personal trainer*.

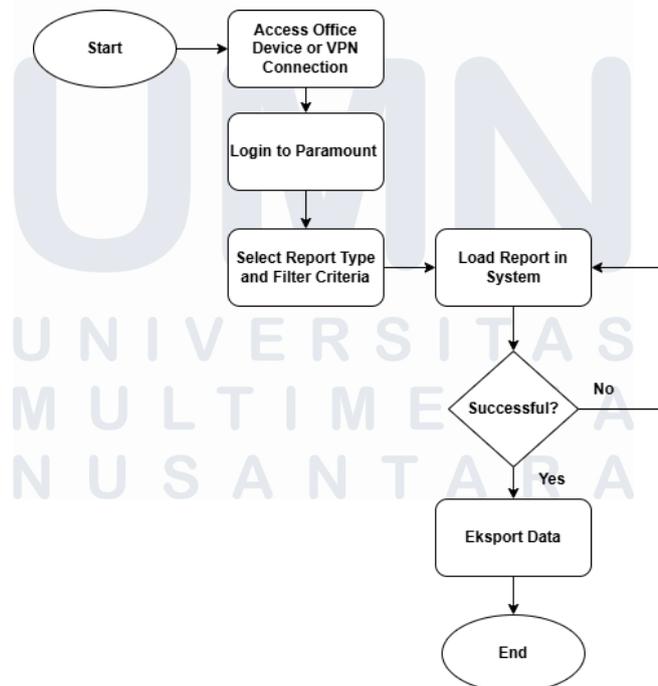
### 3.1 Teknik Pengumpulan Data

#### 3.1.1 Populasi dan Sampel

Populasi dalam penelitian ini mencakup seluruh pelanggan yang terdaftar di pusat kebugaran dalam sistem keanggotaan perusahaan. Sampel yang digunakan merupakan bagian dari populasi tersebut, yaitu pelanggan yang memiliki data lengkap dan valid berdasarkan kriteria tertentu, seperti aktif dalam periode analisis, memiliki riwayat kunjungan, serta informasi *membership* yang tercatat lengkap. Sampel dipilih secara non-probabilistik berdasarkan kelengkapan dan relevansi data untuk keperluan segmentasi dan prediksi.

#### 3.1.1 Sumber dan Proses Pengumpulan Data

Seluruh data yang digunakan dalam penelitian ini diperoleh dari sistem manajemen data internal perusahaan yang dikelola oleh pihak ketiga bernama Paramount. Sistem ini hanya dapat diakses melalui jaringan internal kantor atau koneksi VPN resmi, serta memerlukan autentikasi akun pengguna terdaftar. Paramount menyimpan seluruh riwayat transaksi, keanggotaan, dan aktivitas pelanggan yang dilakukan oleh perusahaan.



Gambar 3. 2 Alur Pengumpulan Data

Gambar 3.2 menggambarkan alur proses pengumpulan data melalui sistem Paramount. Proses dimulai dengan mengakses perangkat kantor atau menggunakan koneksi VPN resmi, kemudian dilanjutkan dengan *login* ke dalam sistem Paramount. Setelah berhasil masuk, pengguna akan memilih jenis laporan dan menetapkan filter yang diperlukan, seperti periode waktu dan cakupan data. Sistem kemudian memuat laporan sesuai permintaan. Pada tahap ini, terdapat proses verifikasi keberhasilan pemuatan data. Apabila pemuatan berhasil, data dapat langsung diekspor dalam format .csv atau .xlsx untuk dianalisis lebih lanjut. Namun, apabila proses pemuatan gagal, maka proses perlu diulangi dari tahap pemilihan laporan.

Tabel 3 2 Tabel Ringkasan *Dataset*

No	Nama <i>Dataset</i>	Jumlah Kolom	Jumlah Baris	Periode	Sumber	Deskripsi Singkat
1.	<i>Sales Gym</i>	50	60.480	Jan 2022 - Mar 2025	Paramount	Data penjualan dan informasi <i>membership gym</i> .
2.	<i>Sales PT</i>	42	26.303	Jan 2022 - Mar 2025	Paramount	Data penjualan layanan <i>personal trainer</i> .
3.	<i>Member Freeze</i>	14	29.743	Jan 2022 - Mar 2025	Paramount	Data status pembekuan ( <i>freeze</i> ) <i>membership</i> pelanggan.
4.	<i>Visit History</i>	21	1.048.575	Jan 2022 - Mar 2025	Paramount	Riwayat kunjungan pelanggan ke pusat kebugaran.

