BAB III
METODOLOGI PENELITIAN

Pada bab ini membahas mengenai metodologi yang digunakan untuk menyelesaikan masalah klasifikasi kredit mulai dari identifikasi masalah, tahapan penelitian, dan penjelasan model gabungan metode *Decision Tree – AdaBoost* dan *Logistic Regression - AdaBoost.* Pada penelitian ini akan menggunakan metodologi CRISP-DM.

3.1 Deskripsi Masalah

Mengklasifikasikan kredit menjadi *good credit* dan *bad credit* merupakan hal yang harus dilakukan dalam industri keuangan terutama pada Bank. Dalam hal ini, masalah yang dihadapi adalah mengklasifikasikan kredit menggunakan *dataset* dari *Home Credit*. *Home Credit* merupakan penyedia layanan pembiayaan konsumen internasional yang saat ini telah beroperasi di tujuh negara di Eropa Tengah dan Timur serta Asia.

Penelitian ini berfokus pada klasifikasi kredit dengan mengidentifikasi calon peminjam yang berpotensi memiliki risiko rendah untuk gagal membayar pinjaman. Dengan memahami faktor-faktor yang mempengaruhi keterlambatan pembayaran, perusahaan dapat menggunakan model prediktif untuk melakukan penilaian risiko yang lebih baik bagi calon peminjam.

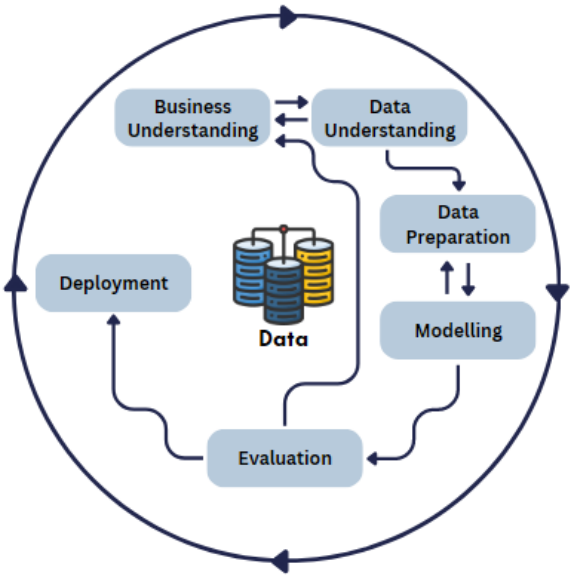
Data yang dimiliki saat ini berisi tentang berbagai informasi pribadi dan keuangan milik individu yang mau mendaftarkan pinjaman dan sebelumnya sudah menerima pinjaman dari *Home Credit*. Setiap atribut akan diseleksi dan diklasifikasikan menggunakan *Decision Tree* dan *Logistic Regression*. Lalu dilanjutkan menggunakan *Adaptive Boosting* yaitu, setiap atribut yang terpilih dimulai dengan memberikan bobot yang sama pada setiap sampel. Jika sebuah sampel gagal atau salah dalam dalam proses pelatihan, bobot yang lebih besar akan diberikan, yang dapat membuat pengklasifikasi pada iterasi berikutnya akan fokus mempelajari sampel yang gagal tersebut.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder, yaitu data yang diperoleh dari situs web *Kaggle* (https://www.kaggle.com/) yang diakses pada tanggal 1 Oktober 2023. Penelitian ini menggunakan 356.255 data peminjam dan 122 variabel atau atribut riwayat kredit peminjam. Dari 122 variabel akan digunakan beberapa variabel yang esensial dalam klasifikasi *credit scoring* dan akan dibahas pada bab 4.

3.3 Langkah-Langkah Penelitian

Langkah awal yang dilakukan yaitu studi literatur mengenai metode *Decision Tree*, *Logtistic Regression* dan *AdaBoost*. Penelitian ini juga menerapkan metode CRISP-DM (*Cross Industry Standard Process for Data Mining*) yaitu sebuah metode yang digunakan dalam proses data *mining*. *Data mining* adalah proses menemukan pola, korelasi, dan tren yang bermakna dari sejumlah besar data dengan menggunakan teknologi pengenalan pola teknik statistik dan matematika (Daniel & Chantal, 2014). Fungsi dari *data mining* menurut Han, et al. (2012), meliputi karakterisasi dan diskriminasi, asosiasi, penggalian pola, klasifikasi dan regresi, analisis klaster, dan mendeteksi outlier. *Data mining* biasanya melibatkan pembersihan data, integrasi data, pemilihan data, transformasi data, penemuan pola, evaluasi pola, dan presentasi pengetahuan. Data mining dapat dilakukan pada jenis data apa pun selama data tersebut bermakna, seperti data *database*, data *warehouse*, data transaksional, dan tipe data lanjutan (Han, et al., 2012). Berikut diagram metode CRISP-DM menurut Brzozowska, et al. (2023).



Gambar 3. 1 CRISP-DM

Metode CRISP-DM terdiri dari enam langkah, yaitu (Brzozowska, et al., 2023):

1. Pemahaman Bisnis (*Busisness Undertanding*)

   Pada tahap ini berfokus pada pemahaman tujuan dari bisnis atau data yang akan digunakan dan masalah yang ingin diselesaikan dengan prinsip *data mining*. Data tersebut selanjutnya akan diolah agar dapat diklasifikasikan menjadi data *good credit* dan *bad credit*. Penelitian ini menggunakan *dataset* berjudul "*Home Credit Default Risk"* yang diambil dari situs web *Kaggle* (https://www.kaggle.com/).

2. Pemahaman Data (*Data Understanding*)

   Tahap ini merupakan tahap pemahaman data mulai dari mengumpulkan data awal, mendeskripsikan data, mengeksplorasi data, dan memverifikasi kualitas data.

3. Persiapan Data (*Data Preparation*)

   Pada tahap persiapan data mencakup seluruh kegiatan untuk menyusun kumpulan data akhir (data yang akan digunakan dalam pemodelan) dari data mentah awal. Persiapan data dimulai dari memilih atribut data, membersihkan data, mengkonstruksi data, memformat data, dan mendeskripsikan kumpulan data akhir.

   Dalam memilih atribut data yang akan digunakan sebagai variabel independen dalam model menggunakan Korelasi Pearson dengan rumus:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{(n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2)(n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2)}}$$

di mana:

r : koefisien korelasi

n : jumlah data

$\sum_{i=1}^{n} X_i Y_i$: jumlah data dari hasil perkalian antara setiap nilai x dan y

$\sum_{i=1}^{n} X_i$: jumlah dari setiap nilai x

$\sum_{i=1}^{n} Y_i$: jumlah dari setiap nilai y

Korelasi Pearson hanya dapat menganalisis data yang berkategori numerik, sehingga untuk data tipe objek harus dikelompokkan terlebih dahulu berdasarkan nilai target 0 dan 1.

4. Pemodelan (*Modelling*)

   Pada tahap pemodelan dimulai dari memilih teknik atau metode pemodelan yang akan digunakan lalu membangun model. Penelitian ini memilih untuk menggunakan algoritma *Decision Tree, Logistic Regression*, dan dilanjutkan dengan algoritma *AdaBoost*. Proses pemodelan dalam penelitian ini menggunakan bantuan *Google Collab* dalam bahasa pemrograman *Python*.

5. Evaluasi (*Evaluation*)

   Pada tahap ini dilakukan evaluasi model untuk mengukur kinerja model. Teori mengenai evaluasi dapat dilihat pada Subbab 2.5.

3.4 Langkah-Langkah Pembentukan Model Gabungan

Pada pembuatan model gabungan metode *Decision Tree – AdaBoost* dan *Logistic Regression – AdaBoost*, proses pelatihan dimulai dengan membuat pohon keputusan awal menggunakan *Decision Tree* dan *Logistic Regression*. Pada tahap awal ini, setiap sampel tidak diberi bobot terlebih dahulu. Setelah pembuatan pohon keputusan pertama, setiap sampel diberi bobot yang disesuaikan dengan metode *AdaBoost*. Berikut adalah langkah-langkah penggabungan metode *Decision Tree – AdaBoost* dan *Logistic Regression – AdaBoost.*

1. Pembuatan model dasar

   Gunakan algoritma *Decision Tree* dan *Logistic Regression* untuk membuat model awal berdasarkan data pelatihan.

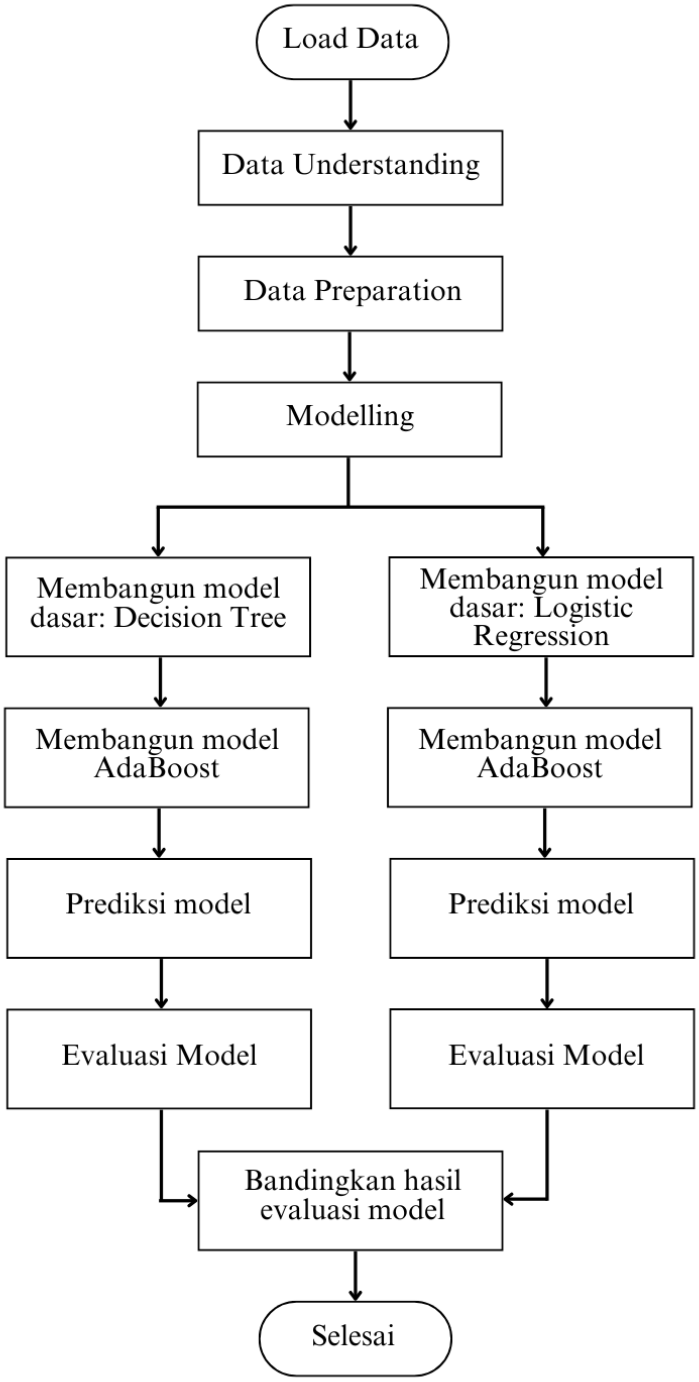
2. Penyesuaian algoritma *Adaboost*
   a. Hitung bobot kesalahan pada setiap sampel dalam data pelatihan.
   b. Jika sampel salah diklasifikasikan, tingkatkan bobot pada setiap sampel tersebut pada iterasi berikutnya.
3. Pembuatan model baru
   a. Gunakan bobot yang telah disesuaian untuk melatih pohon keputusan baru pada data pelatihan.
   b. Proses ini dapat diulang beberapa kali, di mana setiap iterasi fokus memperbaiki kesalahan yang dibuat pada model sebelumnya.
4. Gabungkan model

   Gabungkan semua pohon keputusan yang telah dibuat pada langkah sebelumnya dengan memberikan bobot pada setiap model.

5. Prediksi

   Lakukan prediksi pada setiap model berdasarkan bobotnya dan kombinasikan hasil prediksi untuk menghasilkan prediksi akhir.

Berikut diagram alir penelitian metode *hybrid Decision Tree – AdaBoost* dan *Logistic Regression – AdaBoost*:



Gambar 3. 2 Diagram alir penelitian