# BAB II

# METODE PENELITIAN

## 2.1 Objek Penelitian

Objek yang diteliti pada penelitian ini adalah proses implementasi algoritme C4.5 untuk melakukan prediksi terhadap kelulusan mahasiswa. Dengan kata lain, objek penelitian ini adalah mahasiswa akhir yang sedang menempuh studi di Universitas Muhammadiyah Kalimantan Timur. Objek penelitian ini akan mencakup berbagai variabel atau fitur yang mungkin berpengaruh terhadap kelulusan seperti IPS, SKS, dan sebagainya. Dengan menerapkan algoritme C4.5, penelitian ini akan menghasilkan model prediktif yang dapat digunakan untuk memprediksi kelulusan mahasiswa berdasarkan variabel yang ada.

## 2.2 Alat dan Bahan

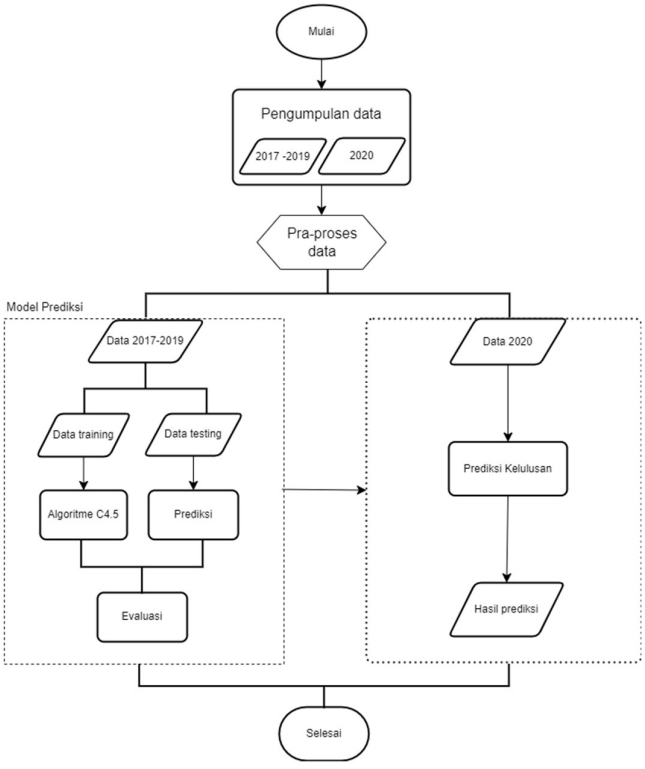
Alat dan bahan yang digunakan dalam penelitian ini meliputi :

1) Laptop Asus Vivobook RAM 4 GB dengan Processor Intel(R) Pentium(R) CPU 5405U 2.30GHz
2) Algoritme C4.5 sebagai algoritme perhitungan untuk menyelesaikan masalah prediksi status kelulusan mahasiswa.
3) *Python* sebagai bahasa pemrograman yang digunakan untuk analisis data, pemrosesan data dan membangun model prediktif.
4) *Google Colab* untuk membuat dan menjalankan kode *Python* secara interaktif, serta menyajikan hasil analisis data dan model prediktif dengan narasi yang terstruktur.
5) Data kelulusan dan data akademik mahasiswa 2017-2019 dan data akademik 2020 sebagai bahan atau informasi yang dianalisis dalam penelitian ini.

6) *Microsoft Excel* yang akan digunakan sebagai alat untuk mengolah data dalam bentuk tabel, melakukan *sorting data, merging data, cleaning data, dan filtering data*.

## 2.3 Prosedur Penelitian

Tujuan utama dari penelitian ini adalah untuk memprediksi kelulusan mahasiswa Program Studi Teknik Informatika Universitas Muhammadiyah Kalimantan Timur angkatan 2020 dengan menggunakan algoritme C4.5. Dalam rangka mencapai tujuan tersebut, maka diperlukan tahapan penelitian yang runtut dan jelas. Tahapan penelitian merupakan serangkaian langkah atau proses yang dilalui dalam melakukan perencanaan, pelaksanaan, menganalisis, dan menyajikan hasil penelitian dengan cara yang sistematis. Tahapan yang ada pada penelitian ini dapat dilihat pada Gambar 2.1




**Gambar 2.1** Alur Tahap Penelitian

Tahap pertama dalam penelitian ini adalah mengumpulkan data kelulusan dan data akademik mahasiswa Program Studi Teknik Informatika Universitas Muhammadiyah Kalimantan Timur angkatan 2017-2020. Selanjutnya melakukan *pre-processing* data sehingga nantinya data dapat di proses menggunakan algoritme C4.5. Berikutnya membuat model prediksi menggunakan data mahasiswa Teknik Informatika angkatan 2017- 2019 dan mengevaluasi model prediksi untuk menentukan apakah model tersebut efektif atau tidak berdasarkan akurasinya. Tahapan terakhir adalah melakukan prediksi kelulusan berdasarkan model prediksi menggunakan data mahasiswa Teknik Informatika angkatan 2020.

### 2.3.1 Pengumpulan Data

Pengumpulan data adalah proses atau langkah pertama yang dilakukan untuk mengumpulkan informasi atau fakta yang relevan untuk tujuan tertentu (Nawassyarif et al., 2020). Penelitian ini menggunakan data sekunder, dimana data sekunder adalah data yang diambil melalui perantara atau instansi. Contoh kasus, ketika akan meneliti terkait kelulusan mahasiswa maka yang diperlukan adalah data mahasiswa serta beberapa variabel pendukung lainnya seperti IPK, masa studi, dan beberapa variabel terkait lainnya (L. U. Khasanah, 2021). Penelitian ini menggunakan data kelulusan Program Studi Teknik Informatika serta data akademik mahasiswa angkatan 2017 -2020 dengan atribut/variabel IP Semester dan juga SKS dari semester 1 hingga semester 7.

### 2.3.2 Pra-proses data

Setelah melakukan pengumpulan data, selanjutnya adalah melakukan pra-proses data. Pra-proses adalah suatu tahapan awal dalam mengelola data untuk membersihkan dari elemen yang tidak digunakan, kemudian merapikan, dan mempersiapkan data sehingga dapat digunakan secara efektif untuk model analisis (Pane & Ramdan, 2022).

Dalam penelitian ini, data yang nantinya didapat dari Program Studi Teknik Informatika digabungkan untuk memperoleh hasil yang akurat. Adapun kegiatan Pra-proses yang dilakukan pada penelitian ini adalah penggabungan data dan pembersihan data yaitu menghapus atribut data yang bukan numerik dan menghapus beberapa atribut.

### 2.3.3 Data *training* dan data *testing*

Berikutnya adalah membagi data menjadi data *training* dan data *testing*. Data *training* adalah data yang digunakan untuk landasan untuk membuat model klasifikasi, dimana algoritme akan mempelajari pola pola dalam data tersebut dan membuat aturan keputusan. Sedangkan data *testing* adalah data yang akan digunakan untuk mengevaluasi seberapa baik model dapat menggeneralisasi pola-pola yang telah dipelajari dari data *training* ke data baru (Darwis et al., 2021). Pada penelitian ini data *training* yang digunakan sebanyak 70% dan data *testing* sebanyak 30% dimana komposisi ini memberikan hasil yang optimal pada beberapa penelitian, salah satunya pada penelitian (Dyah Ardyani Rizqi Azizah Adha et al., 2023).

### 2.3.4 Algoritme C4.5

Klasifikasi data dilakukan dengan menggunakan algoritme C4.5, dimana algoritme ini merupakan pengembangan dari algoritme ID3 oleh *J. Ross Quinlan*. Algoritme C4.5 merupakan metode yang digunakan untuk mengklasifikasikan data dengan menggunakan struktur pohon atau struktur berhierarki. Tahap pertama yang dilakukan dalam algoritme C4.5 adalah memilih atribut/variabel terbaik sebagai akar dengan menghitung *Entropy,* selanjutnya membuat cabang untuk setiap nilai di dalam *node root*. Proses ini dilakukan secara rekursif sampai semua kasus pada cabang mempunyai kelas yang serupa.

*Entropy* adalah parameter yang digunakan untuk menentukan *node* yang paling optimal dalam membagi data menjadi subset yang lebih kecil saat membangun pohon keputusan, semakin tinggi nilai *Entropy* maka akan semakin besar potensi untuk melakukan klasifikasi yang akurat. Persamaan yang digunakan untuk menghitung nilai *Entrop*y adalah sebagai berikut (Muslim et al., 2019, p. 50):

$$Entropy(S) = \sum_{i=1}^{n} - p_i * log_{2P_i} \quad (2.1)$$

Keterangan :

*S :* Himpunan kasus

*n :* Jumlah partisi *S*

$P_i$ *:* Proporsi dari $S_i$ terhadap *S*

Berikut adalah persamaan yang digunakan dalam menghitung $log_{2P_i}$ (Muslim et al., 2019, p. 50):

$$log(X) = \frac{ln(X)}{ln(2)} \quad (2.2)$$

Adapun kriteria yang paling umum digunakan untuk melakukan pemilihan fitur terbaik sebagai pembagi dalam algoritme C4.5 adalah *Gain Ratio*, formulasi *Gain Ratio* terdapat pada persamaan berikut (Muslim et al., 2019, pp. 50–51):

$$GainRatio = \frac{Gain(A)}{SplitInfo(A)} \quad (2.3)$$

Untuk menghitung *Gain* digunakan persamaan berikut (Muslim et al., 2019, p. 51):

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2.4)$$

Keterangan :

$S$ : Himpunan Kasus

$A$ : Atribut

$n$ : Jumlah partisi atribut A

$|Si|$ : Jumlah kasus pada partisi ke-i

$|S|$: Jumlah kasus dalam S

### 2.3.5 Evaluasi Model

Evaluasi model merupakan tahapan yang bertujuan untuk mengetahui model prediksi baik untuk digunakan atau tidak. Kinerja klasifikasi pada dataset tidak selalu mencapai 100% akurat, sehingga perlu dilakukan pengukuran terhadap model klasifikasi. Salah satu teknik yang dapat digunakan untuk mengevaluasi model adalah dengan mengukur tingkat akurasi. Adapun persamaan yang digunakan untuk melakukan pengukuran akurasi adalah sebagai berikut (Muslim et al., 2019, pp. 47–48) :

$$Accuracy = \frac{jumlah\ prediksi\ yang\ benar}{jumlah\ prediksi\ keseluruhan} \times 100 \quad (2.5)$$

### 2.3.6 Prediksi kelulusan

Langkah berikutnya adalah melakukan prediksi kelulusan terhadap mahasiswa Program Studi Teknik Informatika UMKT angkatan 2020. Menurut panduan akademik syarat kelulusan mahasiswa adalah dengan menempuh minimal 7 semester. Sehingga berdasarkan panduan tersebut seharusnya mahasiswa Program Studi Teknik Informatika angkatan 2020 akan lulus tahun ini . Dengan demikian, prediksi kelulusan dalam hal ini berarti memperkirakan apakah mahasiswa tersebut akan lulus tepat waktu atau sebaliknya.