Ariska Kurnia Rachmawati,M.Sc

Riska Ayu Ardani,M.Pd

ANALISIS KINERJA ALGORITMA C5.0 DAN NAIVE BAYES UNTUK PENGENALAN POLA LULUSAN MAHASISWA

DIBIAYAI DENGAN
ANGGARAN DIPA BLU FST
TAHUN 2021

UIN WALISONGO SEMARANG

LAPORAN PENELITIAN

ANALISIS KINERJA ALGORITMA KLASIFIKASI C5.0 DAN NAÏVE BAYES UNTUK PENGENALAN POLA LULUSAN MAHASISWA

Oleh :

Ariska Kurnia Rachmawati,M.Sc
Riska Ayu Ardani, M.Pd

FAKULTAS SAINS DAN TEKNOLOGI
UIN WALISONGO SEMARANG
TAHUN 2021

Abstrak

Salah satu hal yang menjadi perhatian dalam peningkatan mutu pendidikan tinggi di UIN Walisongo adalah mempersiapkan tindakan preventif terhadap mahasiswa yang berpotensi mengalami hambatan dalam proses pembelajaran. Hambatan yang berpotensi dihadapi mahasiswa adalah lamanya masa studi yang beragam, cepat atau lama. Untuk menangani hal tersebut dibangunlah sistem klasifikasi pola lulusan mahasiswa menggunakan model algoritma C5.0 dan Naïve Bayes terhadap data mahasiswa Program Studi Matematika Tahun 2015 dan 2016. Dalam penelitian ini, pola lulusan akan diklasifikasikan menjadi dua yaitu lulus tepat waktu dan tidak lulus tepat waktu. Pengujian dilakukan dengan cara mengukur kinerja dua algoritma tersebut menggunakan metode pengujian *Confussion Matrix* dan nilai kurva ROC. Hasil pengujian menunjukkan bahwa algoritma C5.0 memberikan hasil yang lebih baik daripada Algoritma Naïve Bayes yang ditunjukkan berdasarkan nilai Akurasi sebesar 94,12% dan dibandingkan hasil yang didapatkan algoritma Naïve Bayes yaitu nilai akurasi 91,18%. Secara keseluruhan, variabel yang paling mempengaruhi hasil prediksi pola lulusan adalah IP Semester 1, IP Semester 3, IP Semester 8, jenis kelamin,jumlah sks total dan jalur masuk yang diikuti.

Kata Kunci: klasifikasi, Algoritma C5.0, Algoritma Naïve Bayes, Data Mining, pola lulusan.

KATA PENGANTAR

Puji dan syukur penulis panjatkan atas kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan laporan penelitian yang berjudul "Analisis Kinerja Algoritma Klasifikasi C5.0 Dan Naïve Bayes Untuk Pengenalan Pola Lulusan Mahasiswa". Penelitian ini dilakukan penulis sebagai salah satu bentuk pelaksanaan tugas pokok dosen untuk meneliti selain mengabdi dan mengajar.

Dalam kesempatan ini, penulis menyampaikan terimakasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. Imam Taufik, M.Ag selaku Rektor UIN Walisongo Semarang.
2. Bapak Dr. H. Ismail, M.Ag selaku Dekan Fakultas Sains dan Teknologi UIN Walisongo Semarang.
3. Bapak Dr. H. Akhmad Arif Junaidi, M.Ag selaku Ketua Lembaga Penelitian dan Pengabdian kepada Masyarakat (LP2M) UIN Walisongo Semarang.
4. Para Reviewer yang telah memberikan masukan dan saran.
5. Semua pihak yang telah membantu sehingga penelitian ini dapat terlaksana.

Penulis berharap apa yang telah disusun dapat bermanfaat bagi perkembangan ilmu pengetahuan khususnya bidang Matematika Keuangan. Penulis menyadari laporan penelitian ini masih banyak kelemahan dan kekurangan, maka segala saran dan kritik yang membangun sangat penulis harapkan guna perbaikan ke depannya.

Semarang, November 2021

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Sumber daya manusia merupakan salah satu modal awal pembangunan yang dihasilkan melalui proses pembelajaran pendidikan. Pendidikan mempunyai peran yang sangat penting dalam mencetak generasi penerus bangsa yang unggul dan mampu bersaing. Tujuan pembangunan suatu bangsa meliputi meningkatnya kualitas sumber daya manusia dan berkembangnya kepribadian serta kesejahteraan bangsa dapat dicapai, salah satunya melalui Pendidikan yang bermutu. Mutu Pendidikan mengacu pada proses dan hasil Pendidikan. Mutu proses pendidikan melibatkan beberapa aspek seperti bahan ajar, metodologi, sarana dan prasarana serta factor social lingkungan, dan sumber daya lain yang mendukung suasana akademik menjadi kondusif.

Di Era industry 4.0 , Pendidikan di UIN Walisongo tidak hanya berorientasi pada pragmatism dan materialism namun tujuan yang utama dari Universitas Islam Negeri (UIN) Walisongo adalah menghasilkan lulusan yang memiliki kemampuan akademik dan berakhlakul karimah yang professional dalam mengaplikasikan kesatuan ilmu pengetahuan. Dengan kata lain, lulusan UIN Walisongo setidaknya memiliki capaian pembelajaran sebagaimana capaian kompetensi yang dimiliki seseorang yang mengikuti pelatihan kerja atau pengalaman kerja.

Keberhasilan proses pendidikan tinggi tergambarkan dari lulusan yang berkualitas yang sesuai dengan tuntutan lingkungan khususnya dunia kerja, dimana hal ini menjadi salah satu indikator mutu. Upaya peningkatan mutu lulusan di UIN Walisongo

secara kontinu ditingkatkan dengan memfokuskan pada sumber daya manusia dan teknologi yang meliputi mahasiswa sebagai peserta didik, dosen sebagai pendidik serta sarana prasarana yang mendukung. Salah satu hal yang menjadi perhatian dalam peningkatan mutu pendidikan tinggi di UIN Walisongo adalah mempersiapkan tindakan preventif terhadap mahasiswa yang berpotensi mengalami hambatan dalam proses pembelajaran. Hambatan yang berpotensi dihadapi mahasiswa adalah lamanya masa studi yang beragam, cepat atau lama. Masa studi ini berkaitan dengan keterserapan lulusan dalam dunia kerja, dimana persaingan untuk mendapatkan pekerjaan saat ini lebih ketat dan semakin berat. Monitoring dan Pendampingan intensif mahasiswa oleh dosen pembimbing akademik, program percepatan bimbingan tugas akhir, dan perbaikan sistem manajemen pendidikan merupakan hal yang digaungkan akan menunjang keberhasilan proses pendidikan tinggi di UIN Walisongo.

Data yang beragam terkait lamanya masa studi mahasiswa merupakan kumpulan informasi yang dapat digunakan untuk penentu kebijakan atau memutuskan suatu tindakan pada program studi. Terdapat beberapa faktor yang dapat mempengaruhi mahasiswa lama dalam menyelesaikan studinya. Menurut Samekto (2014), faktor – faktor tersebut dikategorikan menjadi dua, yaitu faktor internal dan faktor eksternal. Faktor internal merupakan faktor yang bersumber dari diri seorang mahasiswa yang menjadi penyebab lamanya masa studi meliputi faktor kecerdasan, minat dan bakat, serta motivasi mahasiswa. Faktor eksternal meliputi faktor yang bersumber dari luar diri mahasiswa, seperti faktor sosial ekonomi keluarga, lingkungan dan pergaulan serta kurikulum pembelajaran. Keberagaman data tersebut sangat bermanfaat apabila ditemukan suatu pola data yang didapatkan dari hasil analisis yang bertujuan untuk menggali informasi dan memiliki

manfaat bagi pemilik data dan pihak eksternal. Dengan beragamnya data di dunia Pendidikan mendorong munculnya ilmu baru untuk mengatasi permasalahan teknik penggalian informasi dari data dalam jumlah yang besar atau disebut *data mining.*

Teknik *data mining* dapat memberikan pengetahuan yang sebelumnya tersembunyi di dalam gudang data sehingga menjadi informasi yang berharga (Santosa, 2007). Data mining adalah serangkaian proses pencarian atau penambangan informasi dari sekumpulan data yang sangat besar untuk menemukan pola baru. Sumber data meliputi database, data gudang, Web, repositori informasi lain, atau data yang dialirkan ke sistem secara dinamis (Jiawei Han, Micheline Kamber & Jian Pei, 2012). Hasil yang diperoleh dari teknik *data mining* dapat digunakan dalam pengambilan keputusan khususnya variabel-variabel yang berpengaruh pada pola lulusan dalam data primer mahasiswa.

Beberapa penelitian telah membahas metode klasifikasi untuk menentukan masa studi peserta didik sebagai salah satu upaya penentuan keputusan dan strategi optimal untuk meminimalisir keterlambatan kelulusan peserta didik. Dalam penelitian terdahulu telah dibuktikan bahwa terdapat beberapa metode yang dianggap cukup baik dan hasil yang cukup akurat diantaranya metode klasifikasi Naive Bayes, Decision Tree dan KNN. Metode Klasifikasi Decision tree yang telah dibahas dalam penelitian terdahulu meliputi algoritma C4.5 yang dikembangkan dan disempurnakan menjadi algoritma C5.0. Penelitian ini bertujuan untuk menyusun model algoritma klasifikasi C5.0 dan Naïve Bayes yang akan digunakan untuk mengenal dan memprediksi pola lulusan peserta didik.

Berdasarkan penjabaran diatas, maka perlu disusun model algoritma data mining yaitu algoritma klasifikasi C5.0 dan Naïve Bayes yang bertujuan untuk

melakukan prediksi terhadap pola kelulusan mahasiswa khususnya di Jurusan Matematika UIN Walisongo yang selanjutnya dapat menjadi suatu referensi penyusunan strategi dalam proses pembelajaran, serta memprediksi dan menganalisis masa studi mahasiswa. Kebijakan – kebijakan baru sebagai tindakan preventif bagi pemangku kepentingan dapat disusun berdasarkan hasil prediksi dan analisis masa studi mahasiswa untuk mengurangi resiko mahasiswa tidak lulus tepat waktu.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang yang dijabarkan diatas, maka dapat dirumuskan permasalahan sebagai berikut :

a. Bagaimana model algoritma klasifikasi C5.0 untuk memprediksi pola lulusan mahasiswa jurusan matematika UIN Walisongo Semarang ?
b. Bagaimana model algoritma klasifikasi Naïve Bayes untuk memprediksi pola lulusan mahasiswa jurusan matematika UIN Walisongo Semarang ?
c. Bagaimana hasil prediksi model algoritma klasifikasi C5.0 dan Naïve Bayes untuk memprediksi pola masa studi mahasiswa jurusan matematika UIN Walisongo Semarang ?

1.3 PEMBATASAN MASALAH

Beberapa pembatasan masalah dalam penelitian ini agar ruang lingkup tidak terlalu luas diantaranya adalah,

a. Semua proses perhitungan menggunakan teknik data mining dengan algoritma C5.0 dan metode Naïve Bayes.
b. Data yang digunakan untuk menentukan pola lulusan mahasiswa adalah data *training* yang meliputi data alumni mahasiswa prodi matematika dari Angkatan tahun 2015 hingga tahun 2017.

c. Prodi yang menjadi objek penelitian adalah prodi matematika UIN Walisongo Semarang.
d. Hasil prediksi yang didapatkan adalah keputusan tepat atau tidak tepat waktu mahasiswa dalam menyelesaikan masa studinya.

1.4 TUJUAN PENELITIAN

Tujuan dari penelitian ini adalah untuk :

a. Menyusun model algoritma klasifikasi C5.0 untuk memprediksi pola lulusan mahasiswa jurusan matematika UIN Walisongo Semarang.
b. Menyusun model algoritma klasifikasi Naïve Bayes untuk memprediksi pola lulusan mahasiswa jurusan matematika UIN Walisongo Semarang.
c. Untuk mengetahui hasil prediksi yang terbaik dari kedua algoritma klasifikasi yang mempengaruhi pola lulusan mahasiswa mahasiswa jurusan matematika UIN Walisongo Semarang.

BAB II

LANDASAN TEORI

2.1 DATA MINING

Sebagai bidang ilmu yang relative baru, saat ini Data Mining menjadi salah satu perhatian para akademisi maupun praktisi. Cara pandang dan pengetahuan yang berbeda membuat para ahli memberikan definisi yang berbeda tentang Data Mining. Menurut Fayyad et all dalam Suyanto (2018), Data Mining merupakan suatu tahapan yang menganalisa proses penemuan pengetahuan di dalam *database* atau yang disebut *knowledge discovery in databases* yang disingkat KDD.

Secara definisi Data Mining juga diartikan sebagai suatu ilmu yang menganalisis dan meninjau kumpulan data untuk menemukan hubungan yang tidak terduga dan meringkas data dengan cara berbeda dengan sebelumnya, sehingga dapat dipahami dan bermanfaat bagi pemilik data. Data Mining merupakan irisan dari beberapa bidang keilmuan yang menyatukan teknik dari machine learning, pengenalan pola, statistic, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar. (Larose, 2005)

Sedangkan menurut Pramudiono (2006), Data Mining adalah analisis otomatis dari data yang berhitung besar atau kompleks memiliki tujuan untuk mendapatkan pola atau kecenderungan yang sering tidak disadari keberadaanya. Dalam beberapa tahun terakhir, data berkembang semakin pesat secara eksponensial, semakin beragam dan kompleks. Volume data yang besar dan beragam saai ini dikenal dengan istilah *Big Data*. *Big Data* mempunyai beberapa ciri yaitu data berukuran sangat besar *(high volume)*,

memiliki karakter sangat bervariasi (*high variety*), kecepatan pertumbuhannya tinggi (*high velocity*), dan menyangkut kevalidan data dimana semakin besar suatu data makan berakibat semakin tidak akurat data itu (*high varicity*).

Secara umum, data mining dapat dikelompokkan berdasarkan fungsi atau kegunaannya menjadi beberapa kelompok :

a. Deskriptif

Data Mining memiliki peran penting dalam mencari pola-pola yang dapat dipahami oleh manusia untuk menjelaskan karakteristik suatu data. Deskripsi dari pola yang diberikan menjelaskan kemungkinan-kemungkinan suatu pola atau kecenderungan.

b. Prediktif

Prediktif berarti bahwa data mining digunakan untuk membentuk suatu model pengetahuan yang akan digunakan untuk melakukan prediksi.

Data mining mempunyai banyak tugas yang bisa digabungkan dan digunakan bersama sama dalam beberapa kasus untuk menyelesaikan permasalahan yang ada (MacLennan,dkk,2009). Beberapa tugas-tigas dari data mining adalah,

1) Klasifikasi

Fungsi dari klasifikasi adalah untuk menggolongkan suatu data ke dalam kategori.

2) Klustering

Fungsi dari klustering adalah mengelompokkan objek yang memiliki kemiripan atribut pada kelompok yang berbeda.

3) Association

Fungsi dari association adalah untuk mencari aturan asisiasi yang dapat mengidentifikasi item menjadi objek.

4) Regression

Fungsi regression adalah hampir sama dengan klasifikasi yanitu bertujuan untuk mencari prediksi dari suatu pola.

5) Forecasting

Fungsi forecasting adalah untuk melakukan prediksi / peramalan untuk data di waktu yang akan datang berdasarkan trend yang telah terjadi di waktu sebelumnya.

6) Sequence Analysis

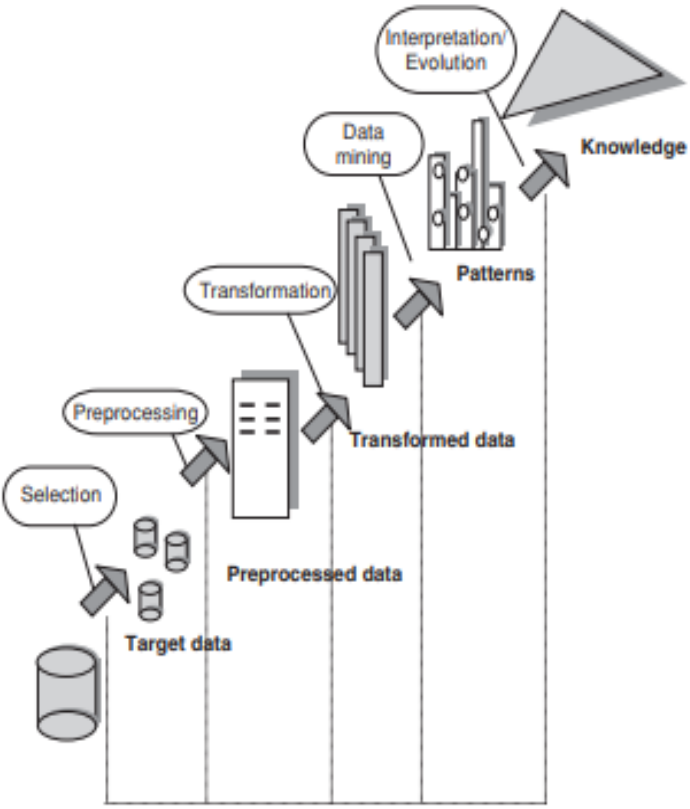
Fungsi dari sequence analysis adalah untuk mencari pola urutan dari rangkaian kejadian.

7) Deviasion Analysis

Deviasion analysis memiliki fungsi untuk mencari kejadian yang jarang ditemukan atau kejadian yang berbeda dari keadaan normal.

2.2 TAHAPAN PROSES DATA MINING

Data Mining merupakan suatu rangkaian proses yang terdiri atas beberapa tahapan. Diagram dibawah ini menggambarkan beberapa tahapan proses yang berlangsung di dalam Data Mining. Fase awal dimulai dari data sumber dan berakhir dengan adanya informasi yang dihasilkan dari beberapa tahapan sebagai berikut :



Gambar 2.1 Tahapan proses Data Mining
(Introduction to Data Mining and Its Application,
S.Sumathi, S.N. Sivanandam)

Tahapan proses dalam Data Mining dapat dijelaskan sebagai berikut :

a. Seleksi Data (*Data Selection*)

Pada tahap ini dilakukan pemilihan (seleksi) data yang ada dalam *database*. Data yang tersaji seringkali tidak seluruhnya dipakai. Oleh karena itu, diperlukan tahap penyeleksian untuk mendapatkan data yang relevan untuk dianalisa. Pada tahap ini dimulai dengan

menentukan tujuan dari proses mining agar proses pengolahan menjadi lebih baik sesuai dengan tujuan penelitian yang akan dicapai.

b. Pemilihan Data (*Preprocessing / Cleaning*)

Pada *preprocessing* mencakup tahapan-tahapan antara lain menghilangkan duplikasi data, menyelidiki data yang sifatnya inkonsisten dan memperbaiki kesalahan data, misalnya terjadi kesalahan cetak / *typography*. Data yang tidak valid maupun data yang tidak diperlukan akan dibuang. Oleh karena itu, tahapan *preprocessing* diperlukan untuk memilih data atau informasi yang relevan dan diperlukan.

c. Transformasi (*Transformation*)

Tahapan transformasi meliputi proses *encoding* atau konversi. Pada tahap ini, dilakukan perubahan atau penggabungan data ke dalam format data yang valid atau siap diaplikasikan.

d. Data Mining

Proses penambangan atau *mining* adalah sutu proses utama yang mengaplikasikan metode, teknik dan algoritma tertentu untuk mengekstrasi pola yang diinginkan.

e. Interpretasi / Evaluasi

Pola data dan informasi yang dihasilkan pada proses penambangan atau *mining* diintrepretasikan dalam bentuk yang mudah dimengerti oleh pihak yang berkaitan. Pada fase ini meliputi identifikasi pola data dan informasi yang ditemukan telah sesuai dengan fakta atau hipotesa. Jika didapatkan bahwa hasil tidak sesuai dengan hipotesa , maka terdapat beberapa alternatif yang dapat dipilih untuk memperbaiki proses data mining atau menggunakan metode data mining yang lain.

Selanjutnya dilakukan proses visualisasi dan interpretasi informasi yang telah digali oleh pengguna. Langkah terakhir adalah memformulasikan keputusan dari hasil analisis yang diperoleh.

2.3 KLASIFIKASI

Klasifikasi merupakan bagian yang sangat penting dalam data mining. Tujuan utama dari klasifikasi adalah mempelajari sekumpulan data sehingga dihasilkan aturan yang dapat mengelompokkan data ke dalam kelas-kelas serta mengenali data – data baru yang belum pernah dipelajari (Suyanto,2019). Secara definisi, klasifikasi adalah suatu proses yang menyatakan suatu objek data sebagai salah satu kategori atau kelas yang telah didefinisikan sebelumnya (Zaki et all,2013). Klasifikasi dapat diartikan juga sebagai proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.

Di dalam Teknik klasisifikasi terdapat dua proses yang digunakan yaitu proses *learning* (tahap *training*) dan proses klasifikasi. Pada proses *learning* , algoritma klasifikasi dibangun untuk menganalisa data *training* yang kemudian direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah proses klasifikasi dimana data *testing* digunakan untuk mengestimasi akurasi dari *rule* klasifikasi.

Model klasifikasi dapat dibangun berdasarkan pengetahuan seorang pakar atau ahli. Namun, pada aplikasinya model klasifikasi lebih sering dibangun menggunakan Teknik pembelajaran dalam bidang *machine learning* dikarenakan himpunan data yang digunakan relative besar. Proses pembelajaran terhadap suatu himpunan data mampu menghasilkan model klasifikasi yang memetakan objek data *x* (input) ke salah satu kelas *y* yang telah didefinisikan sebelumnya. Sehingga , proses pembelajaran ini membutuhjan masukan (input) berupa himpunan data *training* yang memiliki atribut kelas dan menghasilkan *output* berupa model klasifikasi (Suyanto,2019).

2.4 DECISION TREE

Decision Tree merupakan salah satu metode klasifikasi data mining yang paling sering digunakan karena mudah diaplikasikan oleh pengguna dalam interpretasi. *Decision tree* adalah teknik data mining yang membangun representasi aturan klasifikasi berstruktur sekuensial hirarki dengan cara mempartisi himpunan data *training* secara rekursif (Aggarwal,2015). *Decision Tree* menghasilkan *flowchart* dimana strukturnya mirip seperti struktur pohon yang terdiri dari simpul-simpul (*node*) yang membentuk pohon berakar. Setiap simpul atau *node* mempresentasikan atribut yang telah diuji, setiap cabang merupakan hasil uji dan titik akhir merupakan pembagian kelas yang dihasilkan (Han dan Kamber, 2012).

Algoritma *Decision Tree* terdiri dari kumpulan simpul (*node*) yang dihubungkan oleh cabang, dimana cabang tersebut bergerak dari bawah simpul *root* (akar) dan berakhir di simpul *leaf* (daun). Simpul *leaf* memuat suatu keputusan akhir atau kelas target untuk suatu *decision tree*. Sedangkan simpul *root* merupakan titik awal dari suatu *decision tree*. Dan terdapat satu simpul penting yaitu simpul perantara yang menghubungkan dengan suatu pertanyaan atau pengujian (Rokach and Maimon,2015).

Berbagai algoritma *Decision Tree* sudah dikembangkan seperti CART, C4.5 dan ID3. Pada akhir tahun 1970 sampai awal tahun 1980, J. Ross Quinlan seorang peneliti membuat algoritma *Decision Tree* yang dikenal dengan ID3 (*Iterative Dichotomiser*). Quinland kemudian membuat algoritma C4.5 yang merupakan pengembangan dari algoritma ID3 (Han and Kamber,2012). Secara prinsip jenis *Decision Tree* dikelompokkan menjadi *regression tree* dan *Classification Tree*. Tahapan pembentukan *Tree* baik untuk *regression* maupun *classification* adalah,

a) Menentukan *root node.*
b) Dataset diletakkan pada *root node*.
c) Membagi (*splitting*) dataset menjadi subset.
d) Menentukan *decision node*.
e) Proses diulang hingga *stopping criteria* tercapai.

2.4.1 Entropy

Secara istilah, *entropy* adalah keberbedaan atau keberagaman. Dalam data mining *entropy* didefinisikan sebagai suatu parameter untuk mengukur heterogenitas (keberagaman) dalam suatu himpunan data. Semakin heterogeny suatu himpunan data, semakin besar pula nilai *entropy*-nya. Secara matematis, *entropy* dirumuskan sebagai,

$$entropy(S) = \sum_i^c - p_i \, log_2 \, p_i \qquad (2.1)$$

Dimana c adalah jumlah nilai yang terdapat pada atribut target (jumlah kelas). Sedangkan p_i menyatakan porsi atau rasio antara jumlah sampel kelas i dengan jumlah semua sampel pada himpunan data (Aggarwal,2015).

Himpunan data yang memiliki dua kelas dengan jumlah sampel di kelas pertama sama dengan jumlah sampel di kelas kedua akan memiliki *entropy* yang maksimum (yaitu sama dengan 1). Artinya, himpunan data tersebut memiliki keberagaman maksimum. Sebaliknya, himpunan data yang memiliki dua kelas dengan jumlah sampel pada salah satu kelas adalah 0 akan memiliki *entropy* yang minimum (yaitu sama dengan 0). Artinya, himpunan tersebut memiliki keberagaman minimum.

2.4.2 *Information Gain*

Secara definisi, *Information Gain* adalah perolehan informasi. Dalam data mining, *Information Gain* menunjukkan berapa banyak informasi pada suatu

feature tertentu dapat memberikan informasi tentang kelas. *Information Gain* juga diartikan sebagai ukuran efektivitas suatu atribut dalam mengklasifikasikan data (Rokach and Maimon, 2015).

Secara matematis, *Information Gain* dari suatu atribut *A*, dituliskan sebagai :

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2.2)$$

Dimana,

A	= atribut
V	= suatu nilai yang mungkin untuk atribut *A*
$Values(A)$	= himpunan nilai – nilai yang mungkin untuk atribut *A*
$\|S_v\|$	= jumlah sampel untuk nilai *v*
$\|S\|$	= jumlah seluruh sampel data
$Entropy(S_v)$	= entropy utnuk sampel-sampel yang memiliki *v*.

Information Gain akan mengalami maslah untuk atribut yang nilainya sangat bervariasi. Untuk mengatasi permasalahan tersebut dapat digunakan ukuran lain yang disebut *Gain Ratio* yang dihitung berdasarkan *Split Information* yang dirumuskan sebagai berikut :

$$SplitInformation(S, A) = \sum_{i=1}^{c} - \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \quad (2.3)$$

Dimana *S* menyatakan himpunan sampel data, S_i sampai S_c menyatakan subhimpunan sampel data yang terbagi berdasarkan jumlah variasi nilai pada atribut *A*. Selanjutnya *Gain Ratio* dirumuskan sebagai *Information Gain* dibagi dengan *Split Information* sebagai (Mitchell, 1997) :

$$GainRatio(S, A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (2.4).$$

2.5 ALGORITMA C5.0

Algoritma C5.0 adalah salah satu algoritma data mining yang diaplikasikan pada algoritma *decision tree*. Algoritma C5.0 merupakan pengembangan dan penyempurnaan dari algoritma yang dibangun oleh Ross Quinlan pada tahun 1987 yaitu algoritma ID3 dan algoritma C4.5. Algoritma ID3 pertama dikembangkan menjadi algoritma C4.5 mampu menyelesaikan persoalan klasifikasi data mining dengan atribut yang dimiliki bertipe diskrit dan kontinu. Keunggulan dari algoritma C4.5 adalah mampu menangani *missing value* dengan cara memberikan nilai berdasarkan nilai yang paling dominan. Algoritma C4.5 juga mampu memmangkas *decision tree* yang digunakan untuk mengatasi masalah *over fitting* yang terjadi karena data yang digunakan tidak relevan pada data *training* (Benediktus dan Oetama,2020).

Meskipun Algoritma C4.5 memiliki beberapa keunggulan, namun terdapat kelemahan seperti terjadinya *overlapping* terutama pada saat data yang diolah sangat banyak, serta menyebabkan meningkatnya waktu pengambilan keputusan. Oleh karena itu dikembangkanlah algoritma C4.5 menjadi algoritma C5.0, dimana tingkat akurasi yang dimiliki algoritma C5.0 lebih tinggi, waktu pengambilan keputusan lebih cepat dan penggunaan memori jauh lebih rendah dari algoritma sebelumnya (Kumar dan Kiruthika,2015).

Algoritma C5.0 bekerja dengan cara memilih atribut berdasarkan nilai *gain ratio* tertinggi. Algoritma C5.0 memiliki keunggulan dibandingkan algoritma sebelumnya yaitu diantaranya mampu mengatasi masalah seperti *missing value* dan data yang relative besar. Algoritma ini juga dapat melakukan *training* data dalam waktu yang cepat untuk digunakan dalam *testing* data (Myint and Tin,2021). Perbandingan antara

beberapa Algoritma dalam *Decision Tree* tersaji dalam tabel berikut,

Tabel 2.1 Perbandingan Algoritma ID3, C4.5 dan C5.0
Sumber : Kumar dan Kiruthika (2015)

Algoritma	ID3	C4.5	C5.0
Tipe data yang digunakan	Kategorik	Kontinu dan Kategorik	Kontinu, Kategorik, dan periode waktu
Tingkat kecepatan dalam perhitungan	Rendah	Lebih cepat dibanding ID3	Paling Tinggi
Pruning (proses pemangkasan *decision tree*)	Tidak	Pre-pruning	Pre-pruning
Metode Boosting	Tidak didukung	Tidak didukung	Didukung
Formula	Menggunakan *information gain* dan *entropy*	Menggunakan *Split info* dan *gain ratio*	Menggunakan *Split info* dan *gain ratio*

Terdapat beberapa tahapan dalam membuat *decision tree* menggunakan algoritma C5.0 yaitu :

1) Mempersiapkan data *training*. Data *training* seringkali diambil dari data histori yang pernah terjadi sebelumnya dan telah dikelompokkan ke dalam kelas-kelas tertentu.

2) Menentukan akar (*root*) dari pohon. Akar diambil dari atribut yang terpilih dengan cara menghitung nilai *gain* dari masing-masing atribut. Nilai *gain* yang paling tinggi yang akan menjadi akar pertama.
3) Menghitung nilai *gain* menggunakan formula (2.2)
4) Menentukan cabang masing-masing atribut dengan cara yang sama seperti atribut akar pada Langkah (2).
5) Mengulangi Langkah ke-2. Kelas dibagi ke dalam cabang dan jika terdapat cabang yang memiliki dua kelas maka yang terpilih adalah kelas yang terbanyak dan proses diulang untuk masing-masing cabang hingga semua kelas pada cabang memiliki kelasnya masing-masing.

2.6 ALGORITMA NAÏVE BAYES

Algoritma Naïve Bayes merupakan salah satu algoritma klasifikasi yang digunakan untuk memprediksi probabilitas keanggotaan suatu *class* (Han and Kamber,2012). Algoritma Naïve Bayes atau yang disebut Naïve Bayes Classifier berasal dari teorema Bayes yang ditemukan oleh Thomas Bayes pada tahun 1770. Teorema Bayes merupakan sebuah teorema dengan dua penafsiran yang berbeda. Teorema Bayes menyatakan seberapa jauh derajat kepercayaan subjektif harus diubah secara rasional saat diberikan petunjuk baru (Primartha, 2020).

Dalam teorema Bayes, probabilitas atau peluang bersyarat dinyatakan sebagai,

$$P(H|X) = \frac{p(X|H)\ .\ P(H)}{P(X)} \qquad (2.5)$$

Dimana X adalah suatu bukti dan H adalah hipotesis. $P(H|X)$ adalah probabilitas bahwa hipotesis H benar untuk bukti X atau dengan kata lain $P(H|X)$

adalah probabilitas posterior H dengan syarat X. $P(X|H)$ adalah probabilitas bahwa bukti X benar untuk hipotesis H atau probabilitas posterior X dengan syarat H. $P(X)$ adalah probabilitas prior bukti X dan $P(H)$ adalah probabilitas prior hipotesis *H*. Probabilitas prior adalah nilai probabilitas yang diyakini benar sebelum melakukan eksperimen terhadap sesuatu. Apabila setelah dilakukan eksperimen mengakibatkan adanya perubahan terhadap nilai probabilitas tersebut, maka hal ini disebut sebagai probabilitas posterior (Han and Kamber,2012).

Dalam data mining, X menyatakan suatu *tuple* atau objek data, H adalah dugaan atau hipotesis bahwa *tuple* atau objek data X adalah kelas C. Secara spesifik, dalam masalah klasifikasi dapat dihitung $P(H|X)$ sebagai probabilitas bahwa hipotesis *H* benar untuk *tuple X* atau dengan kata lain $P(H|X)$ adalah probabilitas bahwa *tuple X* berada dalam kelas *C*. Selanjutnya, $P(H)$ adalah probbailitas prior bahwa hipotesis *H* benar untuk setiap *tuple* tidak memperdulikan nilai-nilai atributnya sedangkan $P(X)$ adalah probabilitas prior dari *tuple X*.

Tahapan dari algoritma Naïve Bayes adalah sebagai berikut :

1. Mempersiapkan data *training*.

Misalkan *D* adalah himpunan data *training* yang berisi sejumlah *tuple* beserta label kelasnya. Setiap *tuple* berdimensi *n* yang dinyatakan sebagai $X = (x_1, x_2, \cdots, x_n)$ yang didapat dari *n* atribut $A_1, A_2, \cdots, A_n$.

2. Menghitung jumlah kelas dalam data *training*.

Misalkan diketahui terdapat *m* kelas yaitu $C_1, C_2, \cdots, C_m$. Untuk sebuah *tuple* masukkan *X*, Algoritma Naïve Bayes memprediksi bahwa *tuple X* termasuk di dalam kelas C_i jika dan hanya jika

$$P(C_i|X) > P(C_j|X) \; untuk \; \leq j \; \leq m, j \neq i$$

Dengan kata lain, algoritma Naïve Bayes memaksimalkan $P(C_i|X)$. Kelas C_i yang menyebabkan

$P(C_i|X)$ bernilai maksimal disebut dengan *maximum posteriori hyphotesis*. Sehingga estimasi $P(C_i|X)$ menggunakan Teorema Bayes adalah sebagai berikut,

$$P(C_i|X) = \frac{P(X|C_i).P(Ci)}{P(X)} \quad (2.6)$$

3. Menghitung jumlah kasus yang sama dalam kelas yang sama.

Nilai probabilitas $P(X)$ bernilai sama untuk semua kelas artinya *tuple X* mempunyai probabilitas yang sama untuk masuk ke dalam kelas manapun. Oleh karena itu, hanya nilai $P(X|C_i).P(Ci)$ yang perlu dimaksimalkan. Jika probabilitas prior untuk setiap kelas tidak diketahui, maka probabilitas setiap kelas diasumsikan sama, $P(C_1) = P(C_2) = ... = P(C_m)$. Dengan demikian, Naïve Bayes hanya memaksimalkan $P(X|C_i)$. Namun jika probabilitas prior untuk setiap kelas berbeda-beda, maka Naïve Bayes harus memaksimalkan $P(X|C_i).P(Ci)$.

4. Proses Klasifikasi

Dalam tahap ini, jika didalam data *training* memiliki sangat banyak atribut, maka dapat dilakukan reduksi kompleksitas perhitungan $P(X|C_i)$ dengan asumsi bahwa setiap atribut saling independent atau saling bebas. Dengan demikian, Naïve Bayes memaksimalkan

$$P(C_i|X) = \prod_{k=1}^{n} P(x_k|C_i)$$

$$= P(x_1|C_i).P(x_2|C_i)....P(x_n|C_i) \quad (2.7)$$

Berdasarkan *tuple-tuple* pada data *training* dapat diestimasi nilai $P(x_1|C_i), P(x_2|C_i), ..., P(x_n|C_i)$ sesuai tipe atributnya, diantaranya

a) Untuk atribut yang bernilai kategorial, perhitungan digunakan persamaan (2.1). $P(x_k|C_i)$ didefinisikan sebagai jumlah *tuple* di dalam kelas C_i dalam data *training D* yang memiliki nilai x_k

pada atribut A_k dibagi dengan jumlah semua *tuple* di kelas C_i dalam D yang disimbolkan $|C_{i,D}|$.

b) Untuk atribut yang bernilai kontinu digunakan perhitungan distribusi Gaussian. Dalam perhitungan ini, terlebih dahulu dihitung rata-rata μ dan standart deviasi σ sesuai dengan persamaan berikut,

$$P(x_k|C_i) = \frac{1}{\sigma_{C_i}\sqrt{2\pi}}\, e^{-\frac{(x-\mu_{C_i})}{2\,\sigma {C_i}^2}} \quad (2.8)$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2.9)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \quad (2.10)$$

Dimana μ_{C_i} adalah rata-rata dan σ_{C_i} adalah standart deviasi dari nilai-nilai atribut A_k untuk kelas C_i.

