# BAB III
# METODOLOGI PENELITIAN

## 3.1 Gambaran Umum Objek Penelitian

Penelitian ini menggunakan data dari Labskill, sebuah website platform pembelajaran online berbasis keterampilan di Indonesia yang menyediakan beberapa kursus digital seperti Udemy, Coursera, dan juga Dicoding. Data yang digunakan merupakan data transaksi yang diagregasi pada level bulanan untuk periode Januari 2023 hingga Desember 2024, menghasilkan 24 observasi bulanan. Unit analisis penelitian adalah data agregat per bulan, bukan data transaksi individual atau data harian, sehingga prediksi yang dihasilkan juga pada granularitas bulanan.

Data historis yang digunakan mencakup periode 2023–2025 dengan variabel yang meliputi informasi pelanggan, detail program kelas, bootcamp, riwayat transaksi, serta tingkat partisipasi dan engagement peserta. Dengan adanya data ini, penelitian dapat melakukan pendekatan predictive analytics yang lebih komprehensif, mengingat data LabSkill tidak hanya mencerminkan aspek finansial, tetapi juga perilaku konsumen. Data diperoleh dari sistem internal perusahaan dan diproses menggunakan bahasa pemrograman Python yang terintegrasi dengan berbagai pustaka analitik, sehingga memungkinkan eksplorasi variabel-variabel penting yang berpengaruh terhadap pendapatan dan pertumbuhan peserta.

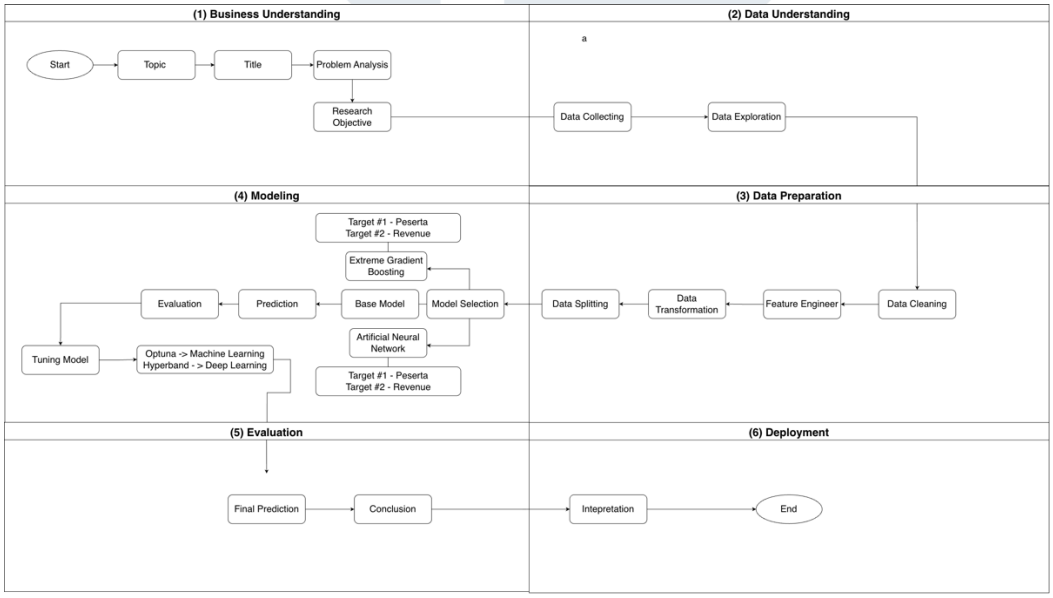
Dalam penelitian ini, dua algoritma utama digunakan untuk membangun model prediksi, yaitu Extreme Gradient Boosting sebagai representasi machine learning, serta Artificial Neural Network sebagai representasi deep learning. Pemilihan dua model ini didasarkan pada temuan penelitian terdahulu yang menunjukkan keunggulan masing-masing algoritma dalam menangani data tabular, data non-linear, maupun data time series. Dengan demikian, LabSkill menjadi kasus yang tepat untuk menguji efektivitas perbandingan model Machine Learning dan Deep Learning dalam memprediksi dua aspek penting, yaitu pendapatan tahunan dan pertumbuhan peserta.

