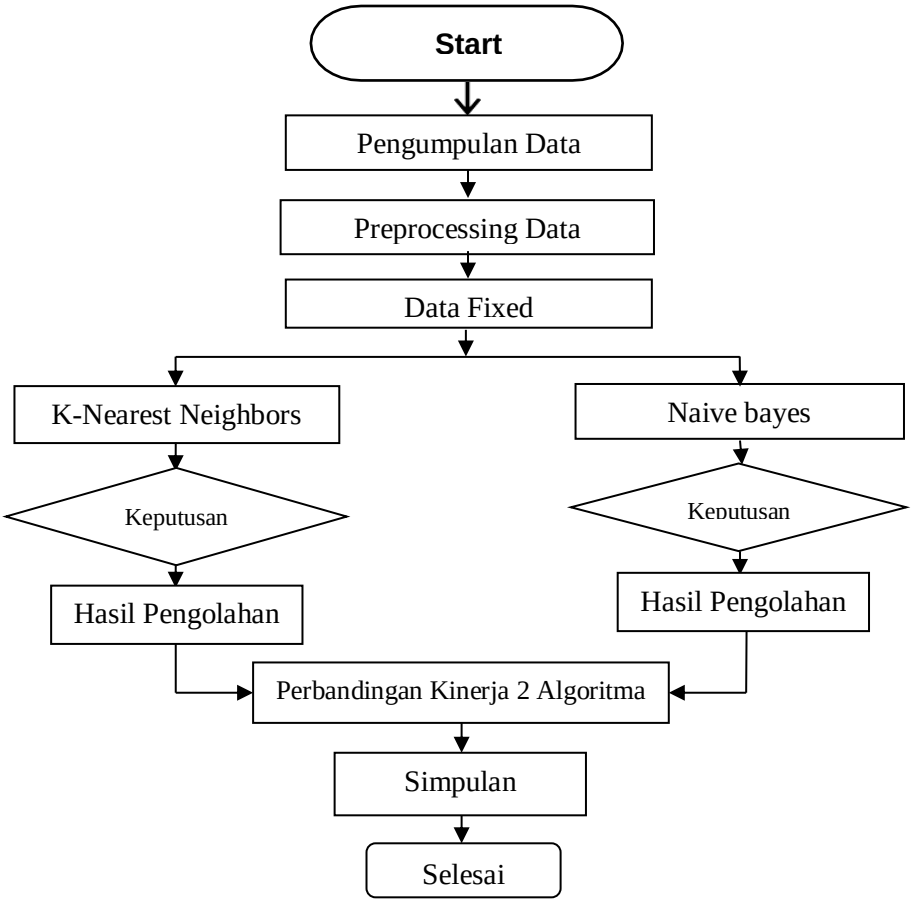
BAB III

METODOLOGI PENELITIAN

Penelitian ini akan dilakukan dengan menerapkan dua metode yaitu Algoritma *K-Nearest Neighbour* dan *Naive Bayes* untuk kelulusan mahasiswa Tepat Waktu dan Tidak Tepat Waktu. Alur tahapan penilitian yang dibuat sebagai kerangka kerja kelulusan mahasiswa dapat dilihat seperti yang ditunjukan pada gambar 3.1 dibawah ini.



***Gambar* 3.1. Diagram Alur**

Berdasarkan Gambar 3.3 dapat dijelaskan tahapan penelitian sebagai berikut:

1) Analisis Data

Pada tahap ini peneliti melakukan analisis data dengan membaca buku, jurnal, dan laporan penelitian yang terkait dengan topik penelitian. Kemudian melakukan pengumpulan data dan informasi seperti melakukan wawancara, dokumentasi dan *observasi* untuk mengambil kebutuhan data. Selanjutnya dilakukan proses definisi kebutuhan dengan malakukan identifikasi data yang dibutuhkan, melihat prosedur yang sedang berjalan, dan menganalisis sistem yang sedang berjalan.