Tabel 3.2 menyajikan daftar *dataset* yang digunakan dalam penelitian ini. Setiap dataset mewakili informasi penting terkait aktivitas pelanggan pusat kebugaran, mulai dari data transaksi *membership*, pembelian layanan *personal trainer* (PT), status pembekuan keanggotaan (*freeze*), hingga riwayat kunjungan pelanggan ke pusat kebugaran. Data mencakup periode dari Januari 2022 hingga Maret 2025, yang dipilih secara khusus untuk merepresentasikan perubahan perilaku pelanggan pascapandemi dan memungkinkan analisis tren musiman serta siklus keanggotaan yang relevan.

Seluruh data bersumber dari sistem internal perusahaan bernama Paramount, maka tidak tersedia tautan URL publik untuk akses eksternal.

Seluruh data bersifat rahasia (*confidential*) dan hanya dapat digunakan oleh pihak internal perusahaan yang memiliki otorisasi.

### **3.1.2 Periode Data**

Periode data yang dianalisis mencakup rentang waktu dari Januari 2022 hingga Maret 2025. Pemilihan periode ini dilakukan secara sengaja untuk memastikan bahwa data yang digunakan bersifat terkini dan mencerminkan dinamika perilaku pelanggan pasca pandemi COVID-19. Tahun 2022 dipilih sebagai titik awal karena mencerminkan masa transisi signifikan dalam pola aktivitas masyarakat, termasuk peningkatan partisipasi dalam kegiatan kebugaran seiring pelonggaran pembatasan sosial. Sementara itu, batas akhir pada Maret 2025 memberikan cakupan data yang cukup panjang untuk menangkap tren musiman, siklus keanggotaan, serta fluktuasi pembelian layanan *Personal Trainer* (PT). Dengan rentang waktu lebih dari tiga tahun, data ini dianggap representatif untuk mengidentifikasi pola historis yang relevan serta membangun model prediksi yang stabil dan dapat diandalkan.

## **3.2 Variabel Penelitian**

Penelitian ini bertujuan untuk melakukan segmentasi pelanggan dan memprediksi potensi pembelian layanan *Personal Trainer* (PT). Oleh karena itu, variabel yang digunakan dibagi menjadi dua kelompok utama: variabel dependen (terikat) yang menjadi target prediksi dan variabel independen (bebas) yang berperan sebagai faktor yang mempengaruhi hasil prediksi.

### **3.2.1 Variabel Dependen**

Variabel dependen dalam penelitian ini adalah pembelian layanan *Personal Trainer* (PT), yang diukur berdasarkan apakah seorang pelanggan pernah membeli layanan PT atau tidak. Variabel ini dipilih sebagai target dalam model prediksi karena keputusan untuk membeli layanan PT dianggap sebagai hasil akhir yang dipengaruhi oleh berbagai faktor yang terkandung dalam variabel independen.

- a. Pembelian Layanan PT (*buy\_pt*):
  - i. Nilai 1: Pelanggan pernah membeli layanan PT.
  - ii. Nilai 0: Pelanggan tidak pernah membeli layanan PT.

### 3.2.2 Variabel Independen

Variabel independen adalah faktor-faktor yang dapat mempengaruhi keputusan pelanggan dalam membeli layanan PT. Beberapa variabel independen yang digunakan dalam penelitian ini antara lain:

- a. *Term Membership (term\_membership)*: Menunjukkan durasi kontrak keanggotaan yang disepakati di awal, seperti 1 bulan, 6 bulan, atau 12 bulan. Variabel ini mencerminkan komitmen awal pelanggan dalam mengikuti program *gym*, yang dapat berpengaruh terhadap kecenderungan membeli layanan tambahan seperti *Personal Trainer*.
- b. *Paid Today Membership (paid\_today\_membership)*: Nilai pembayaran yang dilakukan pelanggan pada saat transaksi *membership* terakhir. Variabel ini digunakan sebagai indikator kemampuan finansial atau *willingness-to-pay* pelanggan, yang berkaitan erat dengan potensi pembelian layanan PT.
- c. *Total Visits (total\_visits)*: Jumlah total kunjungan pelanggan ke *gym* selama masa aktifnya. Variabel ini mencerminkan seberapa aktif pelanggan dalam menggunakan fasilitas *gym* dan menjadi indikator penting dalam memahami *engagement* pelanggan terhadap program kebugaran.
- d. *Member Duration in Months (member\_duration\_months)*: Durasi total keanggotaan pelanggan dalam satuan bulan, yang merefleksikan loyalitas dan keterikatan terhadap layanan *gym*. Pelanggan dengan durasi yang lebih lama cenderung memiliki pengalaman lebih dan potensi lebih tinggi untuk menggunakan layanan tambahan seperti PT.
- e. *Visits per Month (visits\_per\_month)*: Rata-rata frekuensi kunjungan pelanggan per bulan. Fitur ini digunakan untuk

mengetahui intensitas penggunaan layanan *gym* oleh pelanggan dalam jangka waktu yang lebih spesifik.

- f. *Last Visit Gap (last\_visit\_gap\_days)*: Jarak waktu (dalam hari) sejak kunjungan terakhir pelanggan ke *gym*. Semakin kecil gap ini, semakin besar kemungkinan pelanggan masih aktif dan potensial untuk ditawarkan layanan PT.
- g. *Membership Type (membership\_type\_encoded)*: Jenis keanggotaan pelanggan yang telah di-*encode*, yaitu PIF (*Paid in Full*) atau EFT (*Electronic Fund Transfer*/bulanan). Jenis *membership* ini mencerminkan gaya pembayaran dan komitmen finansial pelanggan.
- h. *Gender (gender\_encoded)*: Jenis kelamin pelanggan yang telah dikonversi ke bentuk numerik. *Gender* digunakan sebagai variabel demografis yang bisa mempengaruhi preferensi terhadap layanan PT.
- i. *Peak Hour Category (peak\_hour\_category\_encoded)*: Waktu kunjungan dominan pelanggan yang menunjukkan apakah pelanggan lebih sering datang pada jam sibuk atau tidak. Hal ini memberikan indikasi tentang pola ketersediaan waktu dan keseriusan pelanggan dalam menjadwalkan sesi kebugaran.
- j. *Dominant Day Type (dominant\_day\_type\_encoded)*: Kategori hari kunjungan dominan pelanggan (*weekday/weekend*) yang menunjukkan preferensi waktu kunjungan pelanggan dalam konteks mingguan.
- k. *Dominant Time of Day (dominant\_time\_of\_day\_encoded)*: Preferensi waktu kunjungan pelanggan (pagi, siang, sore, malam) yang mencerminkan pola aktivitas harian pelanggan dan potensi ketersediaan mereka untuk layanan tambahan.
- l. *Freeze History (ever\_frozen\_encoded)*: Status apakah pelanggan pernah melakukan pembekuan *membership (freeze)*. Hal ini mencerminkan kontinuitas keanggotaan dan konsistensi penggunaan layanan *gym*.

- m. *Cluster Segment (cluster\_pt\_purchase\_level)*: Hasil segmentasi pelanggan dari proses *K-Means Clustering* yang dikategorikan ulang berdasarkan tingkat potensi pembelian PT. Fitur ini menjadi variabel strategis untuk menambahkan konteks segmentasi ke dalam model prediktif.

### 3.3 Teknik Analisis Data

Secara umum, data yang telah dikumpulkan dari berbagai sumber internal (*membership*, pembelian PT, pembekuan, dan kunjungan) akan digabungkan dalam satu dataset yang bersih dan siap diproses. Langkah pertama adalah *preprocessing* data yang kemudian akan siap untuk masuk ke proses *Modeling*. Pada tahap *modeling*, *K-Means Clustering* akan diterapkan untuk mengelompokkan pelanggan berdasarkan karakteristik dan pola aktivitas mereka. Hasil segmentasi ini kemudian digunakan sebagai fitur tambahan dalam model *XGBoost* untuk memprediksi kemungkinan pelanggan membeli layanan PT. Setelah model dibangun, evaluasi dilakukan untuk mengukur akurasi dan kualitas prediksi, yang akan dijelaskan lebih lanjut di bab 4.

