BAB III
METODOLOGI PENELITIAN

3.1 Analisis Masalah

Sistem penerima beasiswa merupakan topik utama yang akan dibahas dalam penelitian ini. Berdasarkan pengamatan langsung, bahwasannya penyeleksian penerima beasiswa di lingkungan institut teknologi sumtera masih menerapkan sistem pemberkasan dan diseleksi secara manual. Sistem pemberkasan tentunya tidak sejalan dengan tujuan pembangunan berkelanjutan (*Suntainaility Development Goals*) yang mana akan membuat banyak kertas terbuang. Selain itu penyeleksian secara manual memiliki kemungkinan terjadinya *human-error*. Human-error merupakan kesalahan dari manusia bisa berupa lalai, lupa, keliru dsb. Klasifikasi penerima beasiswa dengan menerapkan konsep data *mining* untuk membuat model penyeleksian dapat mengurangi penggunaan kertas dan terjadinya *human-error*.

Dalam pengembangannya SPK membutuhkan data primer yang berasal dari pihak kampus. Adapun data yang dimaksud adalah data mahasiswa penerima beasiswa dan non-penerima beasiswa dengan atribut selain data identitas pribadi. Nantinya *dataset* akan dilakukan proses learning sehingga mendapatkan sebuah model yang dapat menentukan seseorang layak diterima atau tidak. Untuk meningkatkan akurasi dari model yang dikembangkan maka dilakukan tahap pemangkasan. Setelah itu, untuk mengetahui berapa tingkat akurasi yang didapat maka digunakan perhitungan akurasi dengan *confusion matrix*.

3.2 Alat dan Bahan

Untuk melakukan penelitian penulis membutuhkan alat dan bahan. Adapun alat yang digunakan . meliputi perangkat lunak dan perangkat keras. Perangkat keras yang digunakan berupa laptop dengan spesifikasi seperti pada **Tabel 3.1**. Perangkat lunak yang digunakan yaitu menggunakan bahasa pemrogramman python. Adapun spesifikasi perangkat lunak lainnya dapat dilihat pada **Tabel 3.2**.

Tabel 3.1. Spesifikasi Perangkat Keras

No	Perangkat Keras	Spesifikasi
1	Laptop	TOSHIBA dynabook R734/K
2	*Processor*	Intel Core i5 Generasi ke-4
3	*RAM*	12 GB
4	*Storage*	SSD 120 GB

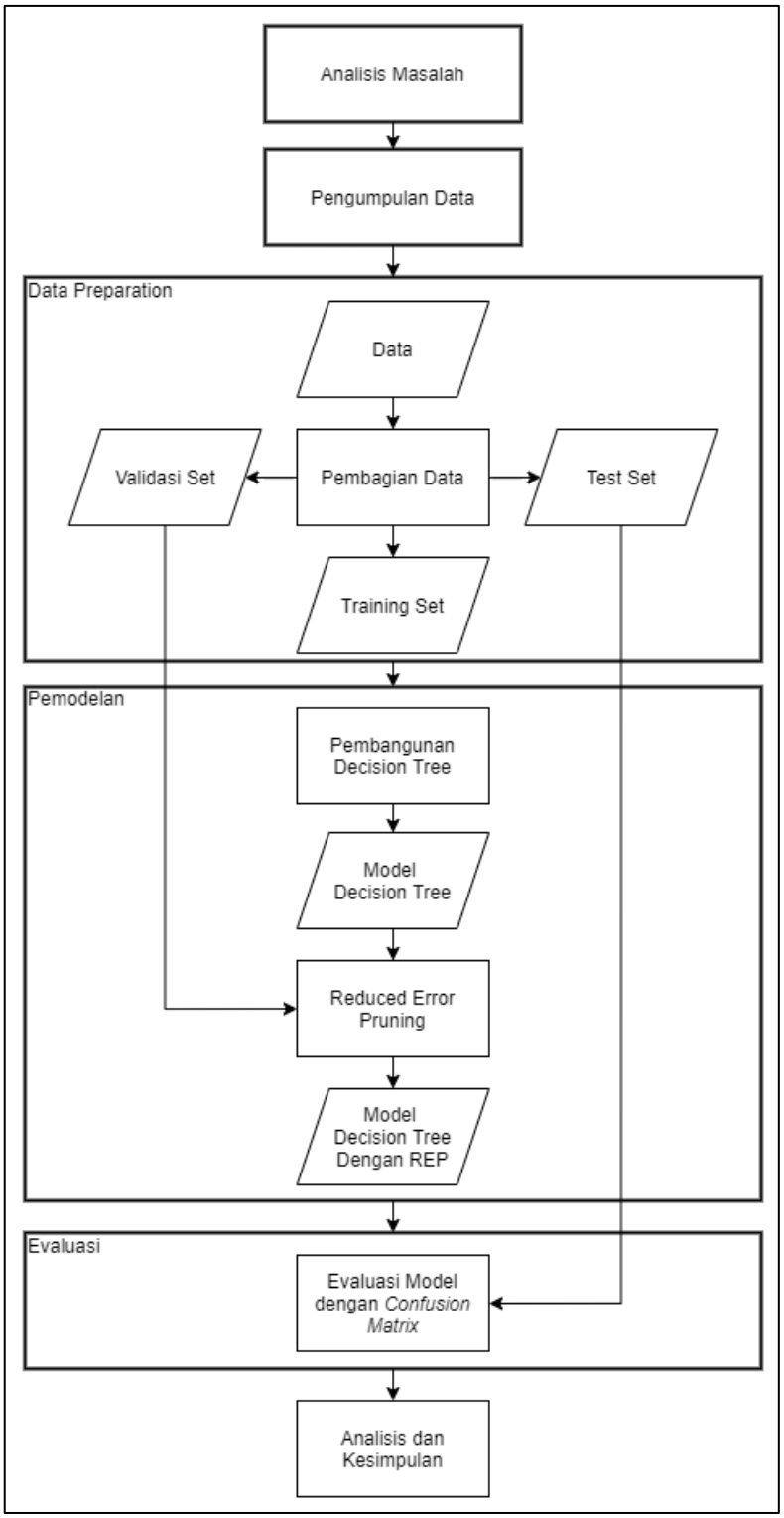
Tabel 3.2. Spesifikasi Perangkat Lunak

No	Perangkat Keras	Spesifikasi
1	Sistem Operasi	Windows 10
2	Bahasa Pemrograman	Python 3.7
3	*Integrated Development Environtment* (IDE)	Jupyter notebook (Anaconda 3)

Bahan yang digunakan merupakan data primer. Data primer yang dimaksud yaitu data Penerima Beasiswa Bank Indonesia tahun 2019-2020 serta data pembanding berupa data mahasiswa yang tidak menerima beasiswa. Data primer didapatkan dari sub bagian akademik dan kemahasiswaan institut teknologi sumatera (Itera).

3.3 Tahapan Peneliatan Klasifikasi Penerima Beasiswa

Pada bagian tahap penelitian penulis akan memaparkan secara keseluruhan serangkaian langkah-langkah yang akan dilaksanakan pada penelitian. Pada penelitian kali ini tahapan merujuk pada CRISP-DM yang dapat dilihat pada **Gambar 2.1**. Namun pada penelitian kali ini tidak sampai pada tahap *deployment*. Serta pada penelitian ini penulis menggunakan istilah yang berbeda. Sehingga langkah-langkah utama yang akan dilakukan yaitu : Analisis Masalah, Pegumpulan Data, Persiapan Data, Pemodelan , dan Evaluasi. pada **Gambar 3.1** tahapan yang merupakan langkah utama penelitian dijelaskan dengan bentuk kotak dan garis tebal. Adapun sub langkah dijelaskan dengan bentuk kotak dan ketebalan garis normal.



Gambar 3.1. Diagram Tahapan Penelitian

3.3.1 Pengumpulan Data

Pengumpulan data merupakan tahapan yang dilakukan untuk mendapatkan

data hingga data dapat diproses secara digital. Setelah mengetahui topik utama permasalahan yang diangkat dalam penelitian. Data yang akan digunakan dalam penelitian ini berupa data primer. Data penelitian ini didapatkan dari sub bagian akadmik dan kemahasiswaan institut teknologi sumatera.

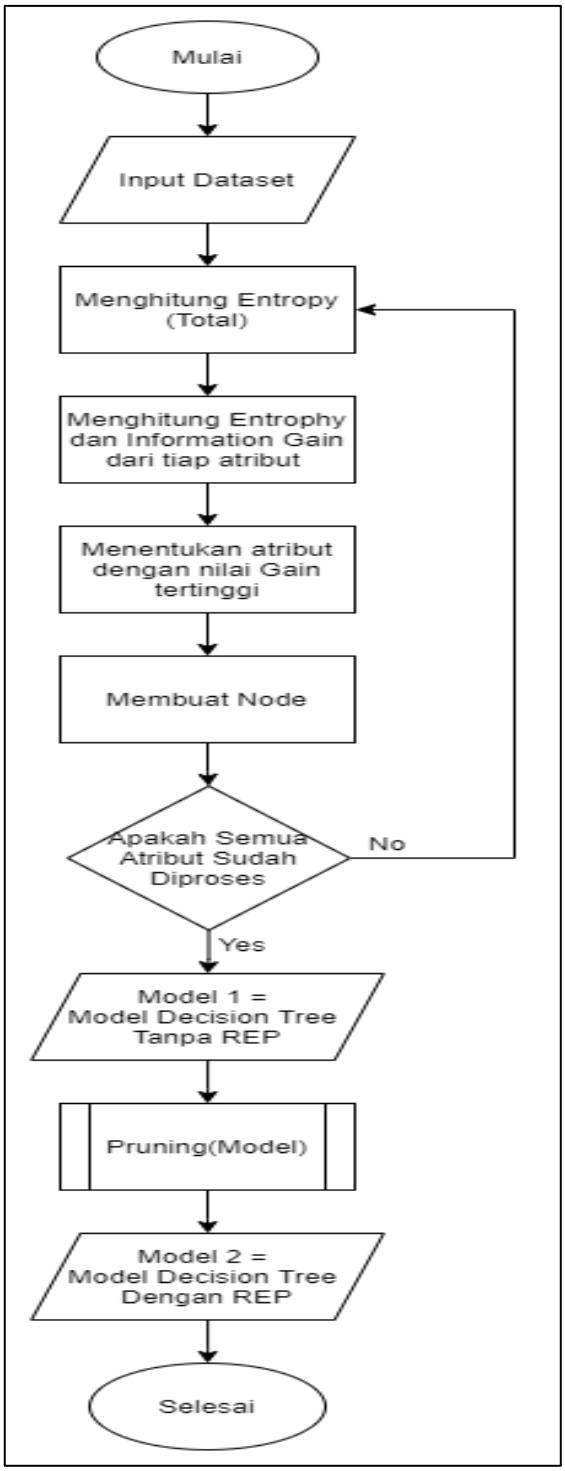
3.3.2 Data *Preparation*

Pada tahap sebelumnya yaitu luaran dari tahap 3.3.1 perlu diuji kualitasnya dengan 3 komponen yang menggambarkan kualitas data yaitu *accuracy*, *completeness* dan *consistency*. Dalam persiapan data, penulis melakukan 2 tahapan. Pertama yaitu memastikan data input berupa csv serta melakukan tahap data *preparation.* Adapun data preparation yaitu mempersiapkan data input sehingga kualitas data meningkat dan siap untuk dimodelkan. Beberapa tahapan pada data preparation yaitu data *integration*, data *cleaning*, data *reduction* dan data *transformation*.

Adapun tahap kedua data akan dibagi menjadi 3 yaitu data training, data validasi dan data test. Data training digunakan untuk membuat model. Data validasi untuk mencari konfigurasi terbaik dalam membuat model. Adapun data test digunakan untuk menilai tingkat akurasi dari sebuah model. Adapun pembagiannya data *training* 70%, *validation* 15% dan data *testing* 15%.

3.3.3 Pemodelan

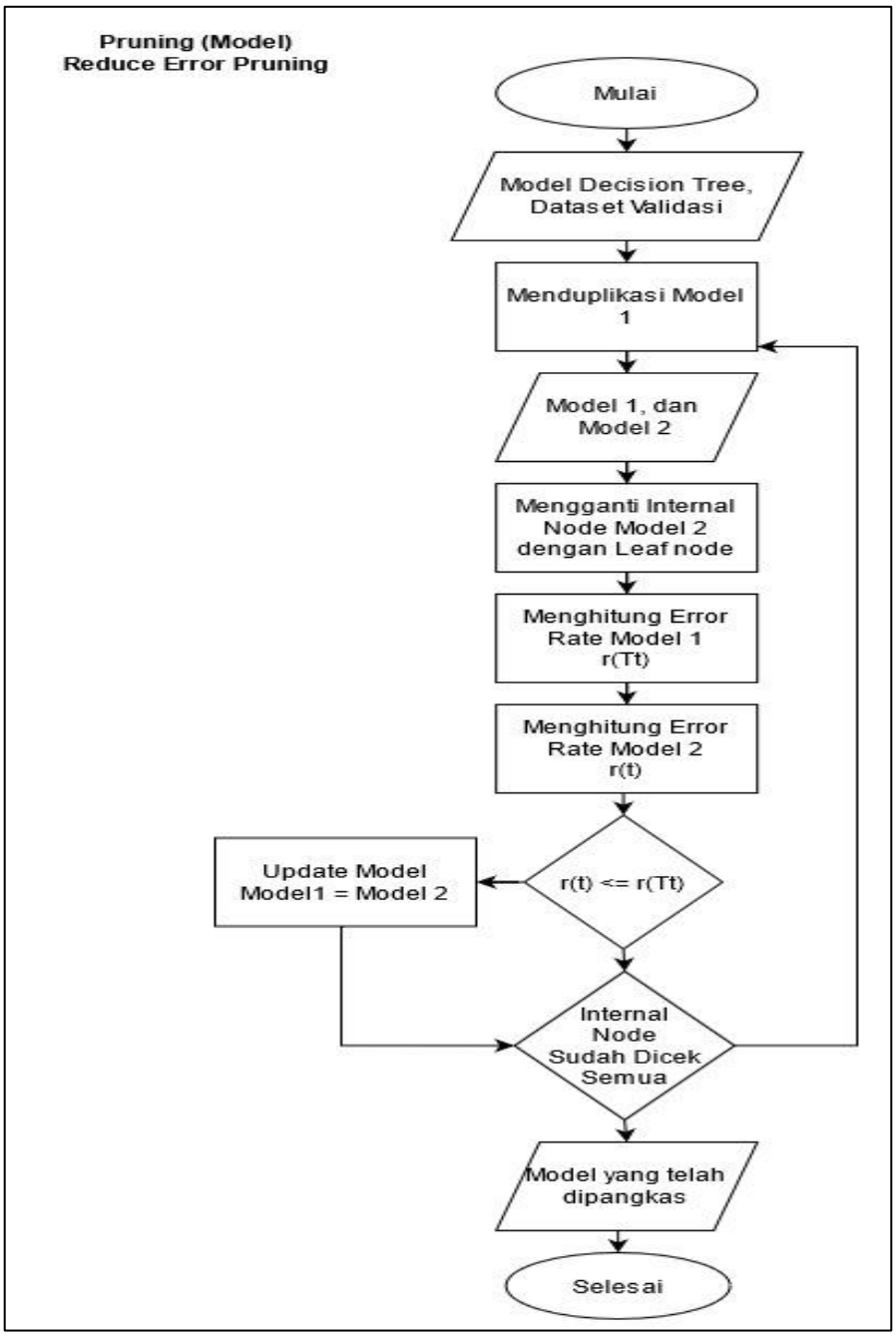
Pada tahap ini penulis mengimplementasikan algoritma *decision tree* C4.5. Adapun langkah langkahnya dapat dilihat pada **Gambar 3.2.** Pemodelan diawali dengan perhitungan *entrophy* total dengan menggunakan **Persamaan 2.3.**. Kemudian menghitung *Gain* dari tiap atribut menggunakan persamaan yang ada pada **Persamaan 2.1.** Setelah itu menentukan atribut untuk menjadi *node* pada *decision tree* dengan membandingkan *gain* dari tiap atribut menggunakan **Persamaan 2.2.** Melakukan hal yang sama hingga seluruh atribut diproses. Setelah didapatkan output berupa *decision tree* yang disebut sebagai model 1. Selanutnya *decision tree* yang telah diproses akan menjadi masukan pada proses REP. Hasil dari REP disebut sebagai model 2.



Gambar 3.2. Tahapan Pemodelan

Adapun tahapan dalam REP telah dijelaskan pada Sub-bab **2.4**. Dalam memudahkan dalam mengilustrasikan kerangka pemangkasan penulis menggambarkankan dengan diagram alir yang dapat dilihat pada **Gambar 3.3**.

Perhitungan *error rate* yang dimaksud dapat dilihat pada **Tabel 2.1** tepatnya pada **Persamaan 2.5.** Sehingga pada tahap pemodelan ini akan didapat 2 model yaitu model tanpa REP dan dengan REP.



Gambar 3.3. Tahap Pemangkasan REP

Tahapan pada REP membutuhkan inputan berupa model *decision tree* dan data untuk memvalidasi. Penulis menggunakan model yang telah dibangun atau

disebut sebagai model 1. Selanjutnya akan ada model 2 hasil duplikasi dari model 1. Pada model 2 akan memangkas sebuah *internal node* yang memiliki *leaf node*. Kemudian membandingkan dan menghitung *error rate* kedua model. Jika *error rate* model 2 lebih kecil atau sama dengan model 1 maka model 1 akan diupdate menjadi model 2. Sebaliknya, jika *error rate* model 2 lebih besar maka model 1 tidak akan diupdate menjadi model 2. Kemudian dilakukan hal yang sama sampai seluruh *internal node* dengan *leaf node* dicek.

3.3.4 *Evaluasi*

Pada tahap evaluasi penulis melakukan pengecekan akurasi dan *error rate* dari model dengan pendekatan *confusion matrix*. Akurasi menilai tingkat ketepatan dalam mengklasifikasikan data input. Error rate menilai tingkat ketidak tepatan model yang dibuat dalam mengklasifikasikan data input. Adapun langkah langkanya yaitu menguji model yang telah dibuat pada tahap 3.2.4 dengan data input berupa data test yang didapat dari sub-bab **3.2.2**. Setelah itu akan di hitung TP, TN, FN, dan FP nya. Kemudian akan menghitung *akurasi* dan *error rate* yang telah dijelaskan pada **Tabel 2.1**. Adapun tahapan evaluasinya sebagai berikut :

1. Menghitung TP, TN, FN dan FP pada model 1 dan model 2.
2. Menghitung Akurasi dan Error rate pada model 1 dan model 2.
3. Membandingkan akurasi dan *error rate* dari kedua model.

3.3.5 Analisis Hasil

Pada tahap ini penulis akan menganalisa model yang telah dilakukan evaluasi. Pada tahap ini penulis akan memberikan gambaran mengenai perbandingan model yang dikembangkan menggunkan C4.5 tanpa REP dan model yang dikembangkan menggunakan C4.5 dengan REP. Berdasarkan pada CRISP-DM penulis tidak melakukan tahap *deployment*. pada tahap ini penulis akan memberikan rekomendasi terkait *deployment* dari model yang telah dibuat.