5. Prediksi label kelas dari *tuple X*.

Langkah selanjutnya adalah menghitung probabilitas $P(X|C_i).P(Ci)$ untuk setiap kelas C_i. Selanjutnya, memaksimalkan probabilitas tersebut yaitu dengan mencari kelas C_i yang menghasilkan probabilitas $P(X|C_i).P(Ci)$ maksimum sebagai kelas keputusan. Secara matematis *tuple X* diberi label kelas C_i jika dan hanya jika

$$P(X|C_i).P(Ci) > P\big(X|C_j\big).P(C_j) \quad (2.11)$$

untuk $1 \leq j \leq m, j \neq i$.

6. Membandingkan hasil per kelas, nilai probabilitas tertinggi ditetapkan sebagai kelas baru (Han and Kamber,2012).

2.7 EVALUASI DAN SELEKSI MODEL KLASIFIKASI

Dalam membangun model klasifikasi terkadang terdapat beberapa masalah dan muncul pertanyaan terkait bagaimana mengevaluasi semua metode klasifikasi tersebut agar diperoleh metode klasifikasi terbaik. Evaluasi terhadap metode klasifikasi umumnya dilakukan menggunakan sebuah himpunan data testing, yang tidak digunakan dalam data set klasifikasi tersebut dengan suatu ukuran tertentu. Terdapat sejumlah ukuran yang dapat digunakan untuk menilai atau mengevaluasi model klasifikasi diantaranya adalah nilai akurasi, nilai kesalahan atau kekeliruan klasifikasi, nilai *recall* atau sensitivitas dan lainnya (Suyanto,2019). Ukuran evaluasi model klasifikasi secara ringkas dapat diilustrasikan pada Tabel 2.2 berikut

Tabel 2.2 Ukuran Evaluasi Model Klasifikasi

No	Ukuran	Rumus
1	Nilai Akurasi	$\frac{TP + TN}{P + N}$
2	Error rate atau Tingkat kesalahan atau kekeliruan klasifikasi	$\frac{FP + FN}{P + N}$
3	*Recall* atau *sensitivity* atau *true positive rate*	$\frac{TP}{P}$
4	*sensitivity* atau *true negative rate*	$\frac{TN}{N}$

5	*Precision*	$\frac{TP}{TP + FP}$
6	F atau F_1 atau F-Score atau rata-rata harmonic dari precision dan recall	$\frac{2 \times precision \times recall}{precision + recall}$
7	F_β dimana β adalah sebuah bilangan riil non negatif	$\frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision\ x\ recall}$

Empat istilah yang penting untuk ukuran evaluasi dalam tabel 2.2 adalah sebagai berikut :

1. TP atau *True Positives* adalah jumlah tuple positif yang dilabeli dengan benar oleh *classifier*. Tuple positif yang dimaksud adalah tuple actual yang berlabel positif.
2. TN atau *True Negatives* adalah jumlah tuple negative yang dilabeli dengan benar oleh *classifier*. Tuple negative yang dimaksud adalah tuple actual yang berlabel negative.
3. *FP* atau *False Positives* adalah jumlah tuple negative yang salah dilabeli oleh *classifier*. Sebagai contoh misalkan tuple atribut X_1 yang diberi label *A* = "Tidak" namun oleh classifier diberi label *A* = "Ya".
4. *FN* atau *False Negatives* adalah jumlah tuple positif yang salah dilabeli oleh *classifier*. Sebagai contoh misalkan tuple atribut X_1 yang diberi label *A* = "Ya" namun oleh classifier diberi label *A* = "Tidak".

Dari ke empat istilah diatas kemudian dibentuk *Confussion Matrix* yang disajikan pada Tabel 2.3.

Tabel 2.3 Confussion Matrix untuk Evaluasi Model Klasifikasi

	Kelas hasil prediksi			
		Ya	Tidak	Jumlah
Kelas Aktual	Ya	TP	FN	P
	Tidak	FP	TN	N
	Jumlah	P’	N’	P + N

Confussion Matrix digunakan untuk menganalisis kualitas *classifier* dalam mengenali tuple-tuple dari kelas yang ada.

2.8 MASA STUDI DAN KELULUSAN MAHASISWA

Masa atau lama studi merupakan waktu yang dibutuhkan oleh mahasiswa untuk menyelesaikan studinya di pendidikan tinggi. Beberapa faktor dapat mempengaruhi lama studi mahasiswa antara lain, faktor internal dan faktor eksternal dari mahasiswa, misalnya faktor keterpaksaan dalam proses kuliah, salah dalam pemilihan jurusan, bahkan juga terlalu aktif dalam organisasi kemahasiswaan. Masa studi mahasiswa UIN Walisongo Semarang untuk program S-1 dapat ditempuh selama 8 semester atau 4 tahun dan paling lama ditempuh selama 14 semester atau 7 tahun. Beban studi mahasiswa program pendidikan S-1 yang harus diambil adalah minimal 144 satuan kredit semester (SKS) dan maksimal 148 satuan kredit semester (SKS).

Mahasiswa diwajibkan untuk menempuh perkuliahan dan kegiatan akademik yang sejenis sesuai dengan rencana studi dengan tertib menurut peraturan dan ketentuan yang diberlakukan. Perkuliahan dapat dibedakan menjadi kegiatan kuliah berbasis teori dan kegiatan kuliah berbasis proyek serta praktikum atau kerja lapangan. Kelulusan mahasiswa ditandai dengan selesainya pendidikan pada jenjang sarjana yang

merupakan akhir pencapaian mahasiswa dalam menempuh studi. Mahasiswa dinyatakan lulus apabila ketentuan – ketentuan tersebut diatas telah dilaksanakan dan tidak terdapat tanggungan administrasi di bagian akademik.

2.9 *SOFTWARE* RAPIDMINER

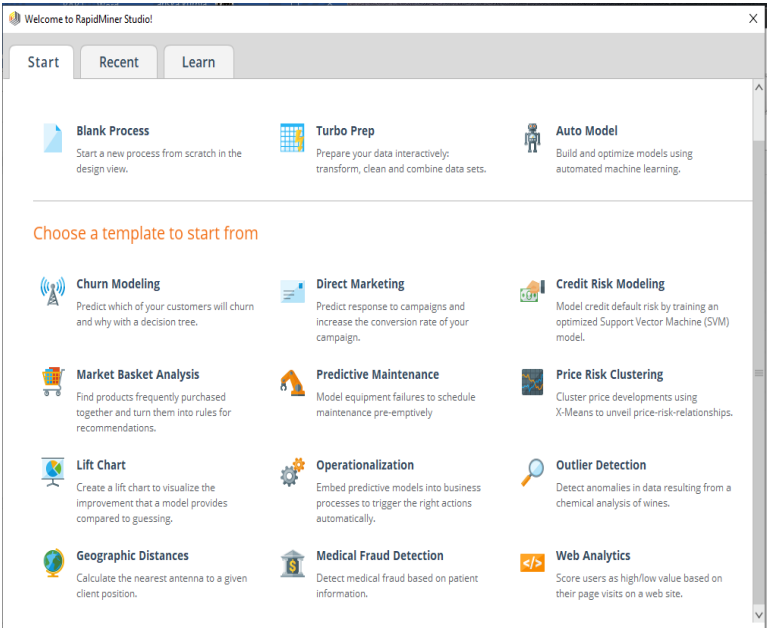
RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). *Software* RapidMiner merupakan sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner merupakan *software* yang menggunakan teknik deskriptif dan prediksi dalam memberikan informasi atau pengetahuan baru kepada *user* sehingga dapat digunakan untuk mengambil keputusan yang terbaik. Sebagai mesin data mining yang terintegrasikan dengan produknya sendiri, *software* RapidMiner mampu berdiri sendiri dalam analisis data dan berbasis Bahasa Java.

Software RapidMiner memiliki beberapa *tools* yang digunakan untuk *pre-processing* data, klasifikasi, *clustering*, regresi aturan asosiasi dan visualisasi. RapidMiner sangat membantu *user* dalam menyelesaikan tahapan-tahapan dalam data mining seperti melakukan *pre-processing* data, menginputkannya dalam skema pembelajaran, dan menganalisa *classifier* yang dihasilkan serta mengukur perfomanya tanpa menuliskan kode program.

RapidMiner mempunyai beberapa sifat diantaranya :

1. Ditulis dengan Bahasa pemrograman Java sehingga dapat dioperasikan dalam berbagai system operasi.
2. Proses penemuan pengetahuan dimodelkan sebagai operator *trees*.
3. Bahasa *scripting* sangat memungkinkan untuk eksperimen skala besar dan otorisasi eksperimen.

4. Memiliki GUI , *command line mode*, dan Java API yang dapat dipanggil dari program yang lain.



Gambar 2.2 Tampilan Awal *Software RapidMiner*

Bagian awal tampilan RapidMiner disambut dengan tampilan *Welcome Perspective*. Pada bagian toolbar, terdapat toolbar Perspectives yang terdiri atas ikon-ikon yang menampilkan perspective dari RapidMiner.

2.10 KAJIAN PENELITIAN TERDAHULU

Literasi sains sudah mulai berkembang luas dan menjadi tujuan utama dari negara berkembang. Beberapa penelitian terkait prediksi pola lulusan mahasiswa telah banyak dilakukan. Beberapa penelitian yang dapat digunakan peneliti sebagai bahan kajian yang akan dilakukan diantaranya,

1. Penelitian oleh Mongan Winny Amelia dkk (2017) dengan judul "Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes". Jurnal Teknik Informatika Vol 11 No 1 ISSN : 2301-8364 menggunakan algoritma Naïve Bayes dalam memprediksi masa studi mahasiswa. Dari hasil penelitian menunjukkan bahwa Algoritma Naïve Bayes cukup baik dalam menentukan prediksi masa studi mahasiswa dengan tingkat akurasi 85,17%. Atribut yang digunakan meliputi program studi, nilai ip, semester dan jumlah sks. Pengujian model menggunakan *Cross Validation 10-Fold* yang diambil dan diuji pada bagian jumlah semester yang dilakukan prediksi. Dari proses mining menggunakan Algoritma Naïve Bayes dikembangkan menjadi suatu aplikasi berbasis Bahasa pemrograman PHP yang dapat dimanfaatkan *user* untuk mengolah data mahasiswa dan memprediksi status kelulusan mahasiswa berdasarkan data history sebelumnya. Berdasarkan penelitian yang dilakukan oleh Mongan Winny Amelia dkk menunjukkan bahwa algoritma Naïve Bayes merupakan algoritma yang cukup tepat dan baik untuk klasifikasi data mining.
2. Penelitian oleh Farida, Ida dan Spits Warnars H.L.H. (2019) dengan judul "*Prediksi Pola Kelulusan Mahasiswa Menggunakan Teknik Data Mining Classification Emerging Pattern".* Jurnal Petir Vol 12 No 1 P-ISSN : 1978-9262 E-ISSN : 2655-5018 menggunakan teknik data mining *Classification Emerging Pattern* untuk menganalisa pola dari data set. Atribut yang digunakan dalam penelitian Ida Farida dkk meliputi jenis kelamin, Angkatan, program studi, system kuliah, IPK, dan status kelulusan. Hal ini bisa dijadikan acuan analisis dari pihak universitas dalam mengambil tindakan strategis dalam menaikan pattern kelulusan tepat waktu pada mahasiswa dengan system kuliah Reguler. Dari hasil tersebut diketahui bahwa

high level management bisa menentukan rencana yang diterapkan untuk meningkatkan kelulusan mahasiswa dengan waktu yang sesuai dengan menganalisa data yang telah ada sebelumnya.

3. Penelitian oleh Reni Pratiwi (2019) dengan judul "Perbandingan Klasifikasi Algoritma C5.0 dan Regression tree (Studi Kasus : Data Sosial Kepala Keluarga Masyarakat Desa Teluk baru Kecamatan Muara Ancalong Tahun 2019" Jurnal Barekeng Vol 14 No 2 P-ISSN: 1978-7227 E-ISSN: 2615-3017 menunjukkan rata-rata tingkat akurasi metode CART sebesar 84,63% sedangkan tingkat akurasi algoritma C5.0 hanya 79,17%. Oleh karena itu dapat disimpulkan bahwa metode CART merupakan metode yang lebih baik dalam pengklasifikasikan data rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019 dibandingkan metode algoritma C5.0. Hasil ketepatan klasifikasi rata-rata pendapatan masyarakat Desa Teluk Baru Kecamatan Muara Ancalong tahun 2019 menggunakan algoritma C5.0 dengan proporsi 90:10 memperoleh tingkat akurasi tertinggi sebesar 90%, sedangkan hasil ketepatan klasifikasi rata-rata pendapatan masyarakat menggunakan metode CART dengan proporsi 50:50 memperoleh tingkat akurasi sebesar 94%. Sehingga dapat dikatakan bahwa metode CART merupakan metode yang lebih baik dalam pengklasifikasian data rata-rata pendapatan masyarakat Desa teluk baru Kecamatan Muara Ancalong dibandingkan dengan metode algoritma C5.0.

4. Penelitian oleh Mudasir Ashraf, Majid Zaman dan Muheet Ahmed (2019) dengan judul "*An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering Approaches*" Science Direct Procedia Computer Science 167 (2020) 1471 – 1483 menggunakan Teknik klasifikasi pembanding meliputi J48, K-Nearest Neighbour dan Naïve Bayes dalam

memprediksi kinerja siswa. Tujuan utama dari penelitian ini adalah untuk menunjukkan dampak dari metode ensemble pada nilai akurasi prediksi pengklasifikasi pembelajaran. Dan berdasarkan hasil yang dicapai, metode klasifikasi yang lebih baik pada dataset adalah menggunakan Algoritma Naïve Bayes dengan nilai akurasi prediksi sebesar 95,50% dibandingkan metode klasifikasi yang lain.

5. Penelitian oleh Zizheng Guo et all (2021) dengan judul "*Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management*" membahas aplikasi algoritma C5.0 dan K-means untuk menghasilkan peta Kawasan longsor regional. Teknik data mining yang digunakan adalah Algoritma C5.0 dan K-Means untuk menganalisa pola dari data set dan menunjukkan zonasi yang rentan dengan bencana longsor. Berdasarkan hasil penelitian Zizheng Guo et all, teknik klasifikasi yang cukup akurat dan tepat dalam memprediksi keputusan digunakan untuk dataset selanjutnya yaitu data induk mahasiswa adalah algoritma C5.0 dan Naïve Bayes untuk memprediksi pola lulusan dan kinerja mahasiwa.

Berdasarkan beberapa hasil penelitian terdahulu, dapat disimpulkan bahwa algoritma Naïve Bayes masih menghasilkan prediksi yang akurat. Namun tidak menutup kemungkinan jika algoritma klasifikasi C5.0 yang merupakan pengembangan dari algoritma C4.5 akan menghasilkan prediksi yang lebih baik dari algoritma Naïve Bayes. Selanjutnya pada penelitian ini, peneliti akan melakukan penelitian "Analisis Kinerja Algoritma Klasifikasi C5.0 Dan Naïve Bayes Untuk Pengenalan Pola Lulusan Mahasiswa".

2.11 KERANGKA BERPIKIR

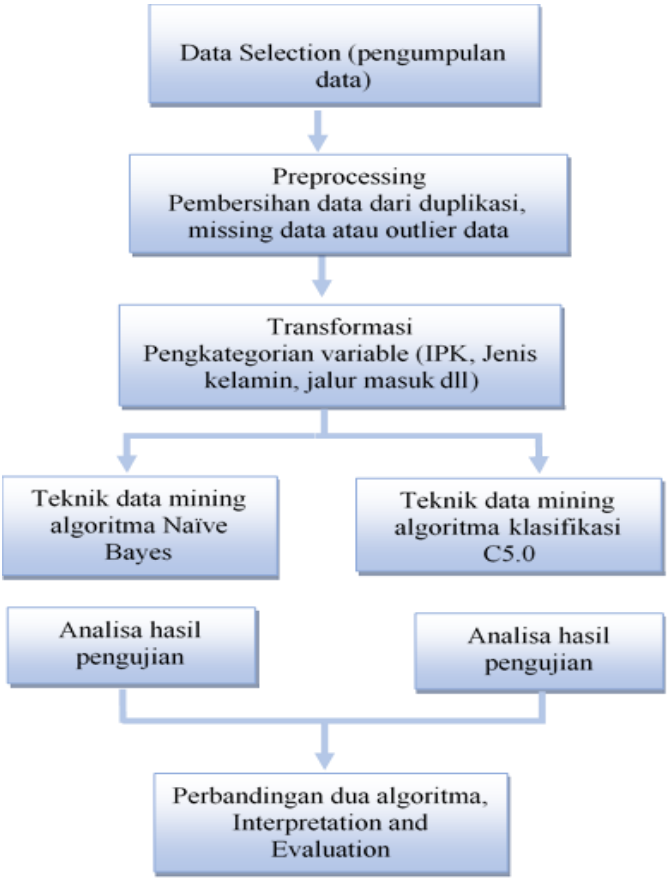
Klasifikasi merupakan bagian yang sangat penting dalam data mining. Tujuan utama dari klasifikasi adalah mempelajari sekumpulan data sehingga dihasilkan aturan yang dapat mengelompokkan data ke dalam kelas-kelas serta mengenali data – data baru yang belum pernah dipelajari (Suyanto,2019). Pada penelitian ini dilakukan beberapa tahapan yang dimulai dengan pengumpulan data berupa data sekunder. Data sekunder terdiri atas data induk akademik mahasiswa dan data biodata mahasiswa Angkatan 2015 dan 2016 yang telah dinyatakan lulus. Objek penelitian ini adalah perbandingan algoritma C5.0 dan Naïve Bayes untuk memprediksi pola lulusan mahasiswa Program Studi Matematika UIN Walisongo Semarang.

Setelah data dikumpulkan dilakukan tahapan *preprocessing* yang merupakan tahapan pembersihan data dari duplikasi, *missing data* dan juga *outlier data*. Data yang telah dibersihkan selanjutnya masuk ke tahapan transformasi. Tahapan transformasi meliputi proses *encoding* atau konversi. Pada tahap ini, dilakukan perubahan atau penggabungan data ke dalam format data yang valid atau siap diaplikasikan. Pengkategorian atribut yang digunakan dalam penelitian ini dibagi menjadi atribut numerik dan atribut nominal. Selanjutanya, masuk pada tahapan proses penambangan atau *mining* yang merupakan suatu proses utama yang mengaplikasikan metode, teknik dan algoritma tertentu untuk mengekstrasi pola yang diinginkan. Algoritma yang dipilih dalam penelitian ini adalah algoritma Naïve Bayes dan algoritma C5.0 yang berdasarkan penelitian terdahulu kedua algoritma ini memiliki tingkat akurasi yang tinggi untuk estimasi atau prediksi.

Kedua algoritma dianalisis dan diinterpretasi serta dievaluasi. Pola data dan informasi yang dihasilkan

pada proses penambangan atau *mining* diintrepretasikan dalam bentuk yang mudah dimengerti oleh pihak yang berkaitan. Pada fase ini meliputi identifikasi pola data dan informasi yang ditemukan telah sesuai dengan fakta atau hipotesa. Jika didapatkan bahwa hasil tidak sesuai dengan hipotesa , maka terdapat beberapa alternatif yang dapat dipilih untuk memperbaiki proses data mining dengan algoritma terpilih. Selanjutnya dilakukan proses visualisasi dan interpretasi informasi yang telah digali oleh pengguna. Langkah terakhir adalah memformulasikan keputusan dari hasil analisis yang diperoleh.

Tahapan terakhir adalah membandingkan kedua algoritma berdasarkan nilai akurasi yang didapatkan dari model algoritma C5.0 dan Naïve Bayes. Seluruh tahapan penelitian dirangkum dalam kerangka berpikir pada Gambar 2.3 berikut,



Gambar 2.3 Alur Penelitian

BAB III

METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian yang dilaksanakan peneliti termasuk penelitian deskriptif kuantitatif. Penelitian deskriptif kuantitatif merupakan penelitian yang bertujuan untuk menggambarkan dan menginterpretasikan objek penelitian dalam hal ini adalah pola lulusan mahasiswa secara sistematis dan factual berdasarkan informasi yang didapat di lapangan dan pengembangan ilmu pengetahuan. Selain itu, penelitian ini juga mengaplikasikan analisis kuantitatif dari data – data pendukung penelitian yang dikumpulkan, diolah dan dianalisis untuk keperluan prediksi hubungan antar variable yang diteliti. Perhitungan yang digunakan untuk memprediksi pola lulusan mahasiswa dengan teknik data mining algoritma C5.0 dan Naïve Bayes berbantuan *Microsoft Excel* dan *Software RapidMiner*.

3.2 Tempat dan Waktu Penelitian

Penelitian dilakukan di Program Studi Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang. Penelitian ini dilaksanakan pada bulan Agustus sampai November 2021.

3.3 Objek Penelitian

Objek penelitian ini adalah prediksi estimasi pola lulusan mahasiswa Program Studi Matematika di Fakultas Sains dan Teknologi UIN Walisongo Semarang menggunakan algoritma C5.0 dan Naïve Bayes. Dari algoritma C5.0 dan Naïve Bayes akan diperoleh hasil dari proses evaluasi dengan menggunakan *Cross Validation, Confusion Matrix,* dan *ROC Curve* untuk mengetahui algoritma klasifikasi

data mining yang paling akurat dalam prediksi pola lulusan mahasiswa Program Studi Matematika.

3.4 Jenis dan Sumber Penelitian

Penelitian ini menggunakan data sekunder yang berasal dari data mahasiswa Program Studi Matematika Angkatan Tahun 2015-2016. Data ini diambil dari laman SIJAMU Walisongo https://sijamu.walisongo.ac.id/ berupa data induk akademik dan data biodata mahasiswa.

3.5 Populasi dan Sampel Penelitian

Pada penelitian ini, yang menjadi populasi adalah mahasiswa Jurusan Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang. Sampel yang digunakan dalam penelitian ini adalah data mahasiswa Program Studi Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang dari Angkatan Tahun 2015 – 2016 berjumlah 49 mahasiswa yang terdiri atas 36 mahasiswa yang telah lulus tepat waktu dan 13 mahasiswa yang tidak lulus tepat waktu. Jumlah lulusan Program Studi Matematika terhitung sebanyak 49 mahasiswa, sehingga seluruh data lulusan digunakan sebagai sampel dalam penelitian ini.

3.6 Metode Pengambilan Sampel Penelitian

Metode yang digunakan untuk pengambilan sampel adalah *purposive sampling*. Metode ini merupakan salah satu metode pengambilan sampel pada populasi yang dilakukan sesuai dengan persyaratan sampel yang dibutuhkan. Pengambilan sampel dilakukan dengan cara mengambil sampel tertentu yang memiliki karakteristik, kriteria atau sifat tertentu. Penelitian ini menentukan karakteristik atau kriteria yang dijadikan bahan pertimbangan dalam pengambilan sampel adalah mahasiswa Program Studi

Matematika Angkatan tahun 2015 dan 2016 yang telah dinyatakan lulus.

3.7 Metode Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini adalah studi literatur dan dokumentasi. Pengumpulan data diawali dengan mencari informasi mengenai masa studi mahasiswa Program Studi Matematika dan atribut-atribut yang mempengaruhi masa studi mahasiswa melalui berbagai sumber seperti buku, jurnal maupun artikel yang ada di internet.

Selanjutnya, dari informasi yang telah dikumpulkan kemudian peneliti mendokumentasikan data-data sekunder yang digunakan dalam penelitian. Data-data tersebut dikumpulkan dari sumber terpercaya yaitu Sistem Akademik Walisongo (WaliSiadik) dan Sistem Penjaminan Mutu (Sijamu) Program Studi Matematika. Data yang diperoleh dari Sistem Akademik Walisongo adalah transkrip nilai mahasiswa yang lulus tepat waktu dan tidak lulus tepat waktu. Sedangkan data yang diperoleh dari Sistem Penjaminan Mutu adalah data terkait identitas mahasiswa diantaranya jenis kelamin dan jalur masuk yang diikuti saat mendaftar. Selanjutnya data tersebut dipelajari dan dianalisis sesuai dengan objek penelitian.

3.8 Proses Data Mining

Data Mining merupakan bidang ilmu yang bertujuan untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang dimiliki. Proses *Knowledge Discovery in Database* (KDD) digunakan untuk menjelaskan penggalian informasi yang tersembunyi dalam suatu basis data yang besar. Langkah awal sebelum menggunakan algoritma data mining dilakukan *preprocessing* data dengan tujuan untuk mempermudah dalam memahami data dan meningkatkan kualitas data. *Preprocessing*

data dapat dilakukan dalam beberapa tahapan diantaranya sebagai berikut :

3.8.1 Pembersihan Data

Tahapan awal proses KDD dilakukan pembersihan data terhadap *noise* yang ditemukan berupa *missing value, inkonsisten* data dan *redundant* data. Seluruh atribut atau variable akan diseleksi untuk mendapatkan atribut atau variabel yang relevan, tidak *missing value* dan *redundant*. Untuk atribut atau variable yang bernilai kosong maka akan dihapus atau dihilangkan.

Data yang digunakan dalam penelitian ini adalah data yang diambil dari dua sumber data yaitu data biodata mahasiswa dan data akademik mahasiswa. Jumlah mahasiswa aktif di Program Studi Matematika sebanyak 270 mahasiswa, dengan rincian 258 mahasiswa aktif dan 12 mahasiswa berstatus cuti atau non aktif. Data yang digunakan adalah data mahasiswa Program Studi Matematika yang telah lulus sebanyak 49 lulusan dari Angkatan Tahun 2015 dan 2016.

3.8.2 Seleksi Data

Penyeleksian data adalah tahapan dipilihnya atribut-atribut yang dibutuhkan dan menghilangkan atribut yang tidak dibutuhkan. Pada Tabel 3.1 merupakan rincian data awal yang diperoleh untuk penelitian sebagai berikut,

Tabel 3.1 Tabel Biodata Mahasiswa Program Studi Matematika

NO	NIM	NAMA	TEMPAT LAHIR	TANGGAL LAHIR	ALAMAT	JENIS KELAMIN	JALUR MASUK
1	15080 46001	Anggit Nurfikria ni	Wonosobo	30/08/1996	Marong Rt.15 Rw.04, Grugu, Kaliwiro, Wonosobo Kaliwiro	Perempuan	SBMPTN
2	15080 46002	Mustaqii m Bariklana	Seputih Mataram,La mpung Tengah	08/04/1997	Perum I Pt. Gpm Blok F.061	Laki-laki	SBMPTN
3	15080 46003	Bambang Puntodew o	Semarang	31/03/1997	Jln. Tugurejo X T.36 No. 81 Tugu	Laki-laki	SBMPTN
4	15080 46004	Uswatun Khasanah	Semarang	28/06/1996	Desa Ciporos Rt 02 Rw 12 Karangpucung	Perempuan	SBMPTN
5	15080 46007	Yunus Nur Rahmawa	Rembang	09/11/1996	Ds.Kemadu Rt1 Rw3, Kec. Sulang, Kab. Rembang	Laki-laki	SBMPTN
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
268	21080 46100	Erlinda Diah Prawati	Rembang	26/01/2002	Ds.Pelemsari Rt 002 Rw 004 Desa/Kel. Ds.Pelemsari Kode Pos 59253 Kec. Sumber	Perempuan	Mandiri
269	21080 46099	Siti Rianita Oktamida	Purworejo	05/10/2002	Desa Seborokrap, Kec. Banyuurip, Kab. Purworejo	Perempuan	Mandiri
270	21080 46098	Achla Fauziyah	Kudus	05/06/2003	Dukuh Kancilan Rt 004 Rw 004 Desa/Kel. Terban Kode Pos 59382 Kec. Jekulo	Perempuan	Mandiri

Dari seluruh informasi pada Tabel biodata mahasiswa terdapat atribut yang digunakan dalam proses analisis data mining yaitu jenis kelamin dan jalur masuk yang diikuti oleh mahasiswa ketika mendaftar ke UIN Walisongo Semarang. Atribut yang terpilih telah mewakili informasi yang dibutuhkan sebagai indikator penelitian.

Tabel 3.2 Tabel Data Akademik Mahasiswa Program Studi Matematika

NO	NIM	NAMA	SEMESTER 1							IP SEMESTER 1	SEMESTER 2							IP SEMESTER 2	SEMESTER 8		IP SEMESTER 8	IPK	TANGGAL KELULUSAN	JUMLAH SKS SELURUH
			PENDIDIKAN PANCASILA	BAHASA INGGRIS	BAHASA ARAB I	LOGIKA PASTI DAN MATEMATIKA	KALKULUS I	METODE STATISTIKA	ALGORITMA DAN PEMROGRAMAN	PENGANTAR AKUNTANSI		PENDIDIKAN KEWARGANEGARAAN	ILMU ALAMIAH	SEJARAH PERADABAN ISLAM	BAHASA INGGRIS	BAHASA INDONESIA	KALKULUS II	ALJABAR LINIER ELEMENTER	LOGIKA ANALITIK	KEWIRAUSAHAAN		KKN	SEMINAR				
1	1500046028	AFRIKHATUL HIMAH	3,4	4,0	4,0	4,0	2,2	2,3	3,2	3,7	3,35	4,0	3,4	3,5	3,6	4,0	3,1	3,2	3,0	3,4	3,5	4,0	3,7	3,9	3,1	20/09/2021	150
2	1500046013	AHMAD SAIFUDIN	3,6	4,0	4,0	3,8	2,1	3,8	2,9	2,8	3,38	3,6	3,5	2,5	3,6	4,0	3,9	3,4	2,3	2,8	3,3	4,0	3,7	3,9	3,0	20/09/2021	149
3	1500046001	ANGGITI MUFIKHANI	3,3	4,0	2,9	3,9	2,8	2,5	3,3	4,0	3,3	3,8	3,1	3,0	4,0	4,0	2,8	2,7	2,8	3,4	3,3	4,0	3,7	3,9	3,1	20/09/2021	149
4	1500046003	PUNTODEVO	3,8	3,8	2,9	3,2	2,0	3,2	3,3	3,0	3,2	3,5	3,1	2,5	4,0	4,0	2,0	2,0	2,4	2,0	2,8	4,0	3,7	3,9	2,9	20/09/2021	149
5	1500046008	DEVI SYAFITRI	4,0	3,9	3,6	2,5	3,0	2,3	3,4	4,0	3,3	3,8	3,2	4,0	3,7	3,9	2,9	3,1	3,5	3,5	3,5	4,0	3,8	3,9	3,0	20/09/2021	149
6	1500046027	IVYANI MUZLULLHUDA	3,8	3,9	4,0	3,0	3,9	2,0	3,2	3,4	3,4	3,6	3,3	3,3	2,2	3,5	2,7	2,6	3,0	3,6	3,1	4,0	3,9	4,0	3,0	20/09/2021	149
7	1500046023	M. MARZUQUR ROHIM	3,5	4,0	4,0	3,4	2,3	4,0	2,9	3,3	3,4	3,6	3,3	3,3	2,5	3,8	3,6	2,3	2,6	2,4	3,0	4,0	3,8	3,9	3,0	20/09/2021	149
8	1500046022	MUHAMMAD ABDURRAHMAN ROIS	4,0	4,0	2,3	4,0	3,5	4,0	3,3	4,0	3,6	4,0	3,4	4,0	4,0	4,0	3,1	3,5	3,7	3,0	3,6	4,0	4,0	4,0	3,4	20/09/2021	150
9	1500046010	MUHAMMAD GILANG FATHURRAHMAN	4,0	4,0	3,8	4,0	4,0	4,0	3,3	4,0	3,9	3,7	3,1	3,5	3,5	4,0	3,7	3,8	3,7	3,3	3,6	4,0	3,7	3,9	3,4	20/09/2021	150
10	1500046002	MUSTAQIM BARWILANA	4,0	3,8	4,0	3,3	2,2	4,0	3,2	2,9	3,4	3,8	3,2	3,2	3,9	4,0	4,0	2,3	3,6	2,8	3,4	4,0	3,8	3,9	3,1	20/09/2021	149
11	1500046024	NAYYAROTUL JANNAH	4,0	4,0	4,0	3,7	2,0	3,2	3,3	4,0	3,5	3,6	3,7	3,5	3,8	4,0	4,0	3,8	2,5	2,9	3,5	4,0	3,7	3,9	3,3	20/09/2021	149
12	1500046009	RAYA	4,0	4,0	3,0	3,7	4,0	4,0	3,3	4,0	3,8	3,8	3,2	3,0	4,0	4,0	4,0	3,7	3,2	3,3	3,6	4,0	3,8	3,9	3,4	20/09/2021	150
13	1500046021	NUR FADAH	3,7	4,0	4,0	4,0	4,0	3,7	3,3	3,8	3,8	3,8	3,4	4,0	4,0	3,9	3,4	2,8	3,3	2,8	3,5	4,0	3,8	3,9	3,2	20/09/2021	149
14	1500046017	SANTOSO	4,0	3,9	4,0	4,0	3,0	3,7	3,2	2,6	3,6	3,9	3,3	3,2	3,9	3,9	2,2	2,0	2,7	3,2	3,1	4,0	3,7	3,9	3,1	20/09/2021	149
15	1500046004	USVATUN KHASANAH	4,0	3,9	3,6	4,0	3,4	4,0	3,3	3,9	3,8	3,8	3,1	3,5	4,0	4,0	3,9	3,6	4,0	3,8	3,7	4,0	3,9	4,0	3,4	20/09/2021	150
16	1500046015	WA MALLIDA	4,0	4,0	4,0	4,0	3,3	3,7	3,3	4,0	3,8	4,0	3,4	3,2	4,0	3,9	4,0	3,7	4,0	3,0	3,7	4,0	3,8	3,9	3,4	20/09/2021	150
17	1500046007	YUMUS NUR RAHMAVA	3,4	3,9	3,2	3,4	4,0	2,7	3,0	2,3	3,3	3,8	3,2	3,5	3,0	3,6	4,0	2,5	2,3	3,1	3,3	4,0	3,6	3,8	3,1	20/09/2021	149
18	1500046025	ZAKARIA BANI KHIYAR	3,3	4,0	4,0	4,0	4,0	4,0	3,6	4,0	3,3	4,0	3,6	4,0	4,0	4,0	4,0	4,0	4,0	3,8	3,9	4,0	3,8	3,9	3,4	20/09/2021	150
19	1500046020	ZHAFRANI MAHFUDH	4,0	4,0	4,0	4,0	4,0	3,2	3,4	4,0	3,8	3,8	3,2	4,0	3,6	4,0	3,9	3,3	3,1	2,9	3,5	4,0	3,8	3,9	3,2	20/09/2021	150
20	1500046037	ANDI WIDODO	3,5	3,5	4,0	4,0	3,2	4,0	4,0	3,5	3,8	3,5	3,5	3,3	3,9	4,0	3,3	3,4	3,5	3,0	3,6	4,0	3,9	4,0	3,3	20/09/2021	150
21	1500046004	ELOK YULIA FAKOH	3,9	3,7	4,0	3,4	3,0	3,7	4,0	4,0	3,7	3,8	3,8	3,3	3,7	4,0	3,1	3,0	3,1	4,0	3,5	4,0	3,8	3,9	3,2	20/09/2021	154
22	1500046012	FITRI NUR ANISA	3,4	3,7	4,0	4,0	3,0	2,7	4,0	3,0	3,5	3,8	3,5	3,5	4,0	3,5	2,8	2,3	3,0	2,2	3,2	4,0	3,9	4,0	3,1	20/09/2021	149
23	1500046030	IQBAL KHABIBI	2,4	2,5	3,8	3,7	2,2	2,1	3,8	3,4	3,0	3,9	3,8	3,7	3,8	3,8	2,7	3,0	2,6	3,6	3,4	4,0	3,7	3,9	2,9	20/09/2021	150
24	1500046031	MUKHOFIUL MIFTAH	3,7	3,5	4,0	4,0	4,0	4,0	4,0	3,1	3,8	3,8	3,5	3,8	3,9	3,8	4,0	4,0	4,0	4,0	3,9	4,0	3,9	4,0	3,4	20/09/2021	150
25	1500046032	MULTIMINGSIH	3,2	4,0	4,0	4,0	3,6	3,9	4,0	4,0	3,8	3,8	3,6	3,5	3,6	3,6	4,0	3,1	2,5	3,8	3,5	4,0	3,9	4,0	3,4	20/09/2021	154
26	1500046010	NAILASH SHOFA	3,8	4,0	4,0	4,0	4,0	4,0	3,4	3,6	3,9	3,7	3,5	3,9	2,7	3,3	4,0	2,9	4,0	2,1	3,3	4,0	3,8	3,9	3,1	20/09/2021	152
27	1500046016	KHASANAH	3,8	4,0	4,0	4,0	3,6	3,1	4,0	4,0	3,8	3,6	3,7	3,5	4,0	3,7	4,0	3,5	3,4	3,7	3,7	4,0	3,7	3,9	3,3	20/09/2021	152
28	1500046020	NOVI SAADAH	3,6	3,7	3,4	4,0	3,0	2,3	4,0	4,0	3,5	3,6	3,8	3,3	3,6	3,8	2,7	3,4	3,3	4,0	3,5	3,9	3,7	3,8	3,3	20/09/2021	154
29	1500046029	RISNA ARDIANY	3,7	4,0	3,4	4,0	3,6	3,7	4,0	4,0	3,8	3,6	4,0	3,8	4,0	4,0	3,7	4,0	4,0	3,5	3,8	4,0	4,0	4,0	3,3	20/09/2021	152

Sumber data kedua adalah data akademik mahasiswa yang meliputi nilai IP tiap semester dari semester 1 hingga semester 8, jumlah SKS yang diambil tiap semester, nilai IPK, tanggal lulus dan total SKS yang diambil.