*Tools* yang digunakan dalam penelitian antara lain adalah Microsoft Excel sebagai tempat penyimpanan data, Visual Studio Code sebagai *code editor* platform, dan Python sebagai bahasa pemrograman yang digunakan dalam melakukan *data mining*.

Tabel 3 3 Tabel Perbandingan *Tools*

Sumber: Sidhu K, 2022 [49]

Indikator	VS Code	R Studio
<b>Fokus utama</b>	Editor kode serbaguna untuk banyak bahasa pemrograman, termasuk R.	IDE khusus untuk bahasa R dan analisis data statistik.
<b>Bahasa pemrograman</b>	Mendukung banyak bahasa seperti Python, JavaScript, C++, R, dan lainnya lewat ekstensi.	Fokus utama hanya pada bahasa R.
<b>Tampilan antarmuka</b>	Modern, fleksibel, bisa dikustomisasi dengan tema dan ekstensi.	Tampilan standar tetapi dapat dikustomisasi dengan tema dan <i>plugin</i> khusus R.
<b>Fitur <i>debugging</i></b>	Mendukung <i>debugging</i> lintas bahasa melalui ekstensi dan <i>tools</i> pihak ketiga.	<i>Debugger</i> sederhana dan spesifik untuk bahasa R.

Indikator	VS Code	R Studio
<b>Kelebihan</b>	Serbaguna, ringan, fleksibel untuk berbagai bahasa dan kebutuhan pengembangan lintas platform.	Terintegrasi penuh dengan paket R, mendukung analisis data, visualisasi, dan akses langsung ke lingkungan kerja R.
<b>Kekurangan</b>	Konfigurasi manual bisa kompleks, membutuhkan ekstensi tambahan, dan konsumsi memori tinggi jika terlalu banyak <i>plugin</i> digunakan.	Terbatas hanya untuk R dan kurang fleksibel untuk kebutuhan lintas bahasa.

Tabel 3.3 menunjukkan perbandingan antara dua *Integrated Development Environment* (IDE) yang umum digunakan dalam analisis data, yaitu Visual Studio Code (VSCode) dan RStudio. VSCode merupakan *editor* kode yang fleksibel, mendukung berbagai bahasa pemrograman seperti Python, JavaScript, dan R melalui sistem ekstensi, serta memiliki antarmuka modern yang dapat disesuaikan. Selain itu, fitur *debugging* di VSCode mendukung berbagai bahasa, yang menjadikannya cocok untuk proyek lintas platform. Namun, fleksibilitas tersebut disertai kebutuhan konfigurasi manual dan konsumsi memori yang dapat meningkat jika banyak ekstensi digunakan.

Sebaliknya, RStudio adalah IDE yang secara khusus dirancang untuk bahasa R dan analisis data statistik. RStudio menawarkan integrasi mendalam dengan ekosistem R, termasuk kemudahan dalam melakukan analisis data dan visualisasi. Namun, IDE ini cenderung kurang fleksibel untuk penggunaan lintas bahasa pemrograman karena hanya difokuskan pada R.

Berdasarkan pertimbangan tersebut, penulis memilih menggunakan VS Code dalam penelitian ini karena kemampuannya untuk mendukung berbagai bahasa pemrograman dalam satu lingkungan kerja yang terintegrasi. Hal ini memudahkan integrasi antara proses analisis data, pengolahan file, dan automasi lainnya. Meskipun memerlukan konfigurasi tambahan, fleksibilitas yang ditawarkan menjadi nilai tambah yang signifikan dalam konteks penelitian ini.

