# BAB III
# METODE PENELITIAN

## A. Jenis Penelitian

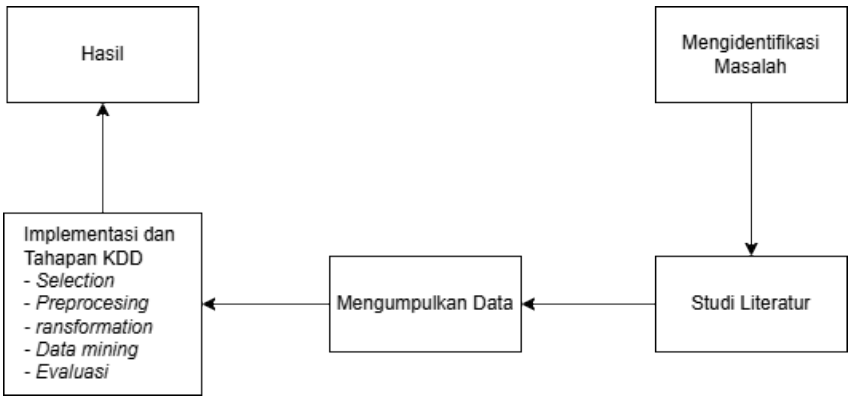
Penelitian ini menggunakan pendekatan kuantitatif yang merupakan jenis penelitian yang menekankan pada pengumpulan dan analisis data dalam bentuk numerik (Abdullah dkk., 2023). Pemilihan pendekatan kuantitatif ini di dasarkan dengan tujuan utama penelitian ini yaitu membandingkan akurasi dari dua algoritma dalam memprediksi yaitu, algoritma C4.5 dan *Naïve Bayes*. Pendekatan kuantitatif ini di anggap sesuai dengan konteks ini karena memungkinkan peneliti untuk mengumpulkan data numerik terkait karakteristik mahasiswa, seperti IPK, jumlah SKS, IPS dan seba*gain*ya, serta menganalisisnya secara statistik untuk menguji kinerja kedua algoritma.

Proses pengumpulan data dilakukan melalui studi literatur yang komprehensif, melibatkan berbagai sumber referensi akademik seperti buku-buku metodologi penelitian, jurnal-jurnal ilmiah terakreditasi, dan artikel-artikel penelitian yang relevan. Sumber-sumber ini dipilih dengan mempertimbangkan kredibilitas, relevansi, dan informasi terkini untuk memastikan kualitas hasil penelitian.



## B. Tahapan Penelitian

Pada penelitian ini terdapat metode yang harus di lakukan agar penelitian ini dapat memperoleh hasil dari tujuan penelitian. Tahapan penelitian ini merujuk pada suatu kegiatan penelitian yang di dalamnya terdapat proses yang dijalankan secara terstruktur, logis dan sistematis. Penelitian ini di lakukan dengan beberapa tahap seperti yang sudah di gambarkan pada gambar 3.1 di bawah ini:




*Gambar 3. 1* Tahapan Penelitian

Sumber :(Kamil & Cholil, 2020)

Berikut adalah tahapan pada penelitian ini yaitu:

1. Mendefinisikan masalah

Tahap awal dalam penelitian ini difokuskan pada pengkajian mendalam terhadap latar belakang permasalahan yang menjadi dasar dilakukannya penelitian prediksi kelulusan mahasiswa tepat waktu. Proses ini diawali dengan mengidentifikasi berbagai masalah yang terjadi di lingkungan akademik, khususnya terkait fenomena keterlambatan kelulusan mahasiswa. Melalui analisis komprehensif, ditemukan berbagai faktor yang berpotensi mempengaruhi durasi masa studi mahasiswa, serta dampaknya terhadap kualitas pendidikan dan efisiensi operasional institusi.

2. Studi literatur

Tinjauan literatur merupakan tahapan dalam penelitian ini, di mana dilakukan pengkajian secara sistematis terhadap berbagai penelitian terdahulu yang berkaitan dengan prediksi kelulusan mahasiswa dan juga perbandingan algoritma C4.5 dan *Naïve Bayes*. Kajian ini mencakup penelusuran terhadap jurnal-jurnal ilmiah dan publikasi akademik lainnya yang membahas tentang metode-metode prediksi kelulusan. Melalui tinjauan ini, diperoleh pemahaman komprehensif mengenai perkembangan terkini dalam bidang prediksi

kelulusan mahasiswa, termasuk berbagai pendekatan dan metodologi yang telah diterapkan oleh peneliti-peneliti sebelumnya.

3. Mengumpulkan Data

Untuk memperkuat pemahaman terhadap permasalahan yang ada, dilakukan pengumpulan data awal melalui studi dokumentasi terhadap data akademik mahasiswa dari periode-periode sebelumnya. Data mahasiswa yang dikumpulkan diambil langsung ke Lokasi penelitian yaitu fakultas Syari'ah UIN Imam Bonjol Padang dan bagian akademik UIN Imam Bonjol Padang. Penulis juga melakukan wawancara langsung pada pihak yang terkait seperti ketua Program studi, mahasiswa, dan tata usaha untuk menganalisis hasil wawancara yang di dapat agar menjadi sebuah data. Hasil analisis terhadap data tersebut menunjukkan adanya pola-pola tertentu yang dapat digunakan sebagai acuan dalam mengembangkan sistem prediksi.

