PENERAPAN FEATURE SELECTION INFORMATION GAIN RATIO PADA ALGORITMA NAIVE BAYES UNTUK PREDIKSI KELULUSAN MAHASISWA

(STUDI KASUS : TEKNIK INFORMATIKA UIN SUSKA RIAU)

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat

Untuk Memperoleh Gelar Sarjana Teknik

Pada Jurusan Teknik Informatika

Oleh

MUHAMMAD FAUZAN WIJANARKO

11651103693

UIN SUSKA RIAU

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU

PEKANBARU

2020

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perguruan tinggi adalah satuan penyelenggara pendidikan tinggi yang merupakan tingkat lanjutan dari jenjang pendidikan menengah di jalur pendidikan formal. Hal ini sesuai dengan pengertian perguruan tinggi menurut (UU No. 20 Tentang Sistem Pendidikan Nasional, 2003) yang menyatakan bahwa : perguruan tinggi merupakan jenjang pendidikan setelah pendidikan menengah mencakup program pendidikan diploma, sarjana, magister, spesialis, dan doktor yang diselenggarakan oleh perguruan tinggi. Perguruan tinggi juga perlu mendeteksi perilaku mahasiswa, sehingga dapat diketahui faktor yang menyebabkan kegagalan seorang mahasiswa untuk lulus atau lulus dengan masa studi yang telah ditetapkan, seperti rendahnya kemampuan akademik, usia masuk, indeks prestasi maupun faktor-faktor lainnya (Romadhona, A., Suprapedi , S. dan Himawan, 2017).

Menurut (BAN-PT, 2019) salah satu indikator yang menjadi tolak ukur keberhasilan perguruan tinggi dalam melakukan proses kegiatan belajar mengajar (KBM) adalah angka kelulusan. Angka kelulusan yang tinggi dianggap sebagai sebuah prestasi pada perguruan tinggi yang bersangkutan. Angka kelulusan yang tinggi bisa digunakan sebagai bahan promosi untuk menarik minat calon mahasiswa baru. Jika perguruan tinggi dapat mencapai tingkat kelulusan 100% maka dapat dikatakan perguruan tinggi tersebut sukses dan berhasil.

Setiap universitas memiliki standar kelulusannya masing-masing, tergantung pada kebijakan dan standarisasi dari masing-masing universitas. Keterlambatan kelulusan mahasiswa merupakan masalah yang sering dihadapi oleh setiap universitas. UIN SUSKA RIAU juga tidak luput dari masalah tersebut, salah satu jurusan yang terdapat di UIN SUSKA RIAU adalah teknik informatika.

Tabel 1.1 Data Kelulusan Mahasiswa

Sumber	Hasil		
PTIPD UNIVESITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU	Tahun	Masuk	Lulus
	2019	247	204
	2018	146	177
	2017	212	126
	2016	175	29

Tabel diatas menyatakan data empat tahun terakhir jumlah kelulusan di jurusan Teknik Informatika. Dapat dilihat setiap tahunnya mengalami peningkatan kelulusan. Tetapi, jika di kalkulasikan jumlah mahasiswa yang masuk dalam empat tahun terakhir berjumlah 780 tidak sama dengan jumlah mahasiswa yang lulus dengan total 536 mahasiswa. Timbulnya permasalahan ini menuntut pihak jurusan untuk memiliki keunggulan dalam hal pemanfaatan sumber daya sarana, prasarana dan manusia. Dengan bantuan sistem informasi untuk menunjang kegiatan pengambilan keputusan yang memanfaatkan gudang data, diperlukan juga analisis data untuk menggali informasi yang tersedia.

Terkait kelulusan mahasiswa telah banyak penelitian yang dilakukan, salah satunya pada penelitian (Imaslihkah et al., 2013) tentang faktor-faktor yang mempengaruhi predikat kelulusan mahasiswa menggunakan analisis *Regresi Logistik*. Pengujian dilakukan menggunakan dua cara yaitu pengujian secara serentak dan pengujian secara individu. Pengujian secara serentak, faktor yang berpengaruh antara lain jalur penerimaan, fakultas, pekerjaan orang tua, jenis kelamin dan pendapatan orang tua. Untuk pengujian individu, faktor yang berpengaruh seperti fakultas, jalur penerimaan, pekerjaan orang tua dan pendapatan. Akurasi yang diperoleh dari model pengujian serentak yaitu sebesar 77,41% yang dirasa sudah cukup baik. Lalu penelitian (Sulistio, 2017) pengimplementasian prediksi kelulusan mahasiswa menggunakan metode *discriminant analysis* berbasis web. Metode *discriminant analysis* ini mengklasifikasikan suatu objek dari objek lain menuju kelas masing-masing.

Objek akan dianggap sebagai siswa, jadi metode ini akan memisahkan sekelompok siswa yang nantinya akan di letakkan pada kelas masing-masing. Hasil yang diperoleh dengan menggunakan 100 data *testing* tejadi kegagalan prediksi sebanyak 7, sehingga dapat disimpulkan akurasi yang diperoleh sebesar 93%.

Penelitian lain oleh (Widaningsih, 2019) untuk membandingkan empat algoritma yaitu C4.5, *Support vector machine* (SVM), *k-nearest neigbor* (kNN,) dan Naïve Bayes untuk memprediksi nilai dan waktu kelulusan mahasiswa. Variable yang digunakan yaitu jenis kelamin dan nilai indeks prestasi dari semester 3 sampai semester 6. *Software* yang digunakan pada penelitian ini yaitu *Rapidminer*. Hasil yang diperoleh dari perbandingan antara empat algoritma tersebut, bahwa algoritma *Naive bayes* merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa tepat waktu dengan IPK lebih dari 3 dan nilai akurasi sebesar 76,79%.

Beberapa penelitian menggunakan *naïve bayes* untuk kasus memprediksi kelulusan mahasiswa, diantaranya : (Setiyani et al., 2020) menghasilkan akurasi diatas 90% dengan jumlah atribut yang berbeda pada setiap literatur, atribut yang terdapat pada semua literatur adalah IPK (indeks prestasi kumulatif). pada penelitian (Prabowo & Kodar, 2019) dengan menggunakan 244 data latih dan 62 data uji pada data kelulusan mahasiswa tahun 2011 sampai 2014, menghasilkan akurasi sebesar 82,26%. Dan juga penelitian yang dilakukan oleh (Siswanto, 2019) penerapan algoritma *naive bayes* menggunakan 14 atribut memperoleh akurasi sebesar 95,14%.

Penelitia yang dilakukan oleh (Natalius, 2011) mengungkapkan terdapat kekurangan pada metode *Naïve Bayes Classifier*, dimana pada metode ini memiliki banyak celah yang mengakibatkan pengurangan keefektifitasannya. Seperti meloloskan atribut-atribut yang tidak layak untuk dilakukan proses mining. Lalu penelitian yang dilakukan oleh (Rosandy, 2016) menemukan kelemahan lainnya pada algoritma *naive bayes* yaitu lama waktu dan tingkat akurasi prediksi yang digunakan untuk melakukan prediksi.

Salah satu teknik yang digunakan untuk mengurangi kompleksitas atribut adalah menggunakan seleksi fitur (*Feature selection*). Teknik ini dilakukan untuk memberitahu *subset* fitur yang paling berpengaruh dalah suatu *dataset*, seleksi fitur juga membentu perngurangan dimensi model, mengurangi fitur domain dan menghilangkan fitur yang berlebihan. Dengan cara ini dapat mempercepat proses pemodelan/pembelajaran (Adnyana, 2019).

Penelitian yang berkaitan dengan kelulusan mahasiswa menggunakan seleksi fitur *information gain* pada *naive bayes* telah dilakukan oleh (doni, 2020) dimana akurasi metode *naive bayes* sebesar 79,25% dapat ditingkatkan menggunakan seleksi fitur *information gain* dengan hasil 86,79%. Terdapat pengembangan dari *information gain* disebut *gain ratio*, *gain ratio* merupakan modifikasi dari *information gain* yang mengurangi biasnya. *Gain ratio* mengambil angka dan ukuran dari cabang kedalam akun ketika memilih sebuah atribut, cara ini akan mengoreksi *information gain* dalam mengambil unsur informasi dari pecahan ke sebuah akun (Priyadarsini et al., 2012). Penelitian yang dilakukan oleh (Ariestya et al., 2016) dimana melakukan perbandingan antara seleksi fitur *information gain*, *gain ratio* dan *gini index* pada *decision tree* untuk menentukan jalur kelulusan mahasiswa, memberikan hasil *gain ratio* tertinggi dengan akurasi sebesar 100% diikuti dengan *information gain* sebesar 90% dan *gini index* sebesar 85%. Dan penelitian (Socrates et al., 2016) untuk mengoptimalkan nilai akurasi *naive bayes* dengan menggunakan fitur seleksi *gain ratio*. Hasil yang diperoleh terbukti dengan menggunakan fitur seleksi *gain ratio* dapat meningkatkan akurasi metode *naive bayes* dari 91% ke 94%.

Berdasarkan permasalahan yang telah di jelaskan di atas, maka dalam hal ini peneliti akan melakukan penelitian tugas akhir yang berjudul "Penerapan *feature selection gen ratio* pada algoritma *Naive Bayes* untuk Prediksi Kelulusan Mahasiswa (Studi Kasus : Teknik Informatika UIN SUSKA Riau)" yang akan di implementasikan menggunakan bahasa pemrograman *Python*.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang diperoleh dan dijelaskan pada latar belakang, maka dapat dirumuskan masalah yang akan dijelaskan pada laporan Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana menerapkan seleksi fitur *Information Gain Ratio* pada algoritma *naive bayes* untuk memprediksi kelulusan mahasiswa di Jurusan Teknik Informatika UIN SUSKA Riau?
2. Bagaimana mengukur tingkat akurasi dalam penerapan seleksi fitur *Information Gain Ratio* pada algoritma *naive bayes* untuk memprediksi kelulusan mahasiswa di Jurusan Teknik Informatika UIN SUSKA Riau?
3. Bagaimana hasil perbandingan tingkat akurasi penerapan seleksi fitur *Information Gain Ratio* pada algoritma *naive bayes* dan tanpa menggunakan seleksi fitur *Information Gain Ratio* untuk memprediksi kelulusan mahasiswa di Jurusan Teknik Informatika UIN SUSKA Riau?

1.3 Batasan Masalah

Ada beberapa batasan masalah dalam penelitian ini, sebagai berikut:

1. Data yang dibutuhkan sebagai datasets utama adalah data akademik mahasiswa Teknik Informatika UIN SUSKA Riau dari tahun 2016-2019 yang sudah dinyatakan lulus sebanyak 530 data.
2. Kelas atau label yang digunakan sebagai hasil prediksi yaitu lulus tepat waktu dan tidak tepat waktu.

1.4 Tujuan Penelitian

Tujuan dalam penelitian ini adalah sebagai berikut :

1. Menerapkan seleksi fitur *Information Gain Ratio* dan algoritma *naive bayes* untuk memprediksi kelulusan mahasiswa di Jurusan Teknik Informatika UIN SUSKA RIAU.
2. Mengukur tingkat akurasi dalam penerapan seleksi fitur *Information Gain Ratio* dan algoritma *naive bayes* untuk memprediksi kelulusan mahasiswa di Jurusan Teknik Informatika UIN SUSKA RIAU.

3. Membandingkan tingkat akurasi penerapan seleksi fitur *Information Gain Ratio* pada algoritma *naive bayes* dan tanpa menggunakan seleksi fitur *Information Gain Ratio* untuk memprediksi kelulusan mahasiswa di Jurusan Teknik Informatika UIN SUSKA Riau

1.5 Sistematika Penelitian

Laporan penelitian ini ditulis dengan sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Bab ini menjelaskan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan yang terdiri dari tujuan umum dan tujuan khusus, serta sistematika penulisan.

BAB II LANDASAN TEORI

Pada bab ini uraikan teori-teori yang relevan dengan penelitian. Teori tersebut akan menjadi literatur bagi peneliti dalam membangun sistem.

BAB III METODOLOGI PENELITIAN

Bab ini menguraikan tahapan-tahapan penelitian yang dilakukan. Tahapan tersebut adalah tahapan pengumpulan data, pengolahan data hingga tahapan pembangunan sistem.

BAB IV ANALISA DAN PERANCANGAN

Bab ini menjelaskan proses analisa terhadap sistem lama dan sistem baru, serta perancangan *database* dan antarmuka dari sistem yang akan dibangun.

BAB V IMPLEMENTASI DAN PENGUJIAN

Bab ini menjelaskan tentang bagaimana mengimplementasikan hasil perancangan ke dalam sebuah sistem berbasis web serta menjelaskan tentang hasil pengujian dari sistem yang sudah dibangun.

BAB VI PENUTUP

Bab ini berisi kesimpulan yang diperoleh dari pembahasan mengenai sistem yang dibangun serta beberapa saran sebagai hasil akhir dari penelitian yang telah dilakukan.

BAB II
LANDASAN TEORI

2.1 Perguruan Tinggi

Perguruan tinggi adalah satuan penyelenggara pendidikan tinggi yang merupakan tingkat lanjutan dari jenjang pendidikan menengah di jalur pendidikan formal. Hal ini sesuai dengan pengertian perguruan tinggi menurut UU No. 20 tahun 2003 pasal 19 ayat 1 yang menyatakan bahwa : perguruan tinggi merupakan jenjang pendidikan setelah pendidikan menengah mencakup program pendidikan diploma, sarjana, magister, spesialis, dan doktor yang diselenggarakan oleh perguruan tinggi. Perguruan tinggi juga perlu mendeteksi perilaku mahasiswa, sehingga dapat diketahui faktor yang menyebabkan kegagalan seorang mahasiswa untuk lulus atau lulus dengan masa studi yang telah ditetapkan, seperti rendahnya kemampuan akademik, usia masuk , indeks prestasi maupun faktor-faktor lainnya (Romadhona, A., Suprapedi , S. dan Himawan, 2017).

2.2 *Data Mining*

Data mining menurut (Larose & Larose, 2014) merupakan korelasi, pola dan arah yang baru dengan menggunakan teknologi pengenalan pola serta statistika dan teknik matematika pada penyaringan sejumlah data yang besar dalam repositori. lalu menurut (Asriningtias & Mardhiyah, 2014) data mining adalah suatu kegitan untuk menemukan pola yang menarik dalam jumlah data yang besar, data dapat disimpan didalam *data base*, *data warehouse* atau penimpanan lainnya. Dan pendapat lain (Jiawei Han, 2012) data mining adalah proses menemukan pola yang menarik dan sejumlah besar pengetahuan dari data yang besar. Sumber data tersebut bisa berupa *database*, gudang data, web, dan repositori informasi maupun data langsung ke sistem.

2.3 Tahapan *Data Mining* dalam KDD

Data Mining sendiri sering disebut sebagai *Knowledge Discovery in Database* (KDD) merupakan suatu kegiatan yang melakukan penngumpulan, pemakaian data masa lampau untuk menemukan informasi seperti hubungan suatu pola dalam *dataset* yang berukuran besar (Handoko & Lesmana, 2018). beberapa Tahapan-tahapan pada *data*

mining ialah seperti pembersihan data (*data cleaning*), seleksi data (*data selection*), transformasi data (*data transformation*), proses mining, dan evaluasi pola (*pattern evaluation*).

2.3.1 Pembersihan Data (*Data Cleaning*)

Data cleaning atau pembersihan data merupakan suatu proses untuk menghilangkan data yang tidak memiliki pengaruh dalam suatu *dataset*. Data yang tidak memiliki pengaruh seperti data yang hilang, data yang tidak valid maupun data yang salah ketik akan dibuang agar meningkatkan performa menjadi lebih ringan dikarenakan jumlah data yang diproses akan berkurang dan kerumitan data menjadi lebih gampang.

2.3.2 Seleksi Data (*Data Selection*)

Data selection atau seleksi data adalah proses memilah data yang akan digunakan, karena tidak semua data pada *database* akan digunakan dalam proses penelitian. Hanya data yang memiliki kriteria sesuai akan digunakan dalam *database* untuk diteliti.

2.3.3 Transformasi Data (*Data Transformation*)

Data transformation atau transformasi data merupakan tahap dimana mengubah format pada *dataset* ke dalam format yang cocok untuk diproses. Beberapa metode dalam *data mining* memiliki jenis format data yang berbeda-beda. Transformasi data pada penelitian ini akan dilakukan dengan cara konversi dan normalisasi data.

2.3.3.1 Konversi Data (*Data Convert*)

Konversi data adalah suatu bentuk teknik mengubah data *string* menjadi angka yang biasa di kenal dengan istilah *encoding*. Dalam hal ini setelah data di seleksi, maka data akan dikonversi dari data pada tipe atribut *non-numerik* ke data *tipe numerik*.

2.3.3.2 Normalisasi Data (*Data Normalization)*

Normalisasi digunakan untuk menghindari adanya duplikasi terhadap tabel pada basis data dan juga merupakan proses untuk menguraikan tabel yang masih memiliki anomali atau ketidakwajaran yang kemudian akan menghasilkan tabel lebih sederhana dan memiliki struktur yang lebih bagus. Dengan memiliki tabel yang tidak terdapat duplikasi, akan memungkinkan pengguna melakukan *insert*, *delete* dan *update* tanpa adanya inkonsistensi data.

Data-data akan dilakukan normalisasi dengan mengubah nilai data tersebut ke dalam nilai range data (nilai data minimum – nilai data maksimum) / 0-1.

$$x_n = \frac{x_0 - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

Keterangan :

Xn = nilai baru untuk variabel X

X0 = nilai lama untuk variabel X

Xmin = nilai minimum dalam dataset

Xmax = nilai maksimum dalam dataset

2.3.4 *Mining* process

Mining process atau proses mining merupakan tahapan proses yang utama. Semua tahapan sebelumnya adalah untuk mendukung tahapan proses ini. Tahapan ini adalah proses menggunakan metode dan algoritma yang ada untuk menemukan pengetahuan dari data yang ada.

2.3.5 Evaluasi Pola (*Pattern Evaluation*)

Pattern evaluation atau evaluasi pola adalah tahapan dalam mengidentifikasi pola – pola yang ada berdasarkan hasil dari proses mining yang telah dilakukan dan menarik kesimpulan berupa pengetahuan untuk menilai apakah hipotesis yang ada telah terpenuhi atau belum.

2.4 Teknik *Data Mining*

Menurut (Maulana & Fajrin, 2018) pengelompokan *data mining* dibagi menjadi beberapa kelompok yaitu :

1. Deskripsi

   Deskripsi merupakan cara untuk menggambarkan pola dan memiliki kecenderungan yang terdapat dalam data yang tersedia.

2. Estimasi

   Memiliki kemiripan dengan klasifikasi, tetapi dalam target variabel lebih kearah numerik ketimbang arah kategori

3. Prediksi

   Prediksi merupakan memperkirakan atau menerka suatu nilai yang belum diketahui pada masa mendatang

4. Clustering

Clustering adalah suatu metode pengelompokan *record*, pengamatan, atau pembentukan kelas objek-objek yang memiliki kemiripan.

5. Asosiasi

Asosiasi merupakan metode yang mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

6. Klasifikasi

Dalam klasifikasi terdapat target kategori variabel, seperti penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu tinggi, sedang, dan rendah.

2.5 Seleksi Fitur (Feature Selection)

Seleksi fitur *(Feature selection*) digunakan untuk mengurangi kompleksitas atribut yang nantinya akan dikelola. Teknik ini dilakukan untuk memberitahu *subset* fitur yang paling berpengaruh dalah suatu *dataset*, seleksi fitur juga membentu perngurangan dimensi model, mengurangi fitur domain dan menghilangkan fitur yang berlebihan. Dengan cara ini dapat mempercepat proses pemodelan/pembelajaran (Adnyana, 2019). Algoritma feature selection dibagi menjadi tiga kelompok : filter, wrappers dan embedded selector (Rahmansyah et al., 2018).

1. Wrappers

Teknik wrapper akan mengambil subset dari suatu set fitur, dengan mengevaluasi kinerja klasifikasi pada subset, lalu subset lain akan di evaluasi menggunakan pengklasifikasi. Subset yang memiliki kinerja paling tinggi pada klasifikasian akan dipilih, dengan artian wrapper lebih dapat diandalkan untuk klasifikasi dengan kepentingan akurasi.

2. Embedded

Algoritma *Decision Tree* mewakili di antara model *Embedded,* yang memilih atribut dengan kemampuan klasifikasi potensial terbesar di setiap *node* untuk membagi subruang. Teknik Embedded melakukan feature selection selama proses mempelajari data sama seperti yang dilakukan jaringan syaraf tiruan.

3. Filter

Metode filter akan mengevaluasi setiap fitur secara bebas dari klasifikasi, Metode *Filter* menggunakan kriteria penilaian yang tepat yang mencakup jarak, informasi,

ketergantungan dan konsistensi. lalu akan memberikan peringkat pada feature yang telah dievaluasi dan mengambil yang memiliki bobot tinggi. Beberapa contoh metode filter yaitu seperti *Information Gain, Symmetrical Uncertainty , Gain Ratio* dan lain-lain.

2.5.1 Algoritma *Information Gain*

Information gain merupakan metode seleksi fitur yang paling sederhana, dimana algoritma ini melakukan perangkingan atribut pada dataset, metode ini banyak digunakan pada aplikasi kategorisasi teks, analilis data dan analisis data citra (Chormunge & Jena, 2016). Pengukuran nilai information gain didapatkan dari nilai entropy sebelum pemisahan dikurangi dengan nilai setelah pemisahan. Pengukuran ini akan digunakan sebagai tahap awal untuk menentukan atribut yang akan digunakan dan dibuang. Atribut yang memiliki kriteria atau bobot tertinggi nantinya akan digunakan untuk proses klasisfikasi (Bimantoro & Uyun, 2017).

2.5.2 Algoritma *Symmetrical Uncertainty*

Metode ini merupakan metode yang paling sering digunakan berbasis filter informasi, metode filter ini dengan berkolerasi dengan cepat untuk menghapus fitur yang tidak relevan dan berlebihan. Pengukuran metode ini digunakan untuk mengukur rudundansi (Piao et al., 2019). Dan merurut (Saikhu et al., 2019) metode Feature Selection yang digunakan untuk mengevaluasi hubungan non-linear antara variabel dan class.

2.5.3 Algoritma *Gain Ratio*

Pengembangan dari *information gain* disebut *gain ratio*, *gain ratio* merupakan modifikasi dari *information gain* yang mengurangi biasnya. *Gain ratio* mengambil angka dan ukuran dari cabang kedalam akun ketika memilih sebuah atribut, cara ini akan mengoreksi *information gain* dalam mengambil unsur informasi dari pecahan ke sebuah akun (Priyadarsini et al., 2012).

Gain Ratio dapat dihitung dengan membagi nilai *Information Gain* dengan nilai *Split Information* dengan rumus sebagai berikut:

$$Gain\ Ratio\ (D,A) = \frac{InGain(D,A)}{SplitInformation(D,A)} \quad (2.2)$$

$$InGain(D) = -\sum_{k=1}^{I} Pk^{log_2 Pk} \quad (2.3)$$

Keterangan:

A : Atribut

Gain (D, A) : *Information* atribut A

$SplitInformation(D, A)$: informasi atribut A

Split Information adalah nilai informasi dari sebuah atribut. *Split Information* dapat dihitung dengan rumus sebagai berikut:

$$SplitInformation(D, A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \, log_2 \frac{|D_j|}{|D|} \quad (2.4)$$

Keterangan:

A : Atribut

v : jumlah partisi atribut A

|Dj| : jumlah kasus pada partisi ke j

|D| : jumlah kasus dalam D

2.6 Klasifikasi

Klasifikasi adalah proses menemukan model (fungsi) yang menggambarkan dan membedakan kelas data atau kosep, untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Proses klasifikasi data, data latih akan dianalisa dengan algoritma klasifikasi. Disini label kelas adalah keputusan yang dipinjam, dan model yang dipelajari atau pengklasifikasi berbentuk aturan klasifikasi. Data uji digunakan untuk memperkirakan akurasi dari aturan klasifikasi. Jika akurasi bisa diterima, maka aturan dapat diterapkan pada klasifikasi data baru. Model yang didapatkan didasarkan pada analisis dari data latih (objek yang label kelasnya diketahui), terdapat banyak metode klasifikasi seperti *support vector machine*, *k-nearest-neighbor*, *naive bayes* dan lain-lain (Jiawei Han, 2012).

2.6.1 Algoritma *Support Vector Machine*

Support Vector Machine (SVM) diperkenalkan pertama kali oleh Vapnik di tahun 1992 sebagai suatu rangkaian dengan konsep yang unggul pada bidang pengenalan pola. Metode ini masih terbilang muda, walapun begitu kemampuan evaluasinya dalam berbagai hal menempatkannya menjadi salah satu tema yang berkembang pesat (Nugroho, 2007). *Support Vector Machine* (SVM) sendiri memiliki prinsip dasar linier clasifier, dimana klasifikasi yang secara linear dapat dipisahkan. Tetapi *Support Vector*

Machine (SVM) telah dikembangkan agar dapat beroperasi pada permasalahan *non-linear* dengan cara memasukkan konsep kernel (Octaviani et al., 2014).

2.6.2 Algoritma *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* (KNN) merupakan salah satu metode yang menerapkan *supervised learning*, dengan kata lain algoritma ini bertujuan untuk menemukan pola baru. Ketepatan akurasi algoritma ini ditentukan oleh ada atau tidaknya data yang tidak relevan. Algoritma ini biasanya digunakan untuk melakukan proses analisis klasifikasi, tetapi belakangan ini metode KNN dapat juga diguanakan untuk prediksi (Bode, 2017). Perhitungan metode ini dilakukan berdasarkan data pembelajaran data yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran akan ditampilkan ke ruang berdimensi banyak, dimana masing masing dimensi menampilkan fitur dari *dataset*. Ruang tersebut akan dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Ruang akan ditandai sebuah titik kelas c, dimana kelas c merupakan klasifikasi yang paling bnyak ditemukan pada buah tetangga terdekat (Yustanti, 2012).

2.6.3 Algoritma *Naïve Bayes*

Naive Bayes classifier (NBC) merupakan salah satu algoritma pada teknik klasifikasi yang ditemukan oleh ilmuan inggris Thomas Bayes. *Naive bayes* merupakan salah satu metode pembelajaran mesin dengan perhitungan probabilitas dan statistik yang memprediksi peluang dimasa depan dengan menggunakan pengalaman dimasa sebelumnya atau yang lebih dikenal dengan teorema *bayes*. Dikombinasikan dengan *naive* dimana bahwa semua atribut independen atau tidak ada ciri tertentu dari sebuah kelas yang berhubungan dengan ciri kelas lainnya (Bustami, 2014).

Dalam penelitian (Saleh, 2015) klasifikasi menggunakan data numerik menggunakan rumus *densitas gauss* :

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2{}_{ij}}} \quad (2.5)$$

Keterangan :

P : Peluang

Xi : Atribut ke i

xi : Nilai atribut ke i

Y : Kelas yang dicari

yi : Sub kelas Y yang dicari

μ : *mean* (rata – rata dari seluruh atribut)

σ : *Deviasi* standar (menyatakan varian dari seluruh atribut)

Alur distribusi gaussian pada algoritma *naive bayes*:

1. Baca data latih
2. jika data bersifat non-numerik hitung jumlah dan probabilitas, namun apabila data numerik maka :
    a. Cari nilai *mean* dan standar *deviasi* dari masing-masing parameter yang merupakan data numerik. Adapun persamaan yang digunakan untuk menghitung nilai rata – rata hitung (*mean*) dapat dilihat sebagai berikut :

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2.6)$$

Keterangan :

μ : rata – rata hitung (*mean*)

xi : nilai sample ke -i

n : jumlah sampel

Dan rumus untuk menghitung nilai simpangan baku (standar deviasi) sebagai berikut:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \quad (2.7)$$

Keterangan :

σ : standar deviasi

xi : nilai x ke -i

μ : rata-rata hitung

n : jumlah sampel

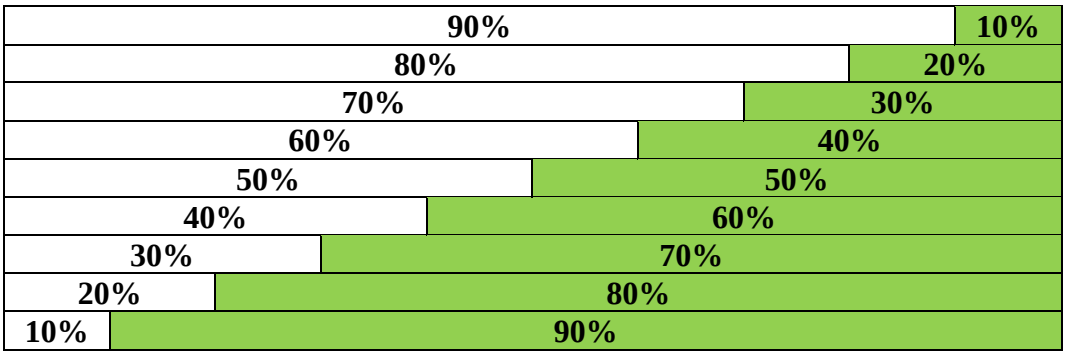
    b. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.
3. Hitung data testing dengan cara menghitung probabilitas distribusi gaussian berdasarkan nilai dalam tabel *mean*, standard *deviasi* dan probabilitas dari masing-

masing atribut, kemudian kalikan semua nilai probabilitas yang telah dihitung pada atribut berdasarkan kelasnya.

4. Nilai dengan bobot terbesar adalah solusinya

2.7 *Split Validation*

Split validation merupakan salah satu operator pada *rapidminer* yang memiliki fungsi untuk melakukan validasi sederhana secara acak kemudian membagi sebuah dataset menjadi 2 bagian yaitu data uji dan data latih. Menggunakan *Split validation* akan dilakukan percobaan *training* berdasarkan rasio *split* yang telah ditetapkan, lalu sisa dari data latih disebut data uji. Data latih merupakan data yang akan digunakan untuk mempelajari pola pada data set, dan data testing adalah sisa data dari data latih, berfungsi untuk menguji keakuratan hasil pembelajaran.



Keterangan:

	=	Data *Training*
	=	Data *Testing*

2.8 *Confusion Matrix*

Confusion matrix merupakan tabel yang memberikan informasi hasil dari klasifikasi yang dilakukan oleh sistem dengan membandingkan hasil klasifikasi sebenarnya. Tabel *confusion matrix* berisi empat kemungkinan dimana *output* sebagai bahan acuan untuk membandingkan antara kejadian sebenarnya dengan prediksi kejadian.

Tabel 2.2 *Confusion Matrix* untuk Klasifikasi Dua Kelas

Actual	***Predicted***	
	Tepat Waktu	**Tidak Tepat Waktu**
Tepat Waktu	TN	FP
Tidak Tepat Waktu	FN	TP

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{2.8}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{2.9}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2.10}$$

Keterangan :

TN = nilai true negatives

TP = nilai true positives

FP = nilai false positives

FN = nilai false negatives

2.9 Penelitian Terkait

Bebrerapa penelitian terkait yang menggunakan algoritma Random Forest adalah sebagai berikut :

1. **Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa**

   Penelitian melakukan pemindaian pada *database* untuk mendapatkan informasi yang dibutuhkan. Aplikasi ini dibangun menggunakan bahasa pemograman *Borland Delphi* 7 dengan *database* SQL Server 2000 untuk tempat penyimpanan data. Hasil yang diperoleh oleh aplikasi ini bahwa atribut yang paling berpengaruh untuk mengetahui tingkat ketepatan waktu kelulusan mahasiswa adalah indeks prestasi kumulatif.

2. **Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin Dan Indeks Prestasi Menggunakan Algoritma Decision Tree**

   Penilitan ini diawali dengan mengambil data dari indeks prestasi (IP) mahasiswa yang mengambil mata kuliah setiap semesternya. Data *training* yang digunakan yaitu

data mahasiswa angkatan 2009 dengan atributnya yaitu usia, jenis kelamin, indeks prestasi selama 4 semester pertama yaitu semester 1, semester 2, semester 3 dan semester 4. Lalu melakukan perhitungan nilai *information gain* pada setiap atribut, atribut dengan nilai tertinggi yaitu indeks prestasi semester 4. Lalu melakukan perbandingan mining algoritma *DecisionTree C4.5*, *ID3* dan *Chaid* untuk mengetahui algoritma mana yang paling cocok untuk melakukan prediksi kelulusan mahasiswa. Hasil yang diproleh yaitu bahwa algoritma decision tree memiliki kinerja lebih baik dari pada algoritma lainnya.

3. **Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naive Bayes : Systematic Review**

   Pada penelitian ini melakukan prediksi mengenai kelulusan mahasiswa yang tepat waktu menggunakan atribut dari database perguruan tinggi. Hasil yang diperoleh memberikan hasil akurasi diatas 90% pada ketiga literatur walaupun menggunakan jumlah atribut dan aplikasi data mining yang berbeda.

4. **Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa**

   Pada penelitian ini melakukan pengaruh mata kuliah terhadap lama studi mahasiswa menggunakan teknik feature selection Correlation Based, Information Gain Based, dan Learner Based, hasil akurasi dari feature selection tersebut akan diukur menggunakan algoritma naive bayes. Hasil yang diperoleh menunjukkan hasil mampu meningkatan akurasi klasifikasi pada algoritma naive bayes. akurasi tertinggi *dataset* nilai mahasiswa dihasilkan oleh teknik *Learned based* menggunakan model *wrapper*, sedangkan hasil terendah diperoleh teknik *information gain*.

5. **Decision Tree Learning Untuk Penentuan Jalur Kelulusan Mahasiswa**

   Penelitian ini melakukan perbandingan seleksi fitur *information gain*, *gain ratio* dan *gini index* pada algoritma *Iterative Dichotomiser* 3 (ID3) untuk penentuan jalur kelulusan mahasiswa. Model yang dihasilkan oleh ketiga seleksi fitur memiliki hasil akurasi diatas 85%, dimana hasil tertinggi dihasilkan oleh *gain ratio* dengan akurasi 100% dan hasil terendah dihasilkan oleh gini index dengan akurasi 85%.

6. **Naïve Bayes dan Filtering Feature Selection Information Gain untuk Prediksi Ketepatan Kelulusan Mahasiswa**

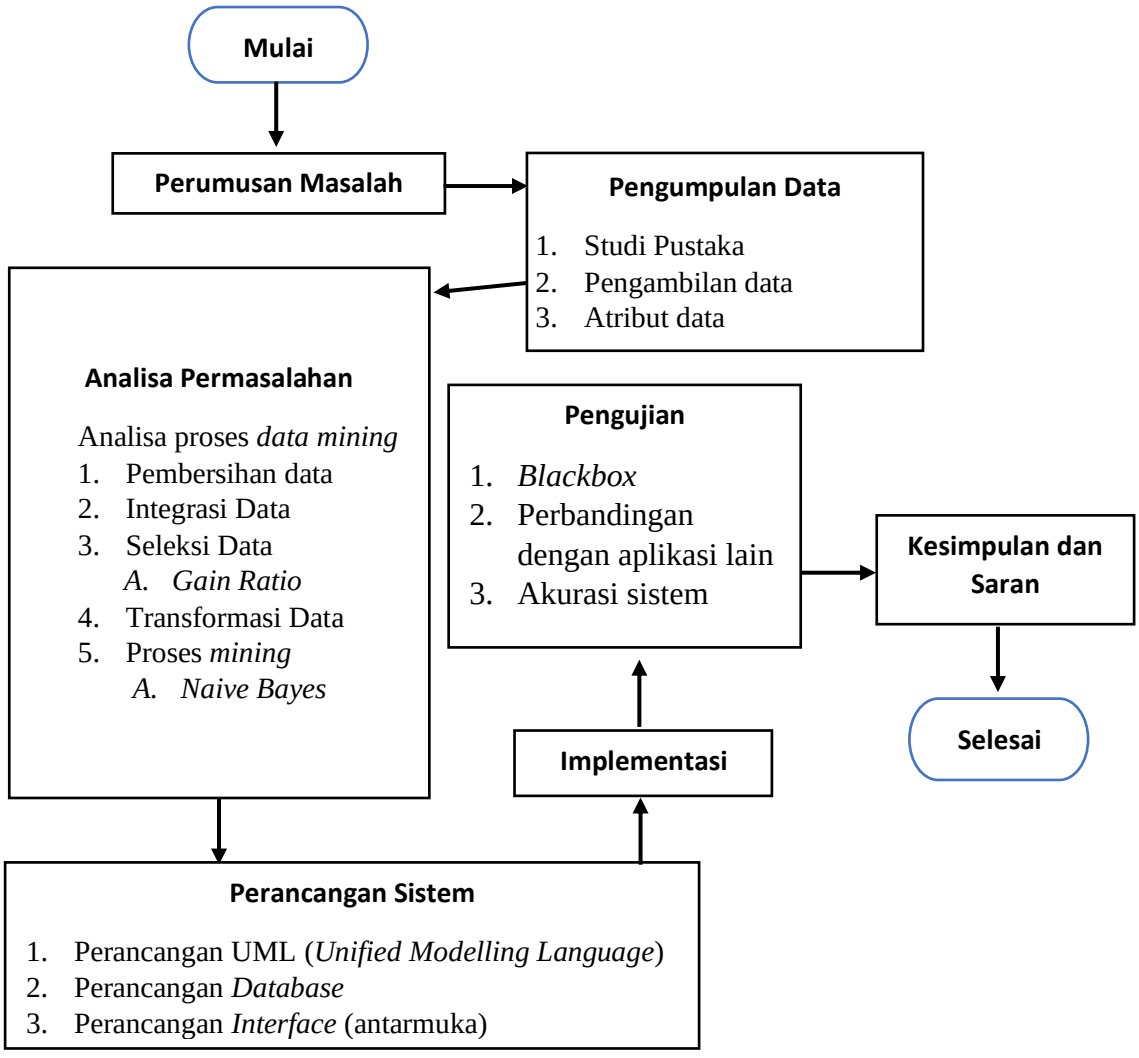
Pada penelitian ini melakukan komparasi antara algoritma *naive bayes* dengan dua algoritma *naive bayes* dan *information gain* untuk memprediksi ketepatan kelulusan mahasiswa. Hasil yang diperoleh penelitian ini menunjukkan akurasi terbaik pada kombinasi algoritma *naive bayes* dan algoritma *feature selection information gain* sebesar 89,79% untuk penggunaan 3 atribut.