2) Pengumpulan Data

Proses pengujian, data dibagi menjadi dua bagian yaitu data *training* dan data *testing* dengan menggunakan algoritma *K-Nearest Neighbor* dan algoritma *Naive Bayes*. Data *training* digunakan untuk membentuk tabel probabilitas dan data *testing* digunakan untuk menguji probabilitas yang telah terbentuk. Pada tahap ini peneliti melakukan pengambilan data kelulusan mahasiswa pada Sekolah Tinggi Teknologi Nusantara Lampung. Data tersebut akan diolah pada tahap pemrosesan data. Atribut yang digunakan dalam memprediksi kelulusan Tepat Waktu dan Tidak Tepat Waktu pada penelitian ini adalah Nomor Induk Mahasiswa (NIM), Nama Mahasiswa, Indeks Prestasi Kumulatif (IPK), Unit Kegiatan Mahasiswa(UKM), Penghasilan Orang Tua dan Test Potensi Akademik, dan kelulusan Tepat Waktu dan Tidak Tepat Waktu.

3) *Preprocessing* Data

Tahap ini peneliti melakukan mengubah data mentah atau biasa dikenal dengan *raw* data yang dikumpulkan dari berbagai sumber menjadi informasi yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya.

4) *K-Nearest Neighbor*s

Algoritma *K-Nearest Neighbors* pada penelitian digunakan untuk menguji dalam menghasilkan hasil pengolahan.

5) *Naïve Bayes*

Pada tahap ini penelitian menggunakan Algoritma *Naïve Bayes* untuk diuji dalam menghasilkan hasil pengolahan.

6) Perbandingan Kinerja 2 Algoritma

Tahap ini peneliti menguji dua Algoritma *Naïve Bayes* dan *K-Nearest Neighbor*s untuk mendapatkan simpulan terhadap prediksi kelulusan mahasiswa Sekolah Tinggi Teknologi Nusantara Lampung.

7) Simpulan

Pada tahap ini peneliti mengambil simpulan terhadap kinerja dua algoritma *Naïve Bayes* dan *K-Nearest Neighbor*s pada kelulusan mahasiswa Sekolah Tinggi Teknologi Nusantara Lampung.

3.1. Metode Penelitian

Data *training* dengan menggunakan pelatihan algoritma sehingga menghasilkan sejumlah aturan dilakukan pada tahapan pemodelan ini (Sabna dan Muhardi,2016). Algoritma *K-Nearest Neighbor* dan *Naive Bayes* pada penelitian ini digunakan dalam melakukan perbandingan.

3.1.1. Metode *K-Nearest Neighbor*

Salah satu metode klasifikasi dalam data mining yang termasuk ke dalam *supervised learning* adaah *K-Nearest Neighbor*. Atribut dan data *training* dilakukan pengklasifikasian, sehingga data baru dilakukan proses klasifikasi berdasarkan perbandingan data *training* pada kesamaan mayoritas. Nilai jarak ditentukan dengan pengujian data *testing* terhadap data *training* dengan menggunakan nilai terkecil dari nilai ketetanggan terdekat pada *K-Nearest Neighbor* (Krisandi et al.,2015). *Euclidean Distance* atau penghitungan jarak umumnya menggunakan jarak sebagai berikut:

$$\mathbf{Dxy} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Di mana :

D = Jarak Kedekatan

x = Data Training

y = Data Testing

i = Fitur ke –

n = Jumlah Fitur

Kesamaan data bisa lebih dari satu, oleh karena itu *K- Nearest Neighbour* dapat mengklasifikasikan sejumlah k data yang serupa juga merupakan data yang memiliki banyak kemiripan. Jarak digunakan dalam *K- Nearest Neighbour* disebut jarak Euclidean. Selain mudah ditemukan jarak terpendek antar data, keuntungan dari algoritma ini adalah dapat menggeneralisasi data pelatihan relatif kecil

(Satriya & Santoso, 2018), Adapun langkah-langkahnya untuk menghitung metode *K- Nearest Neighbour* adalah:

a) Penentuan awal k . parameter

b) Tentukan jarak antara data yang akan dievaluasi

c) Urutkan jarak yang didapat

d) Tentukan jarak terdekat dengan orde nilai k

e) Pasangkan dengan kelas yang sesuai

f) Temukan jumlah kelas di mana mayoritas dan dapat diprediksi adalah kategori obyek.