Selanjutnya, penelitian ini juga menekankan pada optimasi hyperparameter sebagai salah satu langkah krusial untuk meningkatkan akurasi dan efisiensi model. Teknik optimasi modern seperti Optuna dan Hyperband digunakan agar performa model dapat ditingkatkan tanpa harus membangun jaringan ensemble atau hybrid yang kompleks dan mahal secara komputasi. Proses tuning ini diharapkan dapat menghasilkan model prediktif yang tidak hanya akurat, tetapi juga efisien serta lebih sesuai dengan keterbatasan sumber daya perusahaan. Pada tahap akhir, hasil prediksi akan dievaluasi untuk menentukan model terbaik yang dapat digunakan sebagai dasar pengambilan keputusan strategis LabSkill dalam aspek pendapatan maupun retensi peserta.

## 3.2 Metode Penelitian

### 3.2.1 Alur Penelitian

Gambar 3.1 menyajikan alur penelitian dalam bentuk diagram flowchart. Dengan visualisasi alur penelitian ini, diharapkan pemahaman mengenai prosespenelitian secara menyeluruh (end-to-end) dapat dipahami dengan jelas.




*Gambar 3. 1 Alur Penelitian*

Tahap (1) Business Understanding dimulai dengan identifikasi topik penelitian, perumusan masalah, dan penetapan objective penelitian yaitu membandingkan

performa Extreme Gradient Boosting dan Artificial Neural Network untuk prediksi pertumbuhan peserta dan pendapatan Labskill.

Tahap (2) Data Understanding meliputi pengumpulan data transaksi Labskill periode 24 bulan (Januari 2023 - Desember 2024) dan eksplorasi karakteristik data untuk memahami pola dan distribusi variabel.

Tahap (3) Data Preparation mencakup data cleaning, feature engineering untuk menghasilkan 8 features, transformasi data, dan pembagian dataset menjadi training set (20 bulan) dan testing set (5 bulan).

Tahap (4) Modeling merupakan inti dari penelitian ini, dimana dikembangkan dua jalur pemodelan paralel: Machine Learning menggunakan Extreme Gradient Boosting dan Deep Learning menggunakan Artificial Neural Network. Setiap algoritma dikembangkan dalam dua versi yaitu Baseline dengan hyperparameter default dan Optimized dengan hyperparameter hasil tuning menggunakan Optuna untuk Extreme Gradient Boosting dan Hyperband untuk Artificial Neural Network. Setiap model dilatih untuk memprediksi dua target peserta dan revenue, menghasilkan total 8 eksperimen.

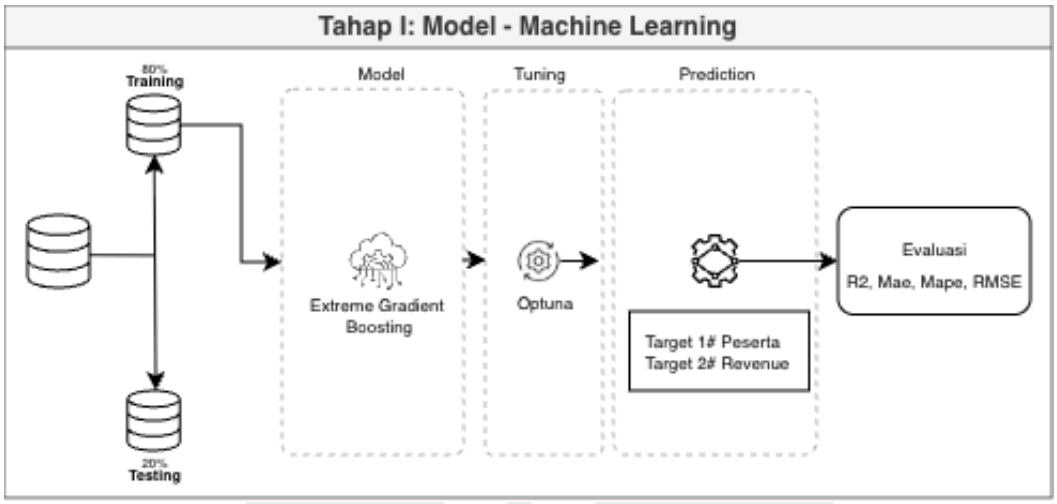
Tahap (5) Evaluation dilakukan dengan membandingkan performa keempat model menggunakan metrik MAE, MAPE, RMSE, dan $R^2$ untuk mengidentifikasi model terbaik.