4. Implementasi dan Tahapan KDD

Dalam tahap ini, tahap pertama dimulai dengan tahapan dalam KDD. Di mana data tersebut dilakukan *selection* terlebih dahulu di lanjut dengan *preprocesing, transformation*, data mining dan evaluasi.

a. selection data untuk memilih atribut yang di perlukan dari data set yang ada serta membuang atribut yang tidak di perlukan. Data di pilih meliputi IPK, SKS, IPS, serta status kelulusan mahasiswa kemudian dibagi menjadi data set (Nur dkk., 2024).
b. *preprocesing* data di lakukan untuk menghilangkan *noise*, *missing Values* dan data yang tidak konsisten.
c. *Transformation* data yang di lakukan untuk mengubah format data yang sudah di seleksi menjadi data yang sama dan akan di proses pada data mining.
d. Data Mining yang menerapkan metode yang di gunakan untuk membentuk pola pada data.

Metode yang di gunakan dalam tahap data mining ini menggunakan algoritma C4.5 dan *Naïve Bayes*. Secara umum tahapan algoritma C4.5 untuk membangun sebuah pohon Keputusan dengan memilih akar, kemudian setiap cabang di hitung nilainya, membagi cabang dan mengulangi semua proses sampai semua cabang memiliki kelas yang sama dan membentuk pohon Keputusan. Ada beberapa tahap dalam membuat sebuah pohon Keputusan yaitu (suriani uci, 2023):

1. Mempersiapkan data *training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah di kelompokan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon keputusan. Akar dipilih dengan menghitung *gain* dari setiap atribut, dan atribut yang memiliki *gain* tertinggi akan menjadi akar pertama dalam pohon keputusan. Sebelum menghitung *gain* persamaan (2), penting untuk terlebih dahulu menghitung nilai *entropy*. Berikut adalah persamaan (1) yang digunakan untuk perhitungan nilai *entropy*.

   *Entropy* (i) = $\sum_{i=1}^{n} -P_i \log_2 P_i$ (1)

   Keterangan :

   S= Himpunan kasus

   n = Jumlah partisi S

   Pi = probabilitas Si terhadap S
3. Hitung nilai *gain* dengan rumus:

   *Gain (S.A)* = $Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$ *(2)*

   S = Himpunan Kasus

   A = Atribut

   n = Jumlah partisi atribut A

   |Si| = Jumlah kasus pada partisi ke-i

   |S| = Jumlah kasus dalam S.
4. Mengulangi langkah ke-2 hingga semua *record* terpartisi
5. Proses partisi pohon keputusan akan berhenti jika:

a) Semua tupel pada *node* N memiliki kelas yang sama
b) Tidak ada atribut dalam tupel yang akan dipartisi lebih lanjut
c) Tidak ada tupel di dalam cabang yang kosong.

*Naïve Bayes* adalah suatu model independen yang membahas mengenai klasifikasi sederhana yang bisa memprediksi probabilitas sebuah *class* yang berdasarkan *Teorema Bayes*. *Naïve Bayes* merupakan suatu algoritma yang dapat mengklasifikasikan suatu variabel tertentu dengan menggunakan metode probabilitas dan statistik. Secara garis besar algoritma dijelaskan seperti persamaan berikut:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

Di mana:

- X adalah kriteria suatu kasus berdasarkan bukti
- H adalah Hipotesis data kelas.
- P(H|X) adalah Probabilitas Hipotesis H terhadap kondisi X (posterior)
- P(X|H) adalah Probabilitas X terhadap kondisi hipotesis H (*likelihood*)
- P(H) adalah Probabilitas Hipotesis H (*prior*)
- P(X) adalah probabilitas X (*evidence*)



Adapun tahapan tahapan menghitung algoritma *Naïve Bayes*

- Mengumpulkan data *training* yang sudah dilabeli kelas/kategorinya.
- Menghitung probabilitas *prior* (*prior probability*) dari masing-masing kelas.
- Menghitung probabilitas bersyarat (*conditional probability*) dari fitur-fitur pada masing-masing kelas.
- Ketika ada data baru yang ingin diklasifikasikan, algoritmanya akan menghitung probabilitas *posterior* (*posterior probability*) dari

data baru tersebut pada masing-masing kelas, menggunakan rumus *Bayes*. Lalu, memilih kelas dengan probabilitas posterior tertinggi sebagai prediksi untuk data baru tersebut.

setelah itu tahapan selanjutnya yaitu melakukan validasi silang menggunakan *cross validation*. Dalam kasus ini, total data set 509 data. Data set ini kemudian akan digunakan dalam proses *K-fold cross validation* dengan k= 5 dan 10, yang berarti data akan dibagi menjadi 5 dan 10.

e. Evaluasi

Tahapan terakhir adalah evaluasi yang membandingkan metode C4.5 dan *Naïve Bayes* untuk memprediksi kelulusan mahasiswa tepat waktu menggunakan *K-fold cross validation*. Setiap iterasi menghasilkan *Confusion matrix* yang mengidentifikasi *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Indikator kinerja yang diukur meliputi *accuracy*, *precision*, *recall*, dan F1-*Score*.