3.1.2. Metode *Naïve Bayes*

Algoritma yang berdasarkan *teorema Bayes* atau *Naive Bayes*, dimana atribut bersifat saling bebas sehingga antar atributnya tidak memiliki hubungan atau ketergantungan. Metode klasifikasi yang menghitung probabilitas suatu kejadian berdasarkan kondisi tertentu merupakan klasifikasi *Naïve Bayes*, berikut *Naïve Bayes* dalam persamaan (Lukito & Chrismanto, 2015).

Naive Bayes adalah algoritma klasifikasi yang cukup sederhana dan mudah diimplementasikan sehingga algoritma ini sangat efektif ketika diuji dengan data set yang benar, terutama jika *Naive Bayes* dikombinasikan dengan seleksi fungsi, sehingga *Naive Bayes* dapat mengurangi redundansi dalam data, selain itu *Naive Bayes* menunjukkan hasil yang baik bila dikombinasikan dengan metode kekelompokan. *Naive Bayes* terbukti memiliki akurasi yang tinggi dibandingkan dengan mendukung mesin vektor (Apandi dan Sugianto, 2019). *Naive Bayes Classifier* termasuk dalam pembelajaran yang diawasi, *Naive Bayes*

memperkirakan probabilitas kelas bersyarat dengan asumsi bahwa atribut bebas bersyarat yang diberikan oleh label y, Asumsi *independensi* bersyarat dapat dinyatakan dalam bentuk berikut (Suyanto, 2017). Perhitungan *Naive Bayes* dengan persamaan berikut.

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

Dimana :

X = Data dengan kelas yang tidak diketahui (bukti)

H = Data hipotesis X adalah spesifikasi kelas

P(H|X) = Probabilitas hipotesis H benar untuk kondisi X (prob posterior)

P(H) = Probabilitas hipotesis H (prob. sebelumnya)

P(X) = Probabilitas sebelum pembuktian X

Adapun ciri-ciri *Naive Bayes* sebagai berikut :

1) Metode *Naïve Bayes* kuat *(robust)* pada data terisolasi yang biasanya data dengan karakteristik yang berbeda *(outliner)*. *Naive Bayes* juga dapat menangani nilai atribut yang salah dengan mengabaikan data pelatihan selama pembuatan model dan proses prediksi.
2) Tangguh dalam menghadapi atribut yang tidak *relevan*.
3) Atribut yang memiliki korelasi bias menurunkan kinerja klasifikasi *Naive Bayes*, karena asumsi independensi atribut tidak ada lagi.
4) Algoritma *Naive Bayes* memiliki kelebihan yaitu relatif mudah untuk diimplementasikan karena tidak menggunakan optimasi numerik, perhitungan matriks dan lain-lain, Efisien dalam pelatihan dan penggunaan,

dapat menggunakan data biner atau polinomial, karena diasumsikan independen itu mungkin metode ini diimplementasikan dengan berbagai kumpulan data, akurasi relatif tinggi.

3.2. Teknik Pengumpulan Data

Data yang digunakan untuk penelitian ini adalah data Badan Administratif dan Kemahasiswaan (BAAK) Sekolah Tinggi Teknologi Nusantara Lampung atau STTN Lampung yang di dapat dari angkatan 2013 sampai 2017, data tersebut merupakan data mahasiswa yang sudah lulus yang akan dijadikan sebagai data *training*, sedangkan pada data *testing* akan digunakan angkatan 2018 S1 Teknik Elektro dan S1 Teknik Industri. Pada penelitian ini, data *training* dan data *testing* yang digunakan merupakan data dengan format.csv yang terdiri dari 917 data mahasiswa yang lulus angkatan 2013 sampai angkatan 2017, Jumlah data terdiri dari masing-masing 378 data mahasiswa S1 Teknik Elektro dan 539 data mahasiswa S1 Teknik Industri dan pada data *testing* menggunakan mahasiswa angkatan 2018 yang berjumlah 100 data mahasiswa yang terdiri dari 40 S1 Teknik Elektro dan 60 S1 Teknik Industri dengan variable yang berisi, yaitu : Nomor Induk Mahasiswa (NIM), Nama Mahasiswa, Jenis Kelamin (JK), Tempat Tanggal Lahir, Alamat, Indeks Prestasi Kumulatif (IPK), Unit Kegiatan Mahasiswa (UKM), Penghasilan Orang Tua, Test Potensi Akademik dan Keterangan Lulus Tepat Waktu dan Tidak Tepat Waktu Mahasiswa yang digunakan untuk variabel utamanya adalah keterangan kelulusan mahasiswa.