Tahap (6) Deployment mengimplementasikan model terbaik yaitu dalam aplikasi Streamlit yang dapat melakukan forecasting 6 bulan ke depan untuk mendukung capacity planning dan budget allocation Labskill.

Gambar 3.2 Kerangka Tahap I Proses Pemilihan Model Machine Learning dan Gambar 3.3 dibawah akan menampilkan Kerangka Tahap II Proses Pemilihan Model Deep Learning, dari situ model yang terbaik akan digunakan untuk proses deployment nanti.

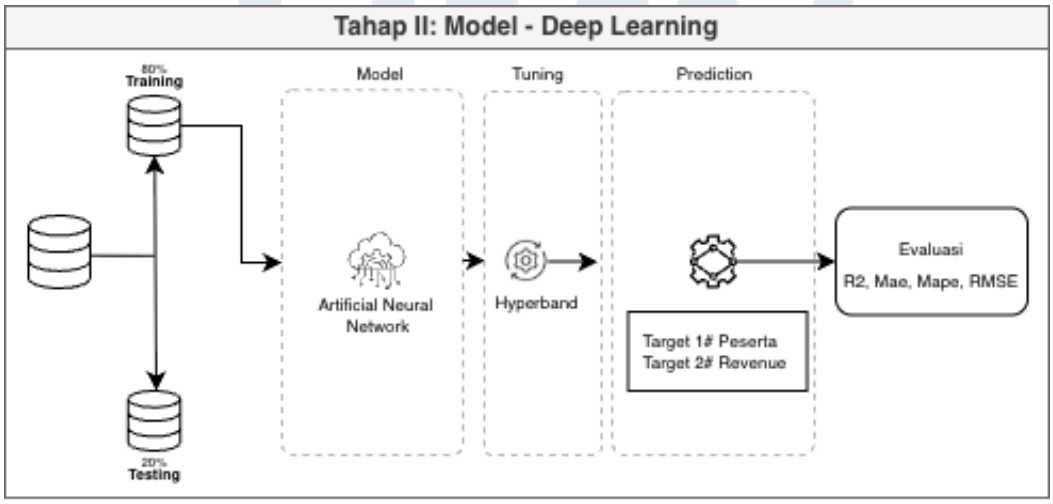
Pendekatan Machine Learning menggunakan algoritma Extreme Gradient Boosting yang merupakan implementasi gradient boosting berbasis decision

tree. Alur pemodelan untuk XG Extreme Gradient Boosting Boost ditunjukkan pada Gambar 3.2.




*Gambar 3. 2 Tahap 1: Prediksi Model - Machine Learning*

Data yang telah dipreparasi dibagi menjadi training set (80% atau 20 bulan) dan testing set (20% atau 5 bulan) menggunakan time-based split untuk menjaga urutan kronologis data time-series. Training set digunakan untuk melatih model, sementara testing set digunakan untuk validasi performa model pada data yang belum pernah dilihat sebelumnya. Sementara Tahap II mengarah terhadap Model deep learning yang akan digunakan dipenelitian ini, yaitu dengan menerapkan model Artificial Neural Network.




*Gambar 3. 3 Tahap 2: Prediksi Model - Deep Learning*

Sama halnya dengan penerapa pada tahap 1 seperti Extreme Gradient Boosting, data dibagi menjadi training set (80%) dan testing set (20%) dengan proporsi yang sama untuk memastikan perbandingan yang fair antara kedua algoritma. Proses modeling Artificial Neural Network mengikuti tahapan yang serupa dengan Extreme Gradient Boosting, hanya terdapat perbedaan pada framework tuningnya menjadi lebih sesuai dengan model Artificial Neural Network.

### 3.2.2 Kerangka Kerja Data Mining dan Data Analysis

Dalam penelitian ini, kerangka kerja yang diterapkan adalah CRISP-DM. Untuk memberikan gambaran perbandingan, berikut adalah tabel yang membandingkan beberapa kerangka kerja data mining, yaitu SEMMA dan CRISP-DM.

*Tabel 3. 1 Kerangka Kerja Data Mining*

| Indikator | SEMMA | CRISP-DM |
|---|---|---|
| Siklus | 1. Sample<br>2. Explore<br>3. Modify<br>4. Model<br>5. Assessment | 1. Business Understanding<br>2. Data Understanding<br>3. Data Preparation<br>4. Modeling<br>5. Evaluation<br>6. Deployment |
| Kelebihan | • Fokus pada analisis statistik dan eksplorasi data | • Lebih detail mengenai karena dimulai dari pemahaman bisnis sampai deployment |
| Kekurangan | • Tidak membahas detail pemahaman bisnis dan deployment | • Lebih kompleks dan membutuhkan waktu lebih lama dibanding SEMMA |

Standarisasi proses dalam data mining merupakan hal yang penting agar penelitian dan implementasi dapat dilakukan secara sistematis. Seiring berkembangnya teknologi data mining, beberapa model proses telah

dikembangkan untuk menjadi acuan, antara lain KDD, SEMMA, dan CRISP-DM [42], [43]. Model KDD merupakan model pertama yang diperkenalkan sebagai kerangka penemuan pengetahuan, sedangkan SEMMA dikembangkan oleh SAS dengan lima tahapan utama: sampling, exploring, modifying, modeling, dan assessing. SEMMA berfokus pada aspek teknis pemodelan, namun memiliki keterbatasan karena minim fleksibilitas dalam iterasi antar tahapan [44].

Sebaliknya, CRISP-DM dikembangkan antara tahun 1996–2000 oleh konsorsium yang didanai Uni Eropa seperti SPSS, dengan tujuan memberikan kerangka standar dan universal untuk proyek data mining. CRISP-DM terdiri dari enam fase iteratif, yaitu business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Sifat iteratif ini menjadikannya lebih adaptif dibandingkan SEMMA yang cenderung linier, karena CRISP-DM memungkinkan adanya feedback loop untuk meningkatkan kualitas hasil [43], [44].

Keunggulan CRISP-DM terletak pada sifatnya yang lebih generik dan independen terhadap domain tertentu, sehingga dapat diterapkan pada berbagai konteks, seperti pemasaran, pengendalian risiko keuangan, maupun diagnosis medis [39]. Lebih jauh, CRISP-DM tidak hanya menekankan aspek teknis pemodelan, tetapi juga menempatkan pemahaman bisnis dan pemahaman data sebagai tahap awal yang penting untuk memastikan kesesuaian hasil dengan tujuan organisasi [45]. Dengan dokumentasi yang komprehensif dan struktur yang jelas, CRISP-DM telah menjadi standar industri yang paling banyak digunakan, melampaui SEMMA yang lebih terikat pada perangkat lunak tertentu [45].

Berdasarkan perbandingan yang telah dibuat diatas, CRISP-DM dinilai lebih sesuai untuk penelitian ini dibandingkan SEMMA, karena menyediakan kerangka kerja yang lebih stabil, terdokumentasi dengan baik, serta fleksibel untuk berbagai jenis data dan tujuan analitik. Pendekatan CRISP-DM yang berfokus pada pemahaman bisnis dan data menjadikannya lebih praktis dan

aplikatif, khususnya dalam konteks penelitian prediksi berbasis Machine Learning dan Deep Learning, di mana iterasi dan optimasi model seringkali dibutuhkan untuk mencapai hasil yang optimal.

## 3.3 Teknik Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan tujuan mendapatkan informasi yang lengkap dan akurat mengenai pendapatan tahunan dan pertumbuhan peserta di platform LabSkill. Data diperoleh langsung dari sistem internal platform, mencakup semua pelanggan yang telah mendaftar, membeli kelas, atau mengikuti program bootcamp.