5. Hasil

Hasil kedua algoritma dibandingkan melalui tabel perbandingan dan perhitungan rata-rata metrik untuk mengidentifikasi kelebihan serta kekurangan setiap metode. Perbandingan metrik kedua model divisualisasikan untuk mempermudah interpretasi. Kesimpulan menunjukkan model yang lebih unggul, kondisi optimalnya, serta rekomendasi implementasi untuk prediksi kelulusan mahasiswa

## C. *Setting* Penelitian

1. Lokasi Penelitian

penelitian ini di lakukan pada program studi hukum keluarga pada UIN Imam Bonjol Padang. Pengambilan dan pengumpulan data yang dibutuhkan untuk penelitian yang di lakukan oleh penulis di lakukan dari bulan pertama hingga selesai.

2. Jadwal Penelitian

Adapun jadwal penelitian yang di rencanakan yang berlangsung selama 5 bulan, dapat di lihat pada tabel di bawah ini:

Tabel 3. 1 Jadwal Penelitian

| Waktu / Kegiatan | Bulan 1 | | | | Bulan 2 | | | | Bulan 3 | | | | Bulan 4 | | | | Bulan 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Mengidentifikasi Masalah | | | | | | | | | | | | | | | | | | | | |
| Studi Literatur | | | | | | | | | | | | | | | | | | | | |
| Mengumpulkan data | | | | | | | | | | | | | | | | | | | | |
| Implementasi | | | | | | | | | | | | | | | | | | | | |
| Evaluasi | | | | | | | | | | | | | | | | | | | | |
| Membuat laporan | | | | | | | | | | | | | | | | | | | | |

## D. Bahan Atau Materi Penelitian

Dalam penelitian ini menggunakan data set sebanyak 509 data. Penelitian ini menggunakan metode *K-fold cross validation* dengan k=10 untuk membagi data menjadi beberapa sub set. Pada data set yang berjumlah 509 dibagi menjadi 10 *fold* dengan rincian *fold* 1 sampai *fold* 9 masing-masing berisi 51 data *fold* 10 berisi 50 data. Sebagai contoh pada iterasi pertama, *fold* 1 digunakan sebagai data testing sedangkan *fold* 2 sampai 10 digunakan sebagai data *training*, kemudian pada iterasi kedua *fold* 2 digunakan sebagai data testing sedangkan *fold* 1, 3, 4, sampai 10 digunakan sebagai data *training*, begitu seterusnya hingga semua *fold* mendapat giliran sebagai data validasi.

Dan juga menggunakan k=5 dengan membagi data menjadi beberapa sub set. Data set di bagi menjadi 5 *fold* dengan rincian *fold* 1 sampai *fold* 4 masing-masing berisi 102 data, dan *fold* 5 berisi 101 data. Sebagai contoh pada iterasi

pertama, *fold* 1 digunakan sebagai data testing sedangkan *fold* 2 sampai 5 digunakan sebagai data *training*, kemudian pada iterasi kedua *fold* 2 digunakan sebagai data testing sedangkan *fold* 1, 3, 4, sampai 5 digunakan sebagai data *training*, begitu seterusnya hingga semua *fold* mendapat giliran sebagai data validasi

Penggunaan 5 dan 10-*fold cross validation* juga membantu memastikan model konsisten dengan berbagai kombinasi data, mengurangi risiko *overfitting*, dan memberikan evaluasi yang lebih handal. Model akan dilatih menggunakan data set, kemudian performa kedua model akan dievaluasi pada setiap *fold*.

Tabel 3. 2 data mahasiswa tahun 2018, 2019 dan 2020

| **Tahun Masuk** | **Program studi Hukum Keluarga** | |
|---|---|---|
| | Jumlah Mahasiswa | Lulus Tepat Waktu |
| 2018 | 146 | 61 |
| 2019 | 169 | 109 |
| 2020 | 133 | 88 |
| jumlah | 448 | 258 |



## E. Teknik Pengumpulan Data

Pada penelitian ini menggunakan Teknik pengumpulan data primer. Di mana data di peroleh langsung dari sumbernya yaitu akademik mahasiswa UIN Imam Bonjol Padang. Yang di berikan langsung oleh Bagian Akademik dan Kemahasiswaan UIN Imam Bonjol Padang. Juga melakukan wawancara langsung kepada pihak yang berkaitan untuk menanyakan kendala-kendala yang di alami sehingga terjadinya lulus tidak tepat waktu.