Tahap studi literatur ini mempelajari tentang semua data dan informasi yang berkaitan dengan Algoritma *K-Nearest Neighbor* dan *Naive Bayes* juga

semua materi yang berhubungan dengan masalah yang akan dibahas, dalam penelitian ini referensi diambil dari berbagai sumber, seperti buku, jurnal, *e-book*, kajian pustaka, wawancara serta sumber-sumber lain yang dinilai dapat memberi tambahan wawasan untuk penelitian ini

3.3. Waktu dan Tempat Penelitian

Waktu yang digunakan pada pelaksanaan penelitian ini adalah tahun ajaran 2021/2022 ganjil pada Sekolah Tinggi Teknologi Nusantara Lampung yang beralamat pada Jalan Pulau Damar Gg. Sapta Marga Kelurahan Waydadi Baru Kecamatan Sukarame Kota Bandar Lampung.

3.4. Tahapan Penelitian

Tahapan yang digunakan dalam penelitian menggunakan model standarisasi data mining yaitu *Cross Industry Standart Process for Data Mining* (CRISP-DM) dengan langkah-langkah sebagai berikut:

1) Fase Pemahaman Bisnis *(Bussiness Understanding Phase)*

Pada tahap ini berfokus pada tujuan penelitian yaitu untuk mengetahui algoritma terbaik untuk kelulusan mahasiswa Sekolah Tinggi Teknologi Nusantara Lampung dengan mencari data-data yang berkaitan pada kelulusan mahasiswa tahun akademik 2013 sampai 2017 dimana mahasiswa merupakan mahasiswa yang sudah lulus yang dijadikan sebagai data training dan 2018 merupakan mahasiswa aktif yang akan dijadikan sebagai data testing, dengan tujuan mendapatkan model terbaik untuk memenuhi dari tujuan penelitian.

2) Fase Pemahaman Data *(Data Understanding Phase)*

Data yang digunakan dalam penelitian merupakan data dari hasil Bagian Akademik Administratif Kemahasiswaan (BAAK) yang didapatkan pada Sekolah Tinggi Teknologi Nusantara Lampung. Beberapa atribut yang akan diuji dapat dilihat pada tabel 3.1 dibawah ini.

Tabel 3.1. Fase Pemahaman Data

No	Kriteria	Keterangan
1	Nama	Nama Mahasiswa
2	NIM	Nomor Induk Mahasiswa
3	IPK	Indeks Prestasi Kumulatif
4	UKM	Unit Kegiatan Mahasiswa
5	Penghasilan Orang Tua	Penghasilan Orang Tua
6	TPA	Test Potensi Akademik
7	Kelulusan	Masa Studi Mahasiswa

3) Data *Intergation* dan Data Informasi

Untuk memudahkan dalam proses maka kriteria pada tabel 3.1 diuraikan dalam tabel 3.2 sebagai berikut :

Tabel 3.2. Data *Intergation* dan Data Informasi

No	Kriteria	Skala	Keterangan
1	Nama	Text	Nama Mahasiswa
2	NIM	Polynominal	Nomor Induk Mahasiswa
3	IPK	Polynominal	> 3,00 = Tinggi 2,70 - 3,00 = Sedang < 2,70 = Rendah
4	UKM	Polynominal	< 1 Tinggi 1 - 2 Sedang > 2 Rendah
5	Penghasilan Orang Tua	Polynominal	> 3.500.000 = Tinggi 2.000.000 - 3.500.000 = Sedang < 2.000.000 =Rendah

6	TPA	Polynominal	> 80 = Tinggi 60 - 80 = Sedang < 60 = Rendah
7	Kelulusan	Binominal	Tepat Waktu Tidak Tepat Waktu

Kriteria-kriteria pada tabel 3.2 untuk nantinya akan digunakan sebagai kriteria dalam menentukan Kelulusan Mahasiswa Sekolah Tinggi Teknologi Nusantara Lampung.

