BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

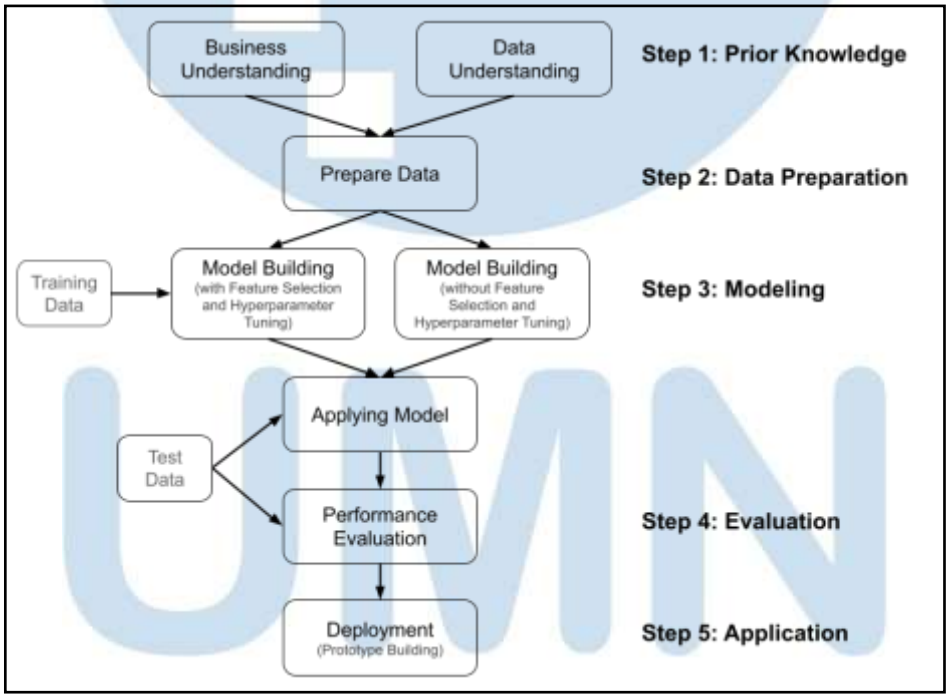
Dalam penelitian ini, fokus utama penelitian diarahkan pada pembangunan model klasifikasi persetujuan pengajuan pinjaman kredit nasabah Bank XY menggunakan beberapa algoritma pembelajaran mesin berdasarkan dataset nasabah Bank XY pada tahun 2021. Objek yang diteliti berupa dataset nasabah Bank XY yang mengandung informasi akan latar belakang, riwayat kredit dan detail aplikasi pinjaman kredit nasabah. Dilakukan peramalan atau memprediksi untuk masa yang akan datang, apakah nasabah Bank XY yang mengajukan pinjaman kredit layak disetujui atau tidak berdasarkan kemampuan nasabah dalam membayar kembali pinjaman beserta dengan bunga secara tepat waktu dalam waktu yang ditentukan. Perbandingan model yang berhasil dibangun menggunakan beberapa macam algoritma dilakukan untuk menentukan algoritma yang memiliki performa terbaik berdasarkan beberapa metrik evaluasi yang dipilih. Melalui ini, peneliti mampu mengetahui efektivitas, keunggulan serta kelemahan masing-masing algoritma yang dipakai. Kemudian, purwarupa (*prototype*) model klasifikasi yang mampu menerima *input* data dari pengguna dan menghasilkan hasil akhir klasifikasi dengan mengimplementasikan algoritma pemenang juga dibangun. *Prototype* merupakan salah satu bentuk *deployment* agar hasil model klasifikasi dapat diterapkan pada dunia nyata oleh pengguna secara langsung.

3.2 Metode Penelitian

Teknik *data mining* yang digunakan pada penelitian ini mengadopsi tahapan-tahapan dalam *data science* process berdasarkan metodologi *framework Cross Industry Standard Process for Data Mining (CRISP-DM)* yang mengalami sedikit modifikasi. *CRISP-DM,* sebuah teknik yang diperkenalkan pada tahun 1996, merupakan kerangka kerja yang paling sering dipakai untuk merumuskan solusi akan permasalahan-permasalahan *data science* [20]. Metode *CRISP-DM* terdiri dari 6 tahapan utama. Langkah pertama merupakan proses *Business*

Understanding, diikuti dengan tahap *Data Understanding* sebagai langkah kedua, tahap *Data Preparation* sebagai langkah ketiga, tahap *Modeling* sebagai langkah keempat, tahap *Evaluation* sebagai langkah kelima dan tahap *Deployment* sebagai langkah terakhir.

Namun, terdapat sedikit modifikasi pada penelitian ini, sejak penelitian hanya dilaksanakan hingga tahap evaluasi. Peneliti tidak melanjutkan mengimplementasi tahap *Knowledge*. Dengan ini, metode penelitian yang diadopsi terdiri dari 5 tahapan utama, yaitu *Prior Knowledge*, *Data Preparation*, *Modeling*, *Evaluation* dan *Application*. Gambar 3.1. menunjukkan *flowchart* proses *data science* berdasarkan metodologi *CRISP-DM* yang dimodifikasi oleh peneliti.



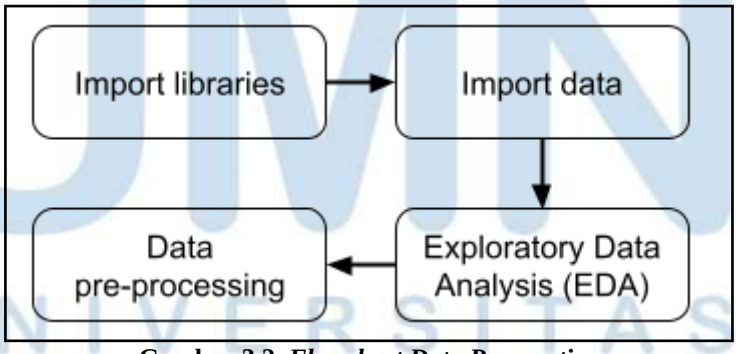
Gambar 3.1. *Flowchart* Metode Penelitian

3.2.1 Prior Knowledge

Dalam tahap ini, dilakukan proses pemahaman bisnis (*business understanding*) dan dataset (*data understanding*) yang digunakan dalam penelitian. Pengetahuan umum terkait bisnis dan data diperlukan agar peneliti mampu memiliki gambaran dasar terkait objek penelitian. Bank XY selaku bisnis yang ingin dipahami dan dataset Bank XY selaku data yang ingin dipelajari merupakan kedua komponen utama dari tahap *Prior Knowledge*. Pemahaman umum yang salah diinterpretasi dapat berakibat fatal pada proses-proses selanjutnya, dimana mampu mengakibatkan tujuan akhir penelitian gagal untuk diraih.

Menurut [20], proses pemahaman bisnis terdiri dari empat aktivitas yaitu menentukan objektif bisnis, menilai situasi, menentukan tujuan yang ingin dicapai melalui *data mining*, serta menghasilkan rencana proyek untuk memenuhi tujuan yang telah ditentukan. Proses pemahaman data menurut [20] juga terdiri dari empat aktivitas, yaitu mengumpulkan data awal, mendeskripsikan data, mengeksplorasi data, dan memastikan kualitas dan sumber data.

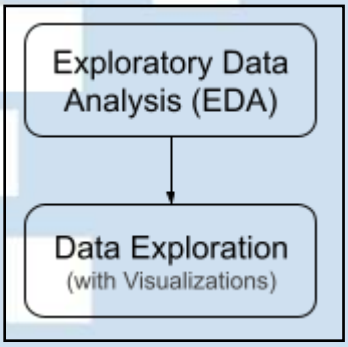
3.2.2 Data Preparation



Gambar 3.2. ***Flowchart Data Preparation***

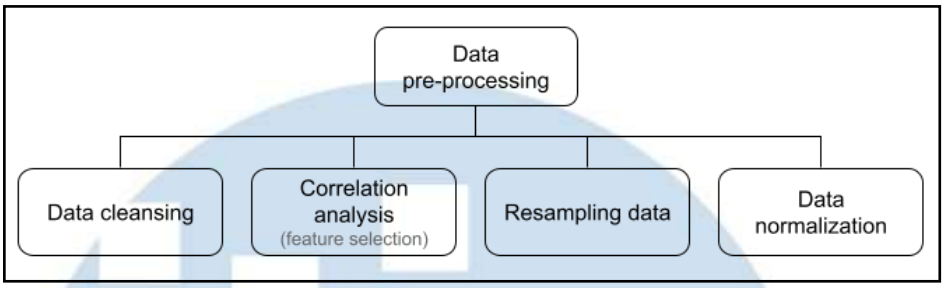
Dalam tahap *data preparation*, dilakukan proses persiapan data yang mencakup kegiatan pembersihan dan pengolahan data mentah yang berhasil diperoleh dan dipahami pada tahap sebelumnya. Tujuan proses persiapan data antara lain adalah guna mentransformasikan data menjadi

data yang berkualitas yang mampu memenuhi kebutuhan proses pembangunan model. Proses preparasi data berperan penting pada performa akhir model yang berhasil dibangun pada tahap selanjutnya. Kesalahan kecil pada tahap pengolahan data berdampak besar pada berkurangnya tingkat akurasi model. Empat proses yang dilaksanakan dalam tahap *Data Preparation* dijelaskan pada gambar 3.2., yang antara lain adalah (1)*Import libraries*, (2)*Import data*, (3)*Exploratory Data Analysis (EDA)* dan (4)*Data pre-processing*.

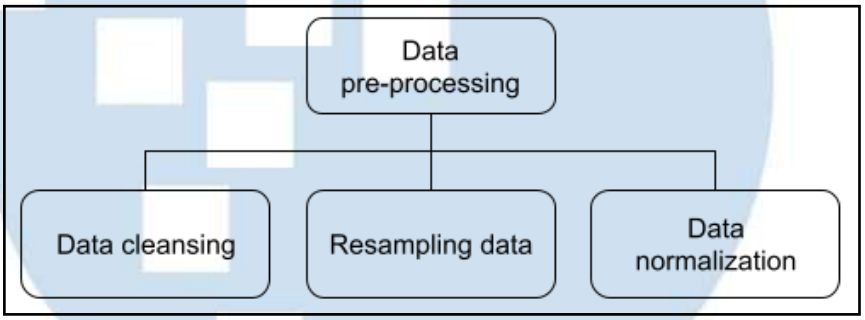


Gambar 3.3. Proses *Exploratory Data Analysis*

Proses *import libraries* dan *import data* melibatkan aktivitas impor *library* atau *package* pendukung dan dataset yang diperlukan ke dalam platform pembangunan model yaitu Google Colab. Proses *Exploratory Data Analysis (EDA)* terdiri dari sebuah kegiatan utama sebagaimana yang diperlihatkan melalui gambar 3.3., yaitu *data exploration* (eksplorasi data). Proses eksplorasi data didukung oleh implementasi grafik-grafik visualisasi untuk membantu menemukan karakteristik umum maupun pola-pola tersembunyi pada dataset.



Gambar 3.4.1. Proses *Data pre-processing* untuk Proses *Modeling* dengan Implementasi *Feature Selection*



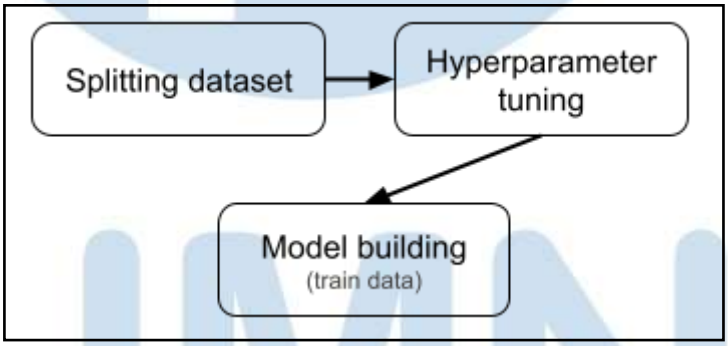
Gambar 3.4.2. Proses *Data pre-processing* untuk Proses *Modeling* tanpa Implementasi *Feature Selection*

Gambar 3.4.1. menjelaskan proses *data pre-processing* untuk proses *modeling* yang mengimplementasikan teknik *feature selection*. Proses *data pre-processing* terdiri dari empat kegiatan utama, yaitu (1)*Data cleansing* (pembersihan data), (2)*Correlation analysis* (analisis korelasi) untuk variabel bertipe data kategorikal dan numerikal, (3)*Resampling data* dan (4)*Data normalization* (normalisasi data). Proses pembersihan data dilaksanakan dimana dataset diolah dan ditransformasi untuk meningkatkan kualitas data berdasarkan temuan-temuan yang dihasilkan melalui proses eksplorasi data. Proses analisis korelasi antar variabel bertipe data kategorikal dan numerikal dengan *target variable* (Loan_Status) dilakukan sebagai bentuk *feature selection* guna menyeleksi variabel-variabel yang diperlukan untuk proses pembangunan model. Teknik analisis korelasi yang digunakan baik antar tipe data kategorikal ataupun numerikal dengan *target variable* adalah teknik Pearson dan Spearman. Proses *resampling data* menggunakan teknik *Synthetic*

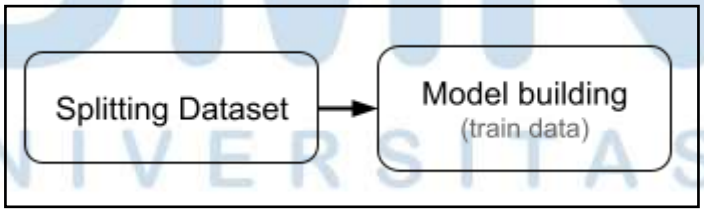
Minority Oversampling Technique (SMOTE) digunakan untuk mengatasi distribusi data yang tidak seimbang pada *target variable* (Loan_Status). *SMOTE* merupakan salah satu metode *oversampling* terpopuler untuk mengatasi masalah ketidakseimbangan pada data. *SMOTE* berfungsi dengan meningkatkan data pada kelas minoritas dengan cara replikasi [55]. Proses normalisasi data dilakukan untuk mengubah nilai-nilai pada variabel numerik ke dalam skala yang umum.

Gambar 3.4.2. menjelaskan proses *data pre-processing* untuk proses *modeling* yang tidak mengimplementasikan teknik *feature selection*, dimana proses *data pre-processing* hanya terdiri dari pembersihan data, *resampling data* dan normalisasi data, tanpa melakukan analisis korelasi.

3.2.3 Modeling



Gambar 3.5.1. *Flowchart Modeling* dengan Implementasi *Feature Selection* dan *Hyperparameter Tuning*



Gambar 3.5.2. *Flowchart Modeling* tanpa Implementasi *Feature Selection* dan *Hyperparameter Tuning*

Dalam tahap *modeling*, dataset yang dihasilkan melalui tahap preparasi data digunakan untuk membangun model klasifikasi persetujuan pengajuan pinjaman kredit. Pembangunan model menggunakan algoritma-algoritma terpilih dengan menerapkan *feature selection* (selanjutnya disingkat dengan *FS*) dan *hyperparameter tuning* (selanjutnya disingkat dengan *HT*) terlebih dahulu, dan model tanpa menerapkan *feature selection* dan *hyperparameter tuning* dilakukan untuk dibandingkan antara satu sama lain. Hal ini bertujuan untuk mengetahui kegunaan teknik *FS* dan *HT*, serta untuk mencari model yang menghasilkan performa terbaik.

Gambar 3.5.1. menjelaskan proses yang dilaksanakan dalam tahap *modeling* dengan implementasi *FS* dan *HT* yang terdiri dari 3 proses utama, yaitu (1)*Splitting dataset*, (2)*Hyperparameter tuning* dan (3)*Model building*. Sebelum mengimplementasikan berbagai macam algoritma terpilih untuk membangun model klasifikasi, dilakukan proses pembagian data (*splitting dataset*) menjadi dua bagian, yaitu *training data* dan *testing data*. *Training data* berlaku sebagai data yang digunakan untuk membangun model klasifikasi, sementara *testing data* adalah data yang digunakan sebagai dasar acuan validasi dan evaluasi yang akan diuji terhadap model yang telah dibangun [56]. Kemudian, proses *HT* dilakukan untuk memilih parameter-parameter yang paling optimal bagi masing-masing algoritma pembelajaran mesin agar seluruh model mampu memiliki akurasi semaksimal mungkin. Teknik *grid search* dan *randomized search* digunakan sebagai metode *HT*. Proses terakhir yaitu pembangunan model (*model building*) melibatkan implementasi algoritma-algoritma *supervised learning* terpilih untuk membangun model menggunakan *training data*, yang antara lain mencakup: *Support Vector Machine, Naïve Bayes, Random Forest, Logistic Regression* dan *K-Nearest Neighbors*.

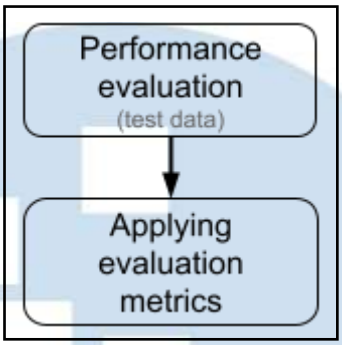
Gambar 3.5.2. menjelaskan proses yang dilaksanakan dalam tahap *modeling* tanpa implementasi *FS* dan *HT*, dimana hanya terdapat dua proses utama yaitu *splitting dataset* dan *model building*.

Seluruh algoritma pembelajaran mesin, baik yang menerapkan pendekatan *supervised learning*, *unsupervised learning* maupun *reinforcement learning* memiliki kelebihan dan kekurangan yang beragam. Berdasarkan [57], [58], [59], [60], [61], [62], dan [63], tabel 3.1.menjelaskan keunggulan (+) dan kekurangan (-) dari setiap algoritma *supervised learning* terpilih untuk membangun model klasifikasi.

Tabel 3.1. Keunggulan dan Kerugian Setiap Algoritma

Algoritma	**+**	**-**
Support Vector Machine	Mampu melakukan generalisasi dengan seimbang, sehingga hasil prediksi tetap baik meski jumlah sampel terbatas	Tidak efisien bila dataset berukuran besar sejak eksekusi memerlukan banyak waktu
Naïve Bayes	Tidak memerlukan jumlah *training data* yang banyak, serta memiliki performa baik dalam *multi-class prediction*	Seringkali memerlukan teknik *smoothing* untuk mengatasi permasalahan *zero frequency* yang ditemui
Random Forest	Default *hyperparameter* cukup untuk memberikan hasil prediksi yang akurat	Jumlah pohon yang banyak mengarah kepada kecepatan jalan model yang lambat
Logistic Regression	Transparan dan simpel, dimana setiap proses dapat diamati dan lebih mudah dipahami, serta output berupa *probabilistic* yakni angka antara 0 dan 1	Rentan akan *noise* dan cenderung menghadapi permasalahan *overfitting*
K-Nearest Neighbors	Variasi *distance criteria* yang dapat dipilih secara bebas oleh pengguna	Penentuan nilai *K* sangat berpengaruh pada hasil akhir model

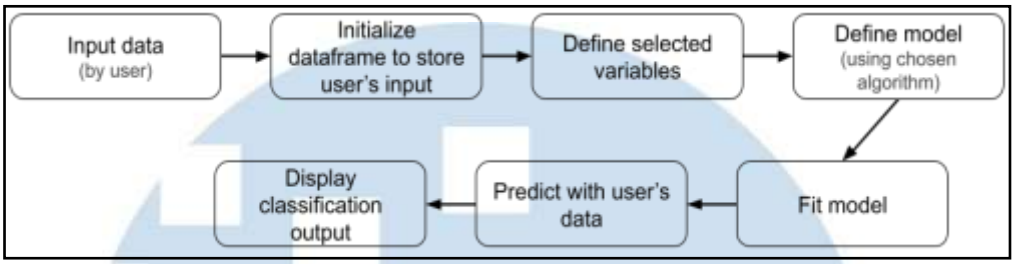
3.2.4 Evaluation



Gambar 3.6. *Flowchart Evaluation*

Dalam tahap *evaluation*, dilakukan pengujian performa model yang telah dibangun untuk divalidasi dan dievaluasi menggunakan *testing data* serta beberapa metrik evaluasi yang ditentukan. *Confusion Matrix, accuracy, precision, recall, F1-score dan ROC (Receiver Operating Characteristic) score* merupakan metrik-metrik evaluasi yang digunakan untuk menilai performa model. Evaluasi model diterapkan baik pada model-model yang mengimplementasikan teknik *FS* dan *HT*, maupun model-model yang tidak mengimplementasikan kedua teknik tersebut. Hasil evaluasi model menunjukkan seberapa baik performa masing-masing model dengan algoritma-algoritma tersendiri yang berhasil dibangun. Dengan ini, peneliti mampu mendapatkan algoritma pemenang yang paling cocok untuk diimplementasikan guna membangun model klasifikasi pengajuan pinjaman kredit pada *future data*.

3.2.5 Application



Gambar 3.7. *Flowchart Prototype* Klasifikasi Persetujuan Pinjaman Kredit

Dalam tahap *application*, dilakukan pembangunan purwarupa (*prototype building*) klasifikasi persetujuan pinjaman kredit berbasis *browser-based platform* Google Colab menggunakan bahasa pemrograman Python, sebagai bentuk peluncuran atau *deployment* akan *prototype* yang telah dibangun untuk uji coba diimplementasikan secara langsung pada dunia nyata. *Prototype* tidak dalam bentuk aplikasi berbasis web (*web apps*) ataupun aplikasi berbasis *mobile* (*mobile apps*), melainkan hanya sebatas program Python yang tidak memiliki *interface*. *Prototype* berfungsi dengan menerima data yang dimasukkan oleh pengguna (*user*), kemudian menjalankan prediksi klasifikasi untuk menentukan apakah pinjaman kredit yang diajukan layak disetujui atau tidak, dan menghasilkan output klasifikasi.

Sebagaimana yang ditunjukkan dalam gambar 3.7. mengenai alur cara kerja *prototype* klasifikasi persetujuan pinjaman kredit yang dirancang, *prototype* terdiri dari 7 proses utama. Pada proses pertama yaitu *input data (by user)*, diperlukan campur tangan pengguna untuk mengisi kolom-kolom data yang diperlukan. Data yang dimasukkan oleh pengguna berlaku sebagai data mentah yang akan diproses atau diklasifikasi. Pada proses kedua yaitu *initialize dataframe to store user's input*, melibatkan inisialisasi dataframe untuk menyimpan data yang telah dimasukkan oleh pengguna. Pada proses ketiga yaitu *define selected variables*, melibatkan penentuan variabel-variabel terpilih sebagai bentuk *feature selection* untuk keperluan pembangunan model klasifikasi. Pada proses keempat yaitu

define model (using chosen algorithm), dilakukan pendefinisian fungsi algoritma beserta dengan parameter-parameter tertentu untuk membangun model klasifikasi menggunakan algoritma pemenang dengan hasil evaluasi terbaik di antara yang lainnya. Pada proses kelima yaitu *fit model*, melibatkan penerapan model pada *data training* untuk melatih model. Pada proses keenam yaitu *predict with user's data*, melibatkan penerapan model pada data yang telah dimasukkan oleh pengguna untuk dilakukan prediksi. Pada tahap ketujuh atau terakhir yaitu *display classification output*, melibatkan penampilan hasil akhir output klasifikasi kepada pengguna, yang menentukan apakah pinjaman kredit yang diajukan layak disetujui atau tidak.

Untuk menjalankan *prototype*, pengguna memerlukan akun Google untuk mengakses Google Colab, serta *source code* mentah yang dihasilkan melalui penelitian ini. Langkah berikutnya setelah memenuhi kebutuhan tersebut hanya memerlukan kontribusi pengguna untuk menekan tombol *run code/program* dan keikutsertaan pengguna dalam memasukkan data yang diperlukan.

3.3 Variabel Penelitian

Penelitian ini memiliki dua tipe variabel, yakni variabel independen dan variabel dependen. Variabel independen merupakan variabel yang memiliki korelasi atau pengaruh (berdampak) pada variabel lain. Variabel dependen merupakan sebuah variabel yang dipengaruhi (terdampak) oleh variabel independen.

3.3.1 Variabel Independen

Dalam penelitian ini, kelayakan nasabah Bank XY yang dinilai melalui informasi latar belakang, riwayat kredit nasabah Bank XY, serta detail aplikasi pinjaman kredit merupakan variabel independen yang ditentukan. Variabel pada dataset yang termasuk ke dalam variabel independen Informasi Latar Belakang Nasabah antara lain adalah "Gender", "Married", "Dependent_No", "Education", "Self_Employed",

"Applicant_Income", "CoApplicant_Income" dan "Property_District". Variabel pada dataset yang termasuk ke dalam variabel independen Riwayat Kredit Nasabah antara lain adalah "Credit History". Variabel pada dataset yang termasuk ke dalam variabel independen Detail Aplikasi Pinjaman Kredit antara lain adalah "Loan_Amount" dan "Loan_Amount_Term".

3.3.2 Variabel Dependen

Dalam penelitian ini, persetujuan pengajuan pinjaman kredit nasabah Bank XY merupakan variabel dependen yang ditentukan. Variabel pada dataset yang termasuk ke dalam variabel dependen Persetujuan Pengajuan Pinjaman Kredit antara lain adalah "Loan_Status".

3.4 Teknik Pengumpulan Data

3.4.1 Sumber Data

Dataset yang digunakan dalam penelitian ini diambil dari situs Kaggle pada Februari 2022, sebuah situs yang mengizinkan pengunjung untuk mencari, mengambil dan meluncurkan dataset secara gratis, yang dipublikasikan pada tanggal 17 Februari 2022. Berikut adalah *link* atau *url* dataset: https://www.kaggle.com/vikramkumar001/partial-bank-loan-dataset.

3.4.2 Metode Pengumpulan Data

Dalam penelitian ini, metode pengumpulan data sekunder digunakan sebagai teknik pengumpulan data. Metode pengumpulan data sekunder adalah proses pengambilan data dimana data yang diperoleh bersumber dari data yang telah ada sebelumnya seperti buku, jurnal, situs web, *scientific papers*, dan dokumen lainnya yang berkaitan dengan topik penelitian [64]. Situs web Kaggle dimanfaatkan sebagai sumber perolehan data. Data-data dikumpulkan melalui sebuah formulir yang dirancang oleh Bank XY, dimana kemudian dibagikan kepada para nasabah untuk diisi secara langsung, jika ingin mengajukan pinjaman kredit.

3.5 Teknik Analisis Data

Berdasarkan [65] dan [66], peneliti menggunakan empat tipe pendekatan sebagai bentuk solusi analitik data, sebagaimana yang dijelaskan dalam tabel 3.2..

Tabel 3.2. Teknik Analisis Data

No	Teknik Analisis Data	Deskripsi
1	Descriptive analytics	Digunakan untuk menganalisis data historikal atau kejadian yang pernah terjadi (masa lalu)
2	Predictive analytics	Digunakan untuk memprediksi data di masa yang akan datang berdasarkan data historikal
3	Prescriptive analytics	Digunakan untuk menghasilkan rekomendasi terkait aksi yang perlu dilakukan berdasarkan hasil analisis
4	Visual analytics	Digunakan untuk merepresentasikan data dalam bentuk visual yang mudah dipahami dan menarik

Tahap pemahaman yang lebih mendalam terkait bisnis dan data dalam penelitian memerlukan teknik analisis deskriptif, dikarenakan fakta serta informasi terkait bisnis Bank XY dan data yang digunakan sebagai objek penelitian merupakan data historikal. Selanjutnya, tahap *modeling* untuk membangun model klasifikasi yang mampu menentukan apakah nasabah Bank XY mampu membayar pinjaman secara tepat waktu atau tidak melibatkan teknik analisis prediktif, dikarenakan hasil akhir model mampu menyediakan prediksi untuk masa yang akan datang. Berdasarkan hasil akhir model yang dibangun, Bank XY mampu memperoleh *insight-insight* yang berguna, dimana hal ini dapat menjadi acuan rekomendasi akan aksi-aksi yang perlu diambil oleh Bank XY. Rekomendasi yang dihasilkan dan diterapkan untuk keuntungan Bank XY merupakan bentuk implementasi analisis preskriptif. Terakhir, untuk membantu pembaca dalam memahami data, serta membuat data menjadi lebih mudah dilihat (*eye-catching*), teknik analisis visual digunakan untuk merepresentasikan data ke dalam bentuk visual yang lebih menarik melalui penggunaan berbagai grafik visualisasi. Tipe-tipe grafik atau *chart* yang digunakan pada penelitian ini meliputi tetapi tidak terbatas pada: *line chart, bar chart, pie chart, cross-table (cross tab), gauge, scatter plot*, dan *heat map*.