Tabel 3 4 Tabel Perbandingan Bahasa Pemrograman

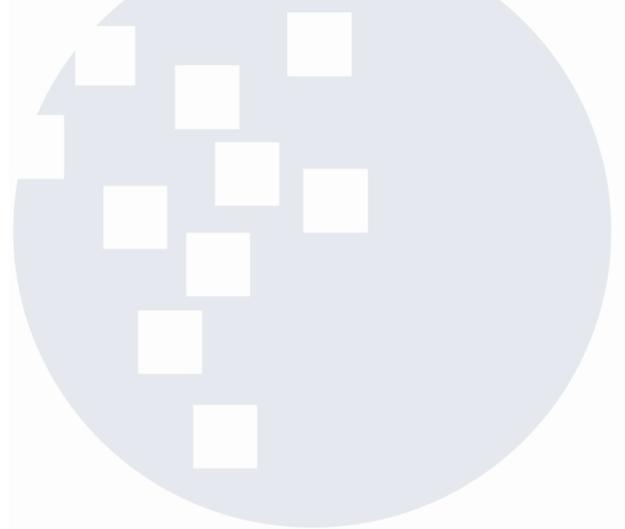
Sumber: Kamingu (2023), Joshi dan Upadhyay (2023), Chavan et al. (2023) [52][53][54]

Indikator	Python	R
<b>Fokus utama</b>	Bahasa serbaguna ( <i>general-purpose</i> ) yang digunakan luas dalam <i>data science</i> , <i>scripting</i> , automasi, dan pengembangan aplikasi.	Bahasa khusus ( <i>domain-specific</i> ) yang dikembangkan untuk keperluan statistik dan visualisasi data.
<b>Sintaks</b>	Sintaks sederhana, mirip bahasa Inggris, mudah dipelajari oleh pemula.	Sintaks lebih kompleks, membutuhkan pemahaman statistik dasar, dan bisa membingungkan bagi pemula.
<b>Ekosistem &amp; Library</b>	Ekosistem luas seperti <i>NumPy</i> , <i>Pandas</i> , <i>scikit-learn</i> , <i>TensorFlow</i> , dan <i>PyTorch</i> mendukung banyak aplikasi dari analisis data hingga <i>deep learning</i> .	Kuat untuk analisis statistik dan visualisasi. .
<b>Penggunaan umum</b>	Digunakan untuk <i>machine learning</i> , analisis prediktif, automasi, <i>API development</i> , dan pemrograman lintas platform.	Umumnya digunakan dalam riset akademik, analisis statistik klasik, dan eksplorasi data.
<b>Kelebihan</b>	Fleksibel, mudah integrasi dengan sistem lain serta memiliki dokumentasi komunitas yang luas.	Sangat kuat dalam pemodelan statistik dan visualisasi data, didukung banyak paket khusus statistik.
<b>Kekurangan</b>	Membutuhkan <i>library</i> tambahan untuk statistik lanjutan.	Kurang fleksibel untuk pengembangan aplikasi di luar konteks statistik atau <i>data science</i> .

Tabel 3.4 menyajikan perbandingan antara bahasa pemrograman Python dan R berdasarkan beberapa aspek utama seperti fokus penggunaan, sintaks, ekosistem *library*, serta kelebihan dan kekurangannya. Python merupakan bahasa serbaguna (*general-purpose*) yang banyak digunakan dalam berbagai bidang seperti *machine learning*, automasi, dan pengembangan aplikasi. Sintaksnya yang sederhana dan ekosistem pustakanya yang sangat luas menjadikannya mudah diakses oleh pemula namun tetap kuat untuk kebutuhan lanjutan. Di sisi lain, R merupakan bahasa pemrograman yang secara khusus dirancang untuk analisis statistik dan visualisasi data. Meskipun sangat kuat dalam konteks statistik klasik dan riset akademik, R memiliki keterbatasan dalam hal fleksibilitas untuk pengembangan aplikasi atau integrasi sistem yang lebih kompleks.

Berdasarkan pertimbangan tersebut, penulis memilih menggunakan Python sebagai bahasa pemrograman utama dalam penelitian ini karena fleksibilitasnya,

ketersediaan pustaka *machine learning* yang lengkap, serta kemampuan integrasi yang luas dengan berbagai platform analisis dan visualisasi modern. Selain Python, penulis juga memanfaatkan Microsoft Excel pada tahap awal untuk keperluan pengolahan data sederhana, seperti *filtering*, *cleaning*, dan *exploratory checks*. Excel dipilih karena kemudahan penggunaannya dan keakrabannya di lingkungan kerja. Namun, karena keterbatasannya dalam menangani data besar dan kebutuhan analisis lanjutan, Excel hanya digunakan sebagai alat bantu awal sebelum proses analisis utama dilakukan dengan Python.



UMMN

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA