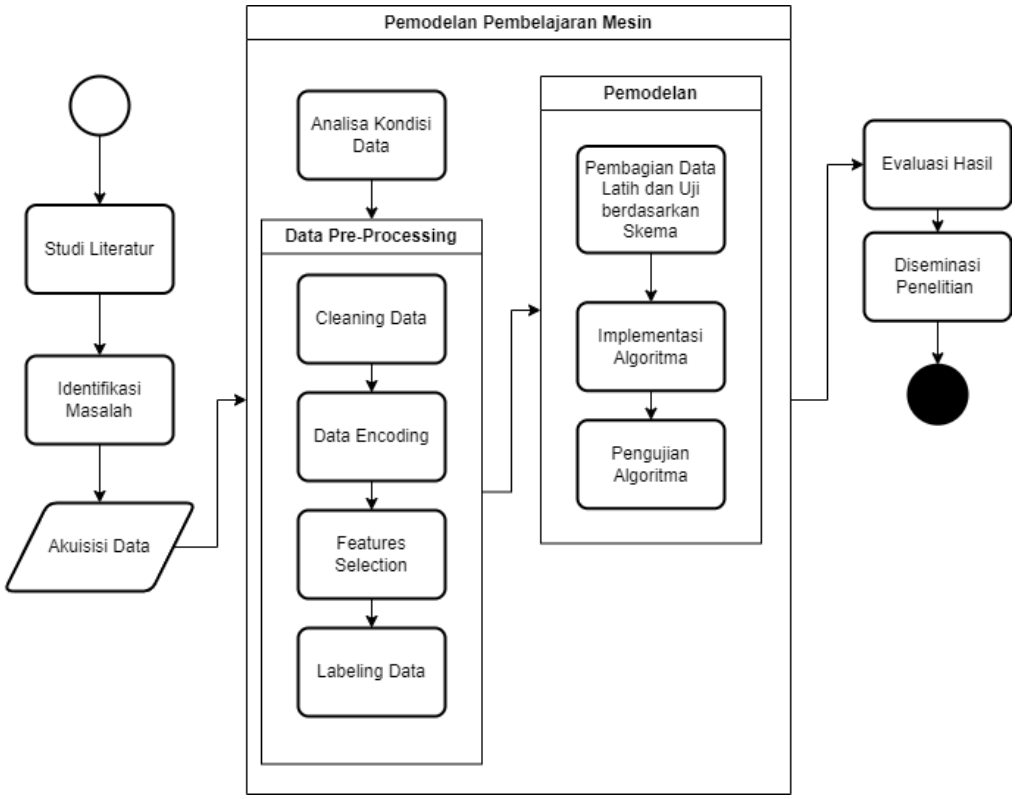
BAB III
METODE PENELITIAN

Bab ini menjelaskan terkait objek penelitian, metode yang digunakan dalam membangun model pembelajaran mesin untuk prediksi performa akademik. Skema rangkaian metodologi yang digunakan dalam penelitian dibuat dengan modifikasi metodologi pembelajaran mesin yang dapat dilihat pada gambar 3.1 sebagai berikut (Hapke and Nelson, 2020).



Gambar 3.1. Metodologi Penelitian Pembelajaran Mesin

3.1. Studi Literatur

Studi literatur dilakukan dengan mengumpulkan penelitian terdahulu dan pendalaman materi yang berkaitan dengan model pembelajaran mesin klasifikasi serta prediksi performa akademik yang kemudian akan menjadi acuan dalam melakukan penelitian ini.

3.2. Identifikasi Masalah

Identifikasi masalah yang dilakukan adalah dengan menganalisa banyaknya data yang telah dikumpulkan dan dapat diolah atau digali untuk menemukan suatu informasi berupa hasil perhitungan akurasi. Dan bagaimana kemampuan dari setiap algoritma dapat menyelesaikan permasalahan yang ada.

3.3. Akuisisi Data

Data yang digunakan dalam penelitian ini merupakan data yang berisi identitas mahasiswa, dan nilai-nilai yang terdapat pada proses SNMPTN. Data ini didapat dari suatu instansi perguruan tinggi negeri di Indonesia, melalui proses wawancara terhadap pengelola data sampai data berhasil diakuisisi. Data yang tersedia untuk digunakan pada penelitian ini merupakan data mahasiswa periode masuk tahun 2018, 2019, dan 2020. Berkas yang didapat berekstensi excel. Data ini memiliki kelas yang dapat ditinjau pada tabel 3.1 sebagai berikut.

Tabel 3.1. Daftar Fitur Dataset SNMPTN

No	**Nama Kolom**
1	Nomor_pendaftaran
2	Nama_siswa
3	NPSN_sekolah
4	Jenis kelamin
5	Tanggal lahir
6	Status penerima beasiswa
7	Asal sekolah
8	Kota/Kabupaten
9	Provinsi
10	Prodi pilihan 1
11	Prodi pilihan 2
12	Ranking sekolah
13	Nilai Mapel Ujian Nasional
14	Lulus pada prodi
15	Lulus pada pilihan
16	Nilai prestasi siswa yang mencakup nilai mata pelajaran (X1)
17	Nilai berdasarkan peringkat siswa di sekolah (X2)
18	Nilai prestasi lain (X3)

19	Nilai portofolio khusus program studi tertentu (X4)
20	Nilai prestasi sekolah berdasarkan akreditasi sekolah (X5)
21	Nilai berdasarkan rasio jumlah pendaftar dan diterima peserta SBMPTN pada tahun sebelumnya (X6)
22	Nilai berdasarkan rasio jumlah pendaftar dan diterima peserta seleksi mandiri pada tahun sebelumnya (X7)
23	Nilai berdasarkan rerata IPK di PTN masing-masing dari alumni asal sekolah pendaftar (X8)
24	Nilai parameter lainnya termasuk parameter aksesibilitas siswa (X9)
25	Nilai total (XT)

Data SNMPTN akan digabungkan dengan data nilai IPK 2 semester awal mahasiswa. Nilai IPK tersebut akan digunakan sebagai bahan pengelompokan klasifikasi mahasiswa yang menjadi label untuk diprediksi oleh model.

3.4. Pengolahan Data

Data diatas akan melalui tahap pemrosesan dengan tujuan menyesuaikan dengan input pada model yang akan dibuat. Pemrosesan dilakukan melalui beberapa tahap, diantaranya:

1. Analisa kondisi data

Pada tahap ini dilakukan proses pengecekan data untuk melihat karakteristik data, distribusi data, adanya *missing value*, adanya redudansi data, korelasi antar fitur, untuk mengetahui langkah yang diperlukan pada proses persiapan data.

2. Pembersihan data

Pada tahap ini dilakukan proses pembersihan data dengan metode substitusi nilai sesuai dengan karakter data ataupun penghapusan data.

3. Encoding data

Pada tahap ini dilakukan pemrosesan data untuk menyetarakan level data pada tipe data bertipe kategori. Tahap ini mengkonversi tipe data atau pun nilai nya sehingga model akan menganggap data ini sama. Salah satu algoritma yang digunakan adalah *One Hot Encoding*. *One Hot Encoding* merupakan tahap dimana data yang semula memiliki nilai kategori dirubah menjadi tipe boolean dengan cara

mentransformasikan nilai didalamnya menjadi kolom-kolom baru. One-hot encoding diterapkan pada kelas data asal Provinsi, lulus_pada_prodi, dan lulus_pada_pil.

4. Features Selection

Pemilihan dilakukan untuk menyesuaikan data dengan input serta output yang diharapkan pada model yang dibangun. Kelas yang berisikan data identitas akan dihapus. Kemudian fitur nilai total (XT) akan dihapus dikarenakan fitur ini merupakan hasil penjumlahan dari fitur nilai-nilai SNMPTN lainnya. Sebelum dilakukan proses pemilihan fitur, dilakukan pengecekan hubungan fitur menggunakan *correlation matrix*.

5. Labeling Data

Proses ini bertujuan untuk memberikan label pada dataset menjadi bentuk klasifikasi. Proses pelabelan menggunakan data IPK semester 1 dan 2 yang dihitung reratanya yang kemudian diklasifikasikan menjadi 4 kelas. Penentuan label ini mengambil referensi dari pengkategorian IPK pada asal institusi dari data yang digunakan pada penelitian ini. Kelas klasifikasi dapat dilihat pada tabel 3.2 sebagai berikut.

Tabel 3.2 Kelas Label Klasifikasi

Kode	**Kategori**	**Nilai rerata IPK**
1	Sangat Kurang	0 <= x <= 1
2	Kurang	1 < x <= 2
3	Baik	2 < x <= 3
4	Sangat Baik	3 < x <= 4

3.5. Pemodelan

1. Pembagian data latih dan data uji

Membagi data menjadi data latih dan uji dari total data, yang akan digunakan untuk evaluasi akurasi model. Skema yang dibuat pada penelitian ini akan menjadi 3 skema yang dapat dilihat pada tabel 3.3.

Tabel 3.3 Skema Pembagian Data Latih dan Data Uji

Skema	Persentase Data Latih	Persentase Data Uji
A	60%	40%
B	70%	30%
C	80%	20%

Skema tersebut dirancang berdasarkan penggunaan umum pada praktikal yang telah terbukti efektif menyelesaikan banyak kasus. Angka tersebut dipakai dengan bergantung pada kasus penggunaan serta ketersediaan data. Terdapat hukum pareto pada bidang studi statistik yang menunjukan bahwa pembagian 80%-20% merupakan titik optimum. Namun tidak ada aturan baku yang menyatakan bahwa terdapat keterkaitan hukum tersebut pada pembagian persentase data latih dan uji model pembelajaran mesin.

2. Implementasi algoritma

Dataset yang telah diolah kemudian diproses menuju tahap pelatihan model dengan menggunakan algoritma Decision Tree, Random Forest, Support Vector Machine, dan Extreme Gradient Boosting.

3. Pengujian Algoritma

Setiap algoritma yang telah diimplementasikan pada model dilakukan proses evaluasi dengan melakukan perbandingan menggunakan *confusion matrix* pada nilai akurasi, presisi, *recall*, dan *f1-score*.

3.6. Evaluasi Hasil

Hasil evaluasi setiap model dikumpulkan kemudian dilakukan proses perbandingan untuk dilakukan analisa. Analisa tersebut bertujuan untuk meninjau hasil yang ditunjukan pada proses pembuatan model. Selain itu juga dilakukan validasi terhadap hasil evaluasi yang dikeluarkan oleh model. Validasi dilakukan dengan membandingkan hasil prediksi model terbaik dengan sample data. Sample data dihitung menggunakan metode pengambilan sampel *Central Limit Theorem* (*CLT*) (Islam, 2018) dengan formula (3.1) berikut.

$$Sample\ Size = (Z\text{-}score^2 * p * (1 - p)) / (E^2 / N) \quad (3.1)$$

Z-score : nilai skor sesuai dengan tingkat persentase kepercayaan yang dipilih (contoh 1,96 untuk tingkat kepercayaan 95%)

p : perkiraan proporsi populasi dengan karakteristik yang diminati (nilai 0,5 apabila tidak terdapat informasi tersedia)

E : margin error

N : ukuran populasi

3.7. Diseminasi Penelitian

Setelah penelitian selesai dilakukan, dibuat laporan Tugas Akhir dan artikel jurnal sebagai bentuk dokumentasi penelitian untuk dipublikasikan sehingga dapat digunakan sebagaimana mestinya.