7. **Analisis Komparasi Algoritma Naive Bayes Dan C4-5 Untuk Waktu Kelulusan Mahasiswa**

Penilitian ini bertujuan untuk membandingkan aloritma *naive bayes* dengan algoritma C4.5 dengan seleksi fitur *gain ratio* pada data kelulusan mahasiswa STMIK Widya Pratama tahun 2011 sampai 2014. Data set memiliki 12 variabel yang akan diolah menggunakan *software* rapidminer yang nantinya akan diuji tingkat akurasinya dengan confusion matriks. Hasil dari penelitian ini adalah algoritma C4.5 mendapatkan hasil yang lebih baik dari pada algoritma *naive bayes* dengan selisih 1.59%.

BAB III
METODOLOGI PENELITIAN

Metodologi penelitian merupakan tahapan proses yang dijadikan pedoman dalam melakukan penelitian untuk mencapai tujuan. Tahapan penelitian yang akan dilakukan ditunujukkan pada gambar berikut.



Gambar 3.1 Metodologi Penelitian

3.1 Identifikasi Masalah

Tahap Pertama yang dilakukan yaitu mengidentifikasi terhadap permasalahan yang akan diangkat sebagai topik penelitian. Tahap ini dilakukan untuk mengidentifikasi permasalahan yang ada sehingga dilakukan penelitian untuk menghasilkan solusi yang diharapkan. Perumusan masalah yang dilakukan pada penelitian ini adalah bagaimana memprediksi kelulusan mahasiswa di Teknik Informatika UIN SUSKA Riau dengan menggunakan algoritma *Naive Bayes* dengan bantuan *feature selection Gain Ratio*.

3.2 Pengumpulan Data

Tahapan Pengumpulan data dilakukan setelah tahap indentifikasi masalah. Tahapan ini dilakukan untuk mendapatkan informasi terhadap data yang berhubungan dengan penelitian yang sedang dilakukan. Tahapan pengumpulan data ini dilakukan dengan dua tahap yaitu dengan studi pustaka dan pengambilan data.

3.2.1 Studi Pustaka

Studi pustaka adalah proses mempelajari dan memahami teori-teori yang berhubungan dengan topik penelitian. Pada proses ini dilakukan pemahaman terhadap beberapa buku, jurnal dan artikel yang memuat teori-teori sebagai referensi dalam melakukan penelitian ini.

3.2.2 Pengambilan Data

Pengambilan data dilakukan dengan mengumpulkan data-data kelulusan mahasiswa Teknik Informatika. Pengumpulan data yang dilakukan pada penelitian ini adalah data sekunder. Data sekuder sendiri merupakan data yang diperoleh menggunakan perantara seperti buku, jurnal dan penelitian sebelumnya. Data yang diperoleh berasal dari Pusat Teknologi Informasi dan Pangkalan Data (PTIPD) Universitas Islam Negeri Sultan Syarif Kasim Riau yang berupa data kelulusan mahasiswa dari tahun 2016 – 2019, data tersebut nantinya akan digunakan sebagai pembentuk model yang nantinya akan diteliti.

3.2.3 Atribut Data

Atribut data yang digunakan merupakan seluruh atribut yang terdapat pada data kelulusan mahasiswa Teknik Informatika Universitas Islam Negeri Sultan Syarif Kasim Riau yang nantinya akan diseleksi terlebih dahulu dengan menggunakan algoritma seleksi fitur (*feature selection*).

Tabel 3.1 Atribut Data Sebelum di Seleksi

No	Atribut	Keterangan
1	nilai_ptik	Nilai Mata Kuliah Pengantar Teknologi Informasi dan Komunikasi
2	nilai_sisdig	Nilai Mata Kuliah Sistem Digital
3	nilai_daspro	Nilai Mata Kuliah Dasar Pemrograman
4	nilai_alpro	Nilai Mata Kuliah Algoritma dan Pemrograman
5	nilai_ecs	Nilai Mata English Communication Skill
6	nilai_matdis	Nilai Mata Kuliah Matematika Diskrit
7	nilai_arkom	Nilai Mata Arsitektur Komputer
8	nilai_basdat	Nilai Mata Kuliah Basis Data
9	nilai_metnum	Nilai Mata Kuliah Metode Numerik
10	nilai_tbo	Nilai Mata Teori Bahasa Otomata
11	nilai_strukdat	Nilai Mata Kuliah Struktur Data
12	nilai_sbd	Nilai Mata Kuliah Sistem Basis Data
13	nilai_so	Nilai Mata Sistem Operasi
14	nilai_jarkom	Nilai Mata Jaringan Komputer
15	nilai_ki	Nilai Mata Keamanan Informasi
16	nilai_rpl	Nilai Mata Rekayasa Perangkat Lunak
17	nilai_si	Nilai Mata Sistem Informasi
18	nilai_kb	Nilai Mata Kecerdasan Buatan
19	nilai_mpti	Nilai Mata Manajemen Proyek Teknologi Informasi
20	nilai_pb	Nilai Mata Kuliah Pemrograman Bergerak
21	nilai_kp	Nilai Mata Kuliah Kerja Praktek
22	IP Semester 1	Indeks Prestasi Semester Mahasiswa Semester 1
23	IP Semester 2	Indeks Prestasi Semester Mahasiswa Semester 2
24	IP Semester 3	Indeks Prestasi Semester Mahasiswa Semester 3
25	IP Semester 4	Indeks Prestasi Semester Mahasiswa Semester 4
26	IP Semester 5	Indeks Prestasi Semester Mahasiswa Semester 5
27	IP Semester 6	Indeks Prestasi Semester Mahasiswa Semester 6
28	Status Kelulusan	Status Kelulusan Mahasiswa Tepat Waktu / Tidak Tepat Waktu

3.3 Analisa Permasalahan

Terdapat beberapa tahapan yang dimiliki *data mining*, masing-masing tahapan ini nantinya akan menghasilkan faktor yang berpengaruh terhadap kelulusan mahasiswa. Berikut tahapan yang akan dilakukan pada penelitian ini :