3.8.3 Transformasi

Tahap transformasi terdiri atas proses *encoding* atau konversi dimana dalam tahap ini dilakukan perubahan data ke dalam format data yang siap untuk diaplikasikan. Data yang diperoleh selanjutnya diidentifikasi mengenai jenis atribut dan rentang nilainya untuk mempermudah dalam proses mining selanjutnya. Proses identifikasi atau penentuan atribut yang akan digunakan dalam penelitian ini adalah sebagai berikut,

1) Jenis kelamin
   Atribut jenis kelamin hanya terdiri dari dua kemungkinan yaitu laki-laki dan perempuan.

Tabel 3.3 Kategori Atribut Jenis Kelamin

NO	Jenis Kelamin	Kategori
1	Laki-Laki	1
2	Perempuan	2

2) Jalur Masuk
   Atribut jalur masuk merupakan variable yang menunjukkan pilihan jalur masuk yang diikuti calon mahasiswa di UIN Walisongo.

Tabel 3.4 Kategori Atribut Jalur Masuk

NO	Jalur Masuk	Kategori
1	SBMPTN	1
2	SNMPTN	2
3	MANDIRI	3

3) Indeks Prestasi Tiap Semester (IPS)
   Atribut indeks prestasi tiap semester (IPS) merupakan nilai Indeks Prestasi (IP) tiap semester yang telah ditempuh oleh mahasiswa. Tipe data pada atribut Indeks Prestasi tiap Semester (IPS) adalah numerik.
4) Indeks Prestasi Kumulatif (IPK)
   Atribut indeks prestasi kumulatif (IPK) merupakan nilai rata-rata Indeks Prestasi (IP) tiap semester yang telah ditempuh oleh mahasiswa. Tipe data pada atribut IP tiap semester adalah numerik.
5) Jumlah SKS Yang Diambil Tiap Semester
   Atribut jumlah SKS menunjukkan jumlah SKS yang diambil oleh mahasiswa tiap semester.

Tabel 3.5 Kategori Atribut Jumlah SKS yang diambil

NO	**Jumlah SKS yang diambil tiap semester**	**Kategori**
1	$\leq$ 20 SKS	1
2	$>$ 20 SKS	2

6) Jumlah Total SKS yang diambil
   Atribut jumlah SKS menunjukkan jumlah total SKS yang diambil oleh mahasiswa dari semester 1 hingga semester akhir.

Tabel 3.6 Kategori Atribut Jumlah Total SKS yang diambil

NO	Jumlah SKS yang diambil tiap semester	Kategori
1	≤ 148 SKS	1
2	> 148 SKS	2

7) Keputusan

Atribut keputusan merupakan data yang berfungsi untuk menentukan hasil keputusan. Dalam data keputusan hanya memiliki 2 nilai yaitu sesuai (lulus tepat waktu) dan tidak sesuai (tidak lulus tepat waktu).

Tabel 3.7 Kategori Atribut Keputusan

NO	Keputusan	Kategori
1	Tidak Lulus Tepat Waktu	1
2	Lulus Tepat waktu	2

Tabel 3.1 dan **Tabel 3.2** merupakan dataset mahasiswa yang belum melalui tahap transformasi. Dataset yang tersaji diidentifikasi jenis data selanjutnya dilakukan proses *encoding* terlebih dahulu kedalam bentuk nominal. Berikut adalah dataset mahasiswa yang telah melalui proses *encoding* dan siap untuk menjadi perangkat pemodelan,

Tabel 3.8 Dataset Mahasiswa setelah proses *encoding*

NO	ID	JENIS KELAMIN	JALUR MASUK	IPS1	SKS1	IPS2	SKS2	1PS3	SKS3	IPS4	SKS4	…	IPK	SKS TOTAL	KETERANGAN
1	1508046001	PEREMPUAN	SBMPTN	3.34	1	3.29	2	3.56	2	3.6	2	…	3.4	148	TEPAT WAKTU
2	1508046002	LAKI-LAKI	SBMPTN	3.43	1	3.42	2	3.09	2	3.73	2	…	3.36	148	TEPAT WAKTU
3	1508046003	LAKI-LAKI	SBMPTN	3.15	1	2.83	2	2.91	2	2.97	2	…	3.02	148	TEPAT WAKTU
4	1508046004	PEREMPUAN	SBMPTN	3.76	1	3.74	2	3.98	2	3.8	2	…	3.82	150	TEPAT WAKTU
5	1508046005	LAKI-LAKI	SBMPTN	3.83	1	3.56	2	3.92	2	3.67	2	…	3.68	144	TIDAK TEPAT WAKTU
6	1508046007	LAKI-LAKI	SBMPTN	3.31	1	3.29	2	3.41	2	3.7	2	…	3.33	148	TEPAT WAKTU
7	1508046008	PEREMPUAN	SBMPTN	3.34	1	3.51	2	3.36	2	3.43	2	…	3.41	148	TEPAT WAKTU
8	1508046009	PEREMPUAN	SBMPTN	3.75	1	3.58	2	3.88	2	3.81	2	…	3.73	150	TEPAT WAKTU
9	1508046010	LAKI-LAKI	SBMPTN	3.89	1	3.59	2	3.89	2	3.73	2	…	3.71	150	TEPAT WAKTU
10	1508046012	PEREMPUAN	SBMPTN	2.99	1	3.6	2	3.16	2	3.3	2	…	3.18	144	TIDAK TEPAT WAKTU
11	1508046013	LAKI-LAKI	SBMPTN	3.38	1	3.33	2	3.51	2	3.49	2	…	3.26	148	TEPAT WAKTU
12	1508046015	PEREMPUAN	SBMPTN	3.79	1	3.69	2	3.87	2	3.79	2	…	3.79	150	TEPAT WAKTU
13	1508046016	PEREMPUAN	SBMPTN	3.16	1	3.36	2	3.13	2	3.46	2	…	3.29	142	TIDAK TEPAT WAKTU
14	1508046017	LAKI-LAKI	SBMPTN	3.55	1	3.14	2	3.02	2	3.56	2	…	3.3	148	TIDAK TEPAT WAKTU
15	1508046018	LAKI-LAKI	SBMPTN	3.75	1	3.56	2	3.77	2	3.21	2	…	3.36	142	TIDAK TEPAT WAKTU
16	1508046020	LAKI-LAKI	SBMPTN	3.83	1	3.53	2	3.81	2	3.77	2	…	3.69	150	TEPAT WAKTU
17	1508046021	PEREMPUAN	MANDIRI	3.81	1	3.49	2	3.7	2	3.64	2	…	3.66	148	TEPAT WAKTU
18	1508046022	LAKI-LAKI	MANDIRI	3.64	1	3.63	2	3.9	2	4.00	2	…	3.84	150	TEPAT WAKTU
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
45	1608046027	PEREMPUAN	MANDIRI	3.76	1	3.85	2	3.74	2	3.65	2	…	3.75	154	TEPAT WAKTU
47	1608046028	PEREMPUAN	MANDIRI	3.3	1	3.31	2	3.4	2	3.23	2	…	3.09	142	TIDAK TEPAT WAKTU
47	1608046029	PEREMPUAN	MANDIRI	3.8	1	3.84	2	3.64	2	3.63	2	…	3.72	152	TEPAT WAKTU
48	1608046030	LAKI-LAKI	MANDIRI	2.98	1	3.43	2	3.46	2	3.2	2	…	3.3	150	TEPAT WAKTU
49	1608046031	LAKI-LAKI	MANDIRI	3.22	1	2.96	2	3.52	2	2.98	2	…	2.85	142	TIDAK TEPAT WAKTU

3.8.4 Data Mining

Proses penambangan atau *mining* adalah sutu proses utama yang mengaplikasikan metode, teknik dan algoritma tertentu untuk mengekstrasi pola yang diinginkan. Pada penelitian ini dipilih algoritma C5.0 dan algoritma Naïve Bayes sebagai metode yang digunakan untuk memprediksi pola data lulusan mahasiswa Program Studi Matematika.

1) Algoritma C5.0

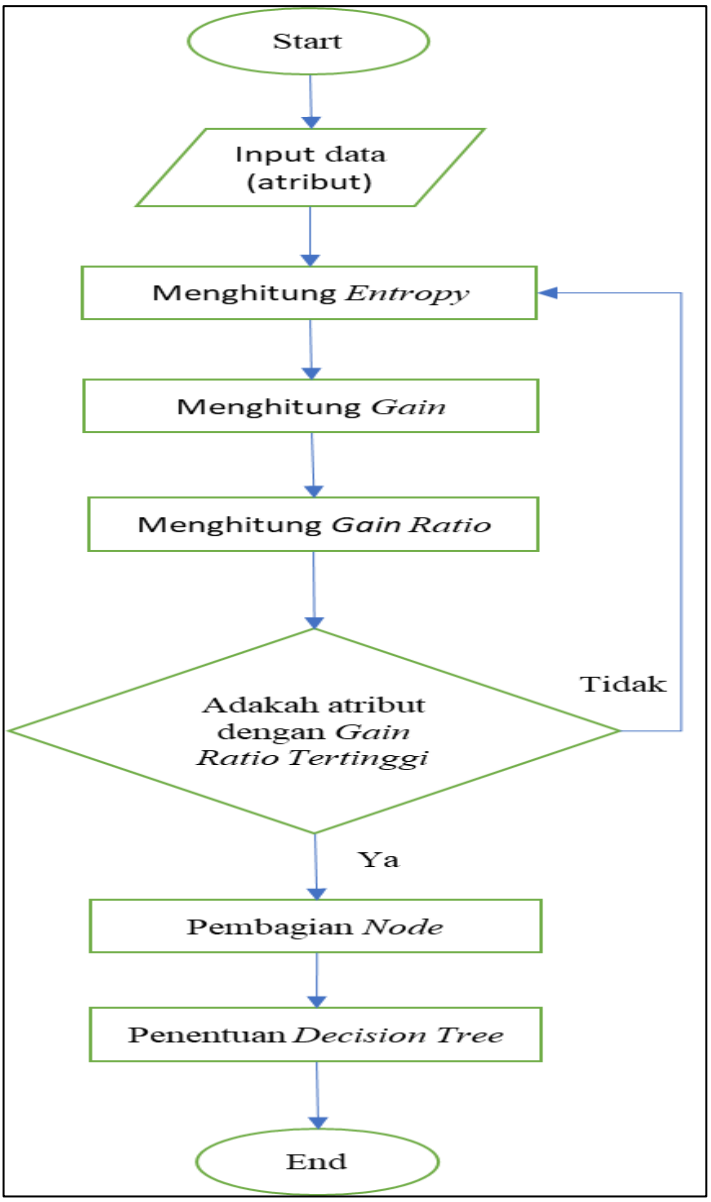
Dalam analisis algoritma C5.0 terdapat beberapa tahapan diantaranya,

a. Mempersiapkan data *training*. Data *training* seringkali diambil dari data histori yang pernah terjadi sebelumnya dan telah dikelompokkan ke dalam kelas-kelas tertentu.
b. Menentukan akar (*root*) dari pohon. Akar diambil dari atribut yang terpilih dengan cara menghitung nilai *gain* dari masing-masing

atribut. Nilai *gain* yang paling tinggi yang akan menjadi akar pertama.

c. Menghitung nilai *gain* menggunakan formula (2.2).
d. Menentukan cabang masing-masing atribut dengan cara yang sama seperti atribut akar pada Langkah (2).
e. Mengulangi Langkah ke-2. Kelas dibagi ke dalam cabang dan jika terdapat cabang yang memiliki dua kelas maka yang terpilih adalah kelas yang terbanyak dan proses diulang untuk masing-masing cabang hingga semua kelas pada cabang memiliki kelasnya masing-masing.

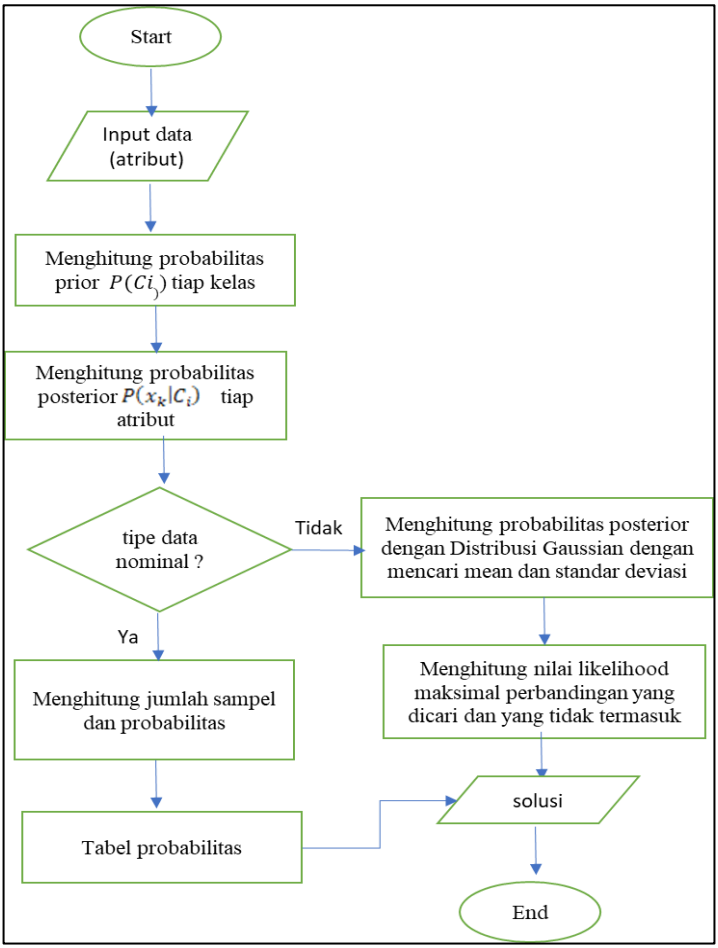
Tahapan analisis algoritma C5.0 lebih jelasnya dapat dilihat pada Gambar 3.1



Gambar 3.1 Alur Algoritma C5.0

2) Algoritma Naïve Bayes

Flowchart (bagan alir) adalah teknik analisis yang digunakan untuk menjelaskan bagian-bagian dari suatu system secara jelas, akurat dan logis. Tahapan analisis dengan Algoritma Naïve Bayes tertuang dalam gambar 3.2 berikut,



Gambar 3.2 Alur Algoritma Naïve Bayes

3.8.5 Evaluasi Pola dan Interpretasi

Pada tahap ini dilakukan pemeriksaan pola atau informasi yang ditemukan telah sesuai dengan fakta atau hipotesa. Jika didapatkan bahwa hasil tidak sesuai dengan hipotesa maka perlu dilakukan Langkah-langkah alternatif untuk memperbaiki pola data.

3.9 Pengujian dan Akurasi Data

Pengujian akurasi dan ketepatan hasil klasifikasi pada penelitian ini dilakukan dengan memanfaatkan data lulusan Angkatan 2015 - 2016 dari Program Studi Matematika. Data lulusan tersebut dibagi menjadi dua partisi dengan 70% data sebagai data training dan 30% data menjadi data testing. Data lulusan tersebut merupakan data *testing* yang diujicobakan kemudian dibandingkan terhadap data yang sebenarnya. Perbedaan antara data sebenarnya dan data *testing* yang telah diujicoba, akan dihitung besaran nilai *error* dan nilai akurasinya. Rasio 70 : 30 antara data training dan testing merupakan model prediksi yang cukup popular dalam klasifikasi data mining dan efektif nilainya.

Metode pengujian algoritma yang digunakan dalam penelitian ini adalah *Confussion Matrix* dan nilai kurva ROC. *Confussion Matrix* adalah suatu matriks berordo 2x2 yang digunakan sebagai alat ukur untuk memperoleh jumlah ketepatan klasifikasi dataset terhadap kelas lulus tepat waktu dan tidak lulus tepat waktu.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Persiapan dan Pengujian Data

Dalam data mining atau penambangan data adalah teknik yang relative cepat dan mudah untuk menemukan pengetahuan, pola dan atau relasi antar data secara otomatis. Data mining dapat memberikan dampak negative maupun positif bergantung pada penggunaannya. Himpunan data dibangun dari objek-objek data, dimana objek data menyatakan sebuat entitas. Dalam penelitian ini, data yang digunakan adalah kumpulan data yang dibagi menjadi data *training* dan data *testing*. Data *training* merupakan data yang digunakan untuk membentuk model klasifikasi. Model klasifikasi ini adalah representasi *knowledge* yang akan digunakan untuk memprediksi kelas data baru yang belum pernah ada. Data *testing* digunakan untuk tahapan pengujian dari *knowledge* yang diperoleh dari data *training*.

Jumlah data keseluruhan yang digunakan dalam proses mining adalah 49 sampel data induk mahasiswa Program Studi Matematika Angkatan tahun 2015-2016 yang telah dinyatakan lulus dengan klasifikasi 36 mahasiswa lulus tepat waktu dan 13 mahasiswa tidak lulus tepat waktu. Objek data yang digunakan dalam penelitian ini merupakan objek data yang disimpan dalam suatu basis data yang disebut *tuple* dimana baris menyatakan objek-objek data dan kolom adalah atribut. Atribut merupakan symbol yang menyatakan identitas atau karakteristik suatu kelompok atau *class*. Atribut Masa atau lama studi merupakan atribut yang menggambarkan waktu yang dibutuhkan oleh mahasiswa untuk menyelesaikan studinya di pendidikan tinggi. Beberapa faktor dapat

mempengaruhi lama studi mahasiswa antara lain, faktor internal dan faktor eksternal dari mahasiswa, misalnya faktor keterpaksaan dalam proses kuliah, salah dalam pemilihan jurusan, bahkan juga terlalu aktif dalam organisasi kemahasiswaan. Masa studi mahasiswa UIN Walisongo Semarang untuk program S-1 dapat ditempuh selama 8 semester atau 4 tahun dan paling lama ditempuh selama 14 semester atau 7 tahun. Beban studi mahasiswa program pendidikan S-1 yang harus diambil adalah minimal 144 satuan kredit semester (SKS) dan maksimal 148 satuan kredit semester (SKS). Dalam penelitian ini, atribut yang digunakan pada antara lain,

Tabel 4.1 Tabel Atribut yang Digunakan

No	Atribut yang digunakan
1	Jenis Kelamin
2	Jalur masuk yang diikuti saat seleksi masuk perguruan tinggi
3	Indeks Prestasi Semester 1 hingga semester 8
4	Indeks Prestasi kumulatif (IPK)
5	Jumlah SKS yang diambil tiap semester
6	Jumlah SKS total
7	Keterangan lulus

Data *training* yang digunakan dalam penelitian ini tersaji dalam tabel berikut,

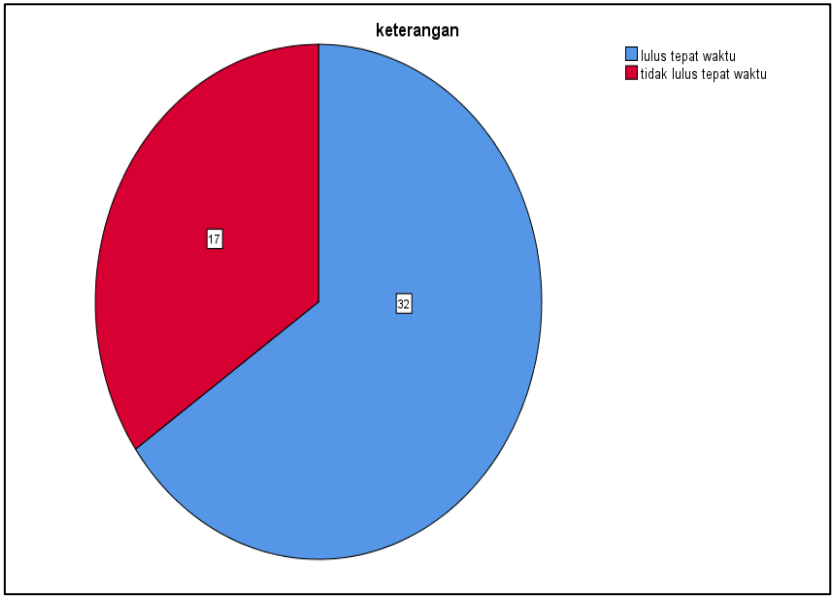
Tabel 4.2 Data *Training*

NO	ID	JENIS KELAMIN	JALUR MASUK	IPS1	SKS1	IPS2	SKS2	IPS3	SKS3	IPS4	SKS4	...	IPS8	SKS8	SKSTOTAL	IPK	KETERANGAN
1	1D01501	PEREMPUAN	SBMPTN	3,34	<= 20 SKS	3,29	> 20 SKS	3,56	> 20 SKS	3,60	> 20 SKS	...	3,85	<= 20 SKS	<=148	3,40	TEPAT WAKTU
2	1D01502	LAKI-LAKI	SBMPTN	3,43	<= 20 SKS	3,42	> 20 SKS	3,09	> 20 SKS	3,73	> 20 SKS	...	3,90	<= 20 SKS	<=148	3,36	TEPAT WAKTU
3	1D01503	LAKI-LAKI	SBMPTN	3,15	<= 20 SKS	2,83	> 20 SKS	2,91	> 20 SKS	2,97	> 20 SKS	...	3,85	<= 20 SKS	<=148	3,02	TEPAT WAKTU
4	1D01504	PEREMPUAN	SBMPTN	3,76	<= 20 SKS	3,74	> 20 SKS	3,98	> 20 SKS	3,80	> 20 SKS	...	3,95	<= 20 SKS	>148	3,82	TEPAT WAKTU
5	1D01505	LAKI-LAKI	SBMPTN	3,83	<= 20 SKS	3,56	> 20 SKS	3,92	> 20 SKS	3,67	> 20 SKS	...	0,00	<= 20 SKS	<=148	3,68	TIDAK TEPAT WAKTU
6	1D01506	LAKI-LAKI	SBMPTN	3,31	<= 20 SKS	3,29	> 20 SKS	3,41	> 20 SKS	3,70	> 20 SKS	...	3,80	<= 20 SKS	<=148	3,33	TEPAT WAKTU
7	1D01507	PEREMPUAN	SBMPTN	3,34	<= 20 SKS	3,51	> 20 SKS	3,36	> 20 SKS	3,43	> 20 SKS	...	3,90	<= 20 SKS	<=148	3,41	TEPAT WAKTU
8	1D01508	PEREMPUAN	SBMPTN	3,75	<= 20 SKS	3,58	> 20 SKS	3,88	> 20 SKS	3,81	> 20 SKS	...	3,90	<= 20 SKS	>148	3,73	TEPAT WAKTU
9	1D01509	LAKI-LAKI	SBMPTN	3,89	<= 20 SKS	3,59	> 