4) Pemodelan *(Modeling Phase)*

Algoritma yang digunakan dalam penelitian ini yaitu algoritma *K-Nearest Neighbour* dan *Naïve Bayes* untuk mengklasifikasikan dalam memperkirakan kelulusan mahasiswa dan untuk memperoleh sebuah model atau fungsi untuk menggambarkan kelas kelulusan tepat waktu dan tidak tepat waktu menggunakan algoritma *K-Nearest Neighbour* dan *Naive Bayes*.

5) Fase Evaluasi *(Evaluation Phase)*

Pada tahap ini dilakukan evaluasi kinerja dari kedua algoritma yaitu Algoritma *K-Nearest Neighbour* dan *Naïve Bayes* dengan membandingkan hasil nilai ratarata akurasi, *recall*, *Precission* dan *error rate* yang terdapat pada tabel *confusion matrix*.

3.5. *Cross Validation*

Cross validation adalah suatu metode tambahan dari teknik data mining yang bertujuan untuk memperoleh hasil akurasi yang maksimal. Metode ini sering juga disebut dengan K-*fold cross validation* dimana percobaan sebanyak K kali

untuk satu model dengan parameter yang sama. Secara umum, kita akan membandingkan n model dalam *cross validation* ini, dalam arti lain fungsi dari penggunaan metode *cross validation* adalah untuk mengetahui *performa* dari suatu model algoritma dengan melakukan percobaan sebanyak K kali untuk menningkatkan tingkat performansi dari model tersebut dan mengolah data set dengan kelas yang seimbang. Dalam kasus klasifikasi, ada yang perlu diperhatikan dalam pembagian set data ke sejumlah K partisi, yaitu harus melakukan *stratification* yang artinya kita akan mempartisi atau membagi set data tersebut ke K partisi denga komposisi kelas yang seimbang disetiap partisinya. Dengan kata lain, distribusi kelas setiap partisi harus sama antar kelas, yang berarti juga sama dengan distribusi kelas di set data originalnya.

3.6. Evaluasi *(Evaluation)*

Cross validation dari algoritma yang digunakan untuk klasifikasi, dilakukan pengujian pada tahapan evaluasi. Nilai akurasi, presisi dan *recall* dari algoritma *K-Nearest Neighbor*, dan *Naive Bayes* dibandingkan pada hasil pengujian tersebut. Agar dapat melihat algoritma yang terbaik pada pengujian dilakukan dengan metode *T-test*.

1) *Cross Validation*

Sebuah pengujian standar yang berfungsi untuk memprediksi *error rate* disebut *cross validation*. Setiap kelas harus memiliki jumlah yang sama yang mewakili jumlah data *training* dan data *testing*, perbandingan yang sama pada setiap kelasnya dilakukan pembagian data secara acak. Tingkat

kesalahan (*error rate)* keseluruhan pada setiap tingkat *iterasi* dihitung rata-ratanya (Hastuti, 2012).

2) *T-Test*

Sampel yang berasal dari objek yang memilikki kesamaan dibagi menjadi dua data. Metode pengujian hipotesis yang memperlakukan satu individu (objek penelitian) dengan dua perlakuan berbeda disebut *T-test*. Dengan membandingkan kondisi objek pertama dan kondisi objek kedua nilai *performance* didapatkan (Hastuti, 2012). Pada *T-test* rumus perhitungan adalah sebagai berikut (Kadafi, 2018).

$$T_{hitung} = \frac{X - \mu_0}{\frac{\sigma}{sn}}$$

Nilai T yang dicari dan menunjukan standar deviasi pada distribusi normal (tabel t) disebut T_{hitung}, nilai rata-rata dari data yang diolah adalah X, rata-rata nilai yang menjadi sampel adalah μ_0, standar *devisiasi* dari populasi yang telah diketahui adalah s, kemudian jumlah populasi yang digunakan dalam penelitian adalah n. Jika Ho ditolak maka nilai uji T lebih kecil daripada a (Huda, 2013).