Data yang dikumpulkan meliputi beberapa aspek penting: transaksi pelanggan partisipasi program, dan engagement pelanggan frekuensi login, interaksi dengan materi, serta skor feedback. Setiap informasi dicatat secara sistematis oleh platform sehingga meminimalkan hilangnya data dan memastikan konsistensi antarentri.

### 3.3.1 Populasi dan Sampel

Dalam penelitian ini, purposive sampling digunakan sebagai prosedur pemilihan sampel, yang umum diterapkan dalam berbagai paradigma penelitian untuk memastikan pemilihan sampel yang relevan dan berkualitas. Pendekatan ini memungkinkan peneliti memperoleh data yang sesuai dengan tujuan penelitian, serta meningkatkan keandalan dan kredibilitas temuan [46],. Purposive sampling difokuskan pada kasus-kasus yang memerlukan informasi detail sehingga dapat memberikan wawasan yang mendalam terhadap penelitian yang diteliti. Ukuran sampel dan kriteria pemilihan ditentukan berdasarkan kesesuaian data dengan tujuan analisis prediksi.

Populasi dalam penelitian ini mencakup seluruh data pelanggan aktif yang tercatat dalam sistem internal perusahaan selama periode 2023 hingga 2025. Dari populasi tersebut, sampel yang digunakan adalah pelanggan yang memiliki data lengkap dengan fokus penelitian, yaitu prediksi pendapatan dan pertumbuhan peserta. Pemilihan sampel memperhatikan kelengkapan variabel numerik maupun demografis, seperti data profil pelanggan, transaksi, serta tingkat partisipasi dalam program bootcamp dan kelas berbasis keterampilan.

*Tabel 3. 2 Periode Data*

| Jangka Waktu Data | Rank | Sumber | Nama Data |
|---|---|---|---|
| Periode 2023–2025 | 1 | Sistem internal perusahaan | Data pelanggan, transaksi, engagement |

Tabel ini merangkum data utama dalam penelitian ini, yang mencakup periode 2023 hingga 2025. Data diperoleh sepenuhnya dari sistem internal perusahaan, terdiri dari informasi pelanggan, catatan transaksi, serta data engagement peserta. Data ini digunakan sebagai dasar dalam pengembangan dan evaluasi model prediksi.

### 3.3.2 Periode Pengambilan Data

Periode pengambilan data dalam penelitian ini dilakukan dengan memanfaatkan data internal perusahaan yang mencakup informasi pelanggan, transaksi, serta tingkat partisipasi peserta dalam program bootcamp dan kelas berbasis keterampilan. Data yang digunakan meliputi variabel numerik dan demografis yang relevan dengan analisis prediksi, seperti profil pelanggan, riwayat transaksi, dan engagement terhadap program.

Data tersebut digunakan sebagai dasar dalam analisis prediktif untuk memproyeksikan pendapatan tahunan dan pertumbuhan peserta pada platform LabSkill. Jangka waktu pengumpulan data adalah periode tahun 2023 hingga 2025, dengan sumber data utama berasal dari sistem internal perusahaan yang telah terdokumentasi secara lengkap. Data historis ini akan menjadi fondasi dalam pembangunan dan evaluasi model Machine Learning dan Deep Learning yang digunakan pada penelitian.

## 3.4 Variabel Penelitian

### 3.4.1 Variabel Independen

Variabel independen dalam penelitian ini mencakup data historis pelanggan dan aktivitas peserta pada platform LabSkill. Data ini meliputi profil demografis pelanggan, riwayat transaksi, jumlah program bootcamp dan juga kelas yang diikuti, serta tingkat engagement peserta terhadap program pembelajaran. Variabel-variabel ini menjadi input utama dalam proses

pemodelan, karena mencerminkan faktor-faktor yang memengaruhi pendapatan serta tingkat pertumbuhan peserta.

Data tersebut digunakan sebagai input untuk pengujian dua algoritma, yakni dua algoritma Machine Learning, Extreme Gradient Boosting serta algoritma Deep Learning, yaitu Artificial Neural Network. Kedua model ini akan dianalisis dan dibandingkan untuk menentukan metode terbaik dalam memprediksi pendapatan tahunan dan pertumbuhan peserta pada platform LabSkill.