3.4.1 Cleaning Data

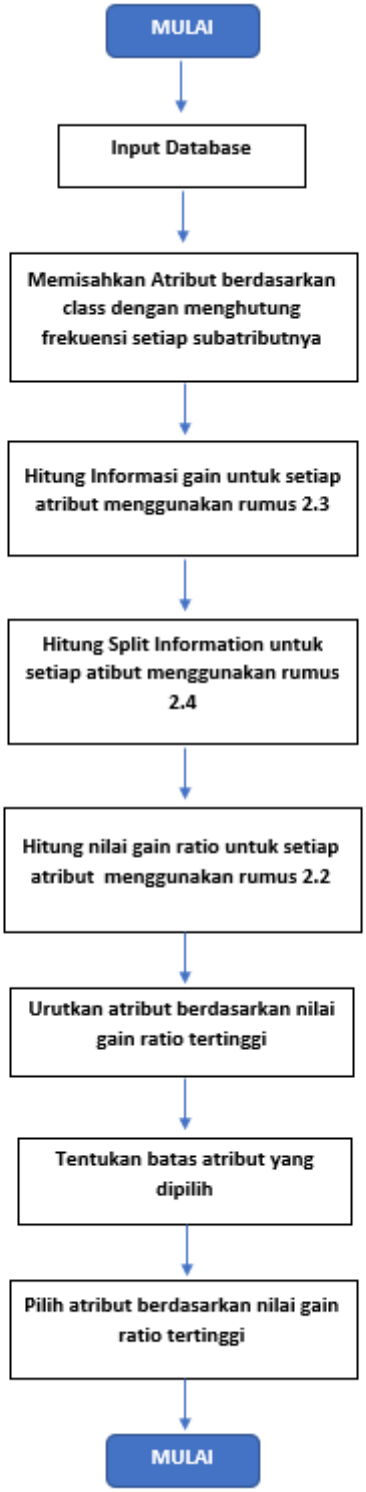
Pembersihan data (cleaning data) merupakan langkah pertama yang akan dilakukan pada data yang kita miliki, pembersihan data akan melakukan proses menyeleksi, memperbaiki maupun menghapus data yang tidak memiliki pengaruh dalam *dataset*. Pembersihan data akan dilakukan dalam data kelulusan mahasiswa dikarenakan sering terdapat kasus dimana terjadinya data yang tidak konsisten dan data yang hilang. Penanganan yang dilakukan jika terdapat data yang hilang maka dilakukan mengisi data yang hilang tersebut berdasarkan data yang sebelumnya. Dengan dilakukannya pembersihan data diharapkan meningkatkan peforma menjadi lebih ringan dikarenakan jumlah data yang diproses berkuran dan tingkat kerumitan data menjadi berkurang.

3.4.2 Seleksi Data

Data selection atau seleksi data adalah proses memilah data yang akan digunakan, karena tidak semua data pada database akan digunakan dalam proses penelitian. Hanya data yang memiliki kriteria sesuai akan digunakan dalam *database* untuk diteliti. Metode *feature selection* yang digunakan yaitu *Information Gain Ratio*.

3.4.2.1 Gain Ratio

Pengembangan dari *information gain* disebut *gain ratio*, *gain ratio* merupakan modifikasi dari *information gain* yang mengurangi biasnya. *Gain ratio* mengambil angka dan ukuran dari cabang kedalam akun ketika memilih sebuah atribut, cara ini akan mengoreksi *information gain* dalam mengambil unsur informasi dari pecahan ke sebuah akun. Tahapan algoritma *Gain Ratio* dapat dilihat pada gambar 3.2.



Gambar 3.2 Tahap Algoritma Gain Ratio

3.4.3 Transformasi Data

Data transformation atau transformasi data merupakan tahap dimana mengubah format pada *dataset* ke dalam format yang cocok untuk diproses. Beberapa metode dalam *data mining* memiliki jenis format data yang berbeda-beda. Atribut yang akan dilakukan transformasi data adalah atribut yang memiliki pengaruh tertinggi dalam kelulusan mahasiswa. Transformasi data pada penelitian ini akan dilakukan dengan cara konversi dan normalisasi data.

3.4.3.1 Konversi Data (*Data Convert*)

Konversi data adalah suatu bentuk teknik mengubah data *string* menjadi angka yang biasa di kenal dengan istilah *encoding*. Dalam hal ini setelah data di seleksi, maka data akan dikonversi dari data pada tipe atribut *non-numerik* ke data *tipe numerik*.

3.4.3.2 Data Normalization

Normalisasi digunakan untuk menghindari adanya duplikasi terhadap tabel pada basis data dan juga merupakan proses untuk menguraikan tabel yang masih memiliki anomali atau ketidak wajaran yang kemudian akan menghasilkan tabel lebih sederhana dan memiliki struktur yang lebih bagus. Dengan memiliki tabel yang tidak terdapat duplikasi, akan memungkinkan pengguna melakukan *insert*, *delete* dan *update* tanpa adanya inkonsistensi data. Data-data akan dilakukan normalisasi dengan mengubah nilai data tersebut ke dalam nilai *range* data (nilai data minimum – nilai data maksimum) / 0-1 pada rumus (2.1).

3.4.4 Pembagian Data

Data yang telah melewati proses *pre-processing* akan melakukan proses pembagian data, pembagian data merupakan mengubah data set menjadi data *training* dan data *testing* dengan metorde *split validation*. Rasio perbandingan yang yang digunakan adalah 90:10, 80:20, 70:30 dan 60:40. Berikut penjelasan bagaimana proses pembagian data :

a. Data *training*, merupakan data kelulusan mahasiswa dari tahun 2016-2019 akan menjadi *inputan* yang akan dibagi secara acak dengan menggunakan metode *split validation*. Data *training* dibuat untuk melatih algoritma untuk mencari model yang pas. Jika rasio yang ditetapkan 90:10, yaitu 90% data kelulusan terdiri dari data latih.

b. Data *testing*, 10% dari data kelulusan yang telah ditetapkan merupakan data *testing*, model yang dihasilkan pada data *training* akan digunakan sebagai acuan untuk mengetahui performa model yang telah didapatkan pada proses *testing*.