### 3.4.2 Variabel dependen

Variabel dependen dalam penelitian ini adalah pendapatan tahunan dari bulanan dan pertumbuhan peserta pada platform LabSkill. Pendapatan bulanan direpresentasikan melalui total transaksi pelanggan dalam periode penelitian, sedangkan pertumbuhan peserta dihitung berdasarkan peningkatan jumlah pengguna atau partisipan program bootcamp dan kelas berbasis keterampilan dari tahun ke tahun.

Nilai-nilai ini berfungsi sebagai output yang diprediksi oleh masing-masing algoritma berdasarkan input dari variabel independen. Hasil dari keempat model akan dibandingkan untuk menentukan model prediksi yang paling akurat dan efisien dalam data perusahaan.

## 3.5 Teknik Analisis Data

Penelitian ini menerapkan teknik analisis data berbasis data mining dan machine learning, dengan tahapan utama meliputi pra-pemrosesan data eksplorasi data awal pembangunan model prediksi, serta evaluasi dan optimasi model. Proses ini mencakup transformasi data mentah menjadi bentuk yang siap dianalisis, identifikasi pola atau korelasi dalam data, serta penerapan algoritma pembelajaran mesin dan pembelajaran mendalam untuk membangun model yang mampu memprediksi performa akademik mahasiswa secara akurat.

Bahasa pemrograman yang digunakan dalam penelitian ini dipilih untuk mendukung proses data mining dan perancangan model. Dari berbagai pilihan yang

tersedia, penelitian ini akan memfokuskan penggunaan pada dua bahasa pemrograman, yaitu Python dan R. Berikut ini adalah perbandingan antara kedua bahasa pemrograman yang akan diterapkan dalam penelitian ini.

*Tabel 3. 3 Perbandingan Bahasa Pemrograman*

| Indikator | Python | R |
|---|---|---|
| Kelebihan | • Serbaguna untuk berbagai proyek<br>• komunitas besar<br>• Memiliki berbagai Pustaka luas seperti Pandas, Scikit-learn | • Unggul dalam analisis statistik<br>• Banyak pustaka untuk analisis data<br>• Cocok untuk eksplorasi pada tahap awal |
| Kekurangan | • Tidak memiliki repo pustaka umum<br>• Sulit mencari fungsi spesifik<br>• Kompleks mengintegrasikan pustaka tertentu | • Sulit dipelajari bagi pemula<br>• Indeks vektor mulai dari satu<br>• Sintaks beberapa fungsi tidak selalu jelas |

Python saat ini menjadi bahasa pemrograman paling populer menurut survei IEEE Spectrum [1]. Hal ini didukung oleh ekosistem pustaka yang sangat luas seperti Pandas, Scikit-learn, TensorFlow, dan PyTorch, yang memungkinkan pengolahan data, pembangunan model prediksi, serta deployment aplikasi dilakukan secara menyeluruh dalam satu lingkungan. Python juga memiliki komunitas global yang besar sehingga pembaruan, dokumentasi, serta dukungan praktis sangat mudah diakses. Fleksibilitas tersebut menjadikan Python unggul dibanding R dalam penelitian ini [47], karena tidak hanya digunakan untuk analisis statistik, tetapi juga untuk implementasi model Machine Learning dan Deep Learning hingga tahap visualisasi dan integrasi aplikasi [48].

Sementara itu, R tetap dikenal kuat dalam bidang analisis statistik tradisional dan eksplorasi awal data. Namun, keterbatasannya dalam integrasi sistem modern dan deployment aplikasi membuat penggunaannya kurang relevan untuk penelitian ini. Sebagai tambahan, berbagai penelitian sebelumnya menunjukkan bahwa Python lebih banyak digunakan dalam evaluasi model prediktif dengan metrik standar seperti root mean square error, mean absolute error, dan mean square error, [47]. Oleh karena itu, Python dipilih sebagai bahasa utama untuk teknik analisis data dalam penelitian ini, karena mampu mendukung kebutuhan end-to-end mulai dari data preprocessing, modeling, evaluasi, hingga visualisasi interaktif [45].