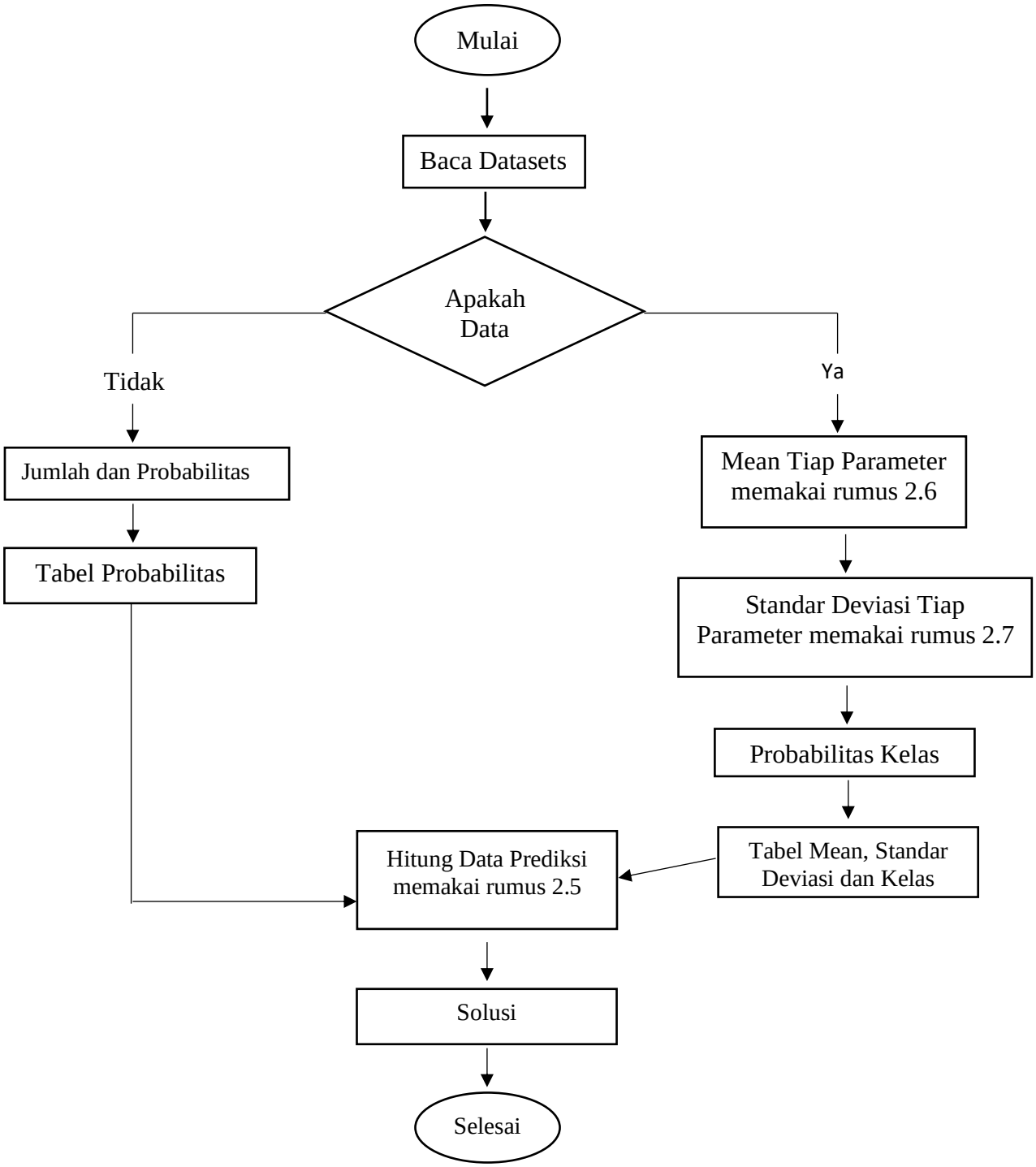
Pembagian data *training* dan data *testing* dilakukan pada masing-masing rasio perbandingan. Sehingga masing-masing perbandingan rasio memiliki akurasi dari pembagian data. Setelah didapat akurasi dari model terbaik yang telah di evaluasi, selanjutnya akan dilakukan proses *data mining* untuk memprediksi data baru mahasiswa dan mengetahui hasil prediksi pada algoritma *Naïve Bayes* berbasis *Information Gain Ratio.*

3.4.5 Mining process

Mining process atau proses *mining* merupakan tahapan proses yang utama. Semua tahapan sebelumnya adalah untuk mendukung tahapan proses ini. Tahapan ini adalah proses menggunakan metode dan algoritma yang ada untuk menemukan pengetahuan dari data yang ada.

3.4.4.1 Naive Bayes

Naive Bayes classifier (NBC) merupakan salah satu algoritma pada teknik klasifikasi yang ditemukan oleh ilmuan inggris Thomas Bayes. *Naive bayes* merupakan salah satu metode pembelajaran mesin dengan perhitungan probabilitas dan statistik yang memprediksi peluang dimasa depan dengan menggunakan pengalaman dimasa sebelumnya atau yang lebih dikenal dengan teorema *bayes*. Dikombinasikan dengan *naive* dimana bahwa semua atribut independen atau tidak ada ciri tertentu dari sebuah kelas yang berhubungan dengan ciri kelas lainnya. Alur distribusi gaussian pada algoritma *naïve bayes* dapat dilihat pada gambar 3.3.



Gambar 3.3 Tahapan Algoritma *Naïve Bayes*

3.4 Perancangan System

Tahap lanjutan dari proses analisa adalah perancangan. Pada tahap ini dilakukan penggambaran dan perencanaan dari sistem yang akan dibangun. Proses perancangan dalam penelitian ini adalah sebagai berikut:

3.5.1 Perancangan UML (Unifield Modelling Languange)

Unififed Modelling Language (UML) merupakan himpunan struktur dan teknik untuk permodelan desain program berorientasi obyek (OOP). Perancangan UML pada penelitian ini meliputi perancangan *Use Case Diagram, Sequence Diagram, dan Class Diagram.*

1. *Use Case Diagram*

   Use case diagram merupakan tahap yang akan digambarkan atau dijelaskan terhadap proses pembangunan sistem.

2. *Sequence Diagram*

   Sequence diagram adalah tahap yang menggambarkan serangkaian pesan yang dilakukan oleh beberapa obyek.

3. *Class Diagram*

   Class diagram adalah tahap yang dilakukan dengan menggambarkan struktur *class* dengan *class* lainnya.

3.5.2 Perancangan *Database*

Perancangan ini bertujuan untuk mendesain *database* sebagai tempat penyimpanan data pada sistem yang akan dibangun. Keberadaan *database* akan mendukung kebutuhan pemrosesan data dan beberapa obyek kinerja sistem.

3.5.3 Perancangan *Interface* (Antarmuka)

User interface merupakan tampilan visual pada produk yang disajikan, yang menghubungkan sistem dengan pengguna. Tampilan pada *interface* dapat berupa bentuk, warna dan tulisan sebagaimana tampilan sebuah produk yang dilihat oleh pengguna.

3.5 Implementasi

Implementasi sistem adalah prosedur tersistematika yang dilakukan dalam menyelesaikan desain pada dokumen yang disetujui. Aplikasi ini membutuhkan

perangkat pendukung yaitu perangkat keras. Spesifikasi perangkat keras adalah sebagai berikut :

Processor : *Intel Core i5*

RAM : 8 GB

SSD : 512 GB

Spesifikasi perangkat lunak:

Platform : *Microsoft Windows 10 Home Single Language*

Database : MySQL

Web Server : *Apache*

Browser : *Google Chrome*

Server : *localhost*

Bahasa Pemrograman : HTML, Python, dan Java Script

Text Editor : Php Storm 2019.3.4

3.6 Pengujian

Menurut kamus besar bahasa indonesi (KBBI) pengujian merupakan proses, cara ataupun perbuatan untuk menguji sesuatu. Tahap pengujian ini dilakukan untuk mengetahui hasil dari model dan sistem yang telah dibuat. Untuk mengetahui kelayakan dan fungsi sistem tersebut maka akan dilakukan pengujian akurasi model dan pengujian sistem menggunakan :

a. Pengujian *black box*, yaitu pengujian yang dilakukan pada fungs-fungsi sistem yang dibangun sesuai dengan yang diharapkan.
b. Pengujian hasil sistem, yaitu pengujian dengan membandingkan hasil pada sistem dengan hasil menggunakan aplikasi rapidminer.
c. Pengujian akurasi, yaitu pengujian yang dilakukan dengan *split validation* untuk mengevaluasi model pada metode yang digunakan dengan menggunakan *confusion matrix*. Pengujian dilakukan berdasarkan pengolahan pada *split validation* pada data setelah di normalisasi, sehingga akan diperoleh akurasi dari masing-masing perbandingan rasio dari pengujian metode *naïve bayes* berbasis *information gain ratio* dan tanpa menggunakan *information gain ratio*. *Confusion matrix* diperoleh dengan melakukan perbandingan pada hasil prediksi dari data

dan hasil sebenarnya. Untuk mengetahui kinerja pada metode akan digunakan rumus *precision* (2.9), *recall* (2.10) dan *akurasi* (2.8).

3.7 Kesimpulan dan Saran

Tahap ini berisi *review* dari penelitian yang telah dilakukan sebelumnya. *Review* ini bertujuan untuk memastikan bahwa sistem yang telah dibangun sudah sesuai dengan yang diharapkan. Kemudian diberikan beberapa saran untuk menjadi acuan dalam pengembangan penelitian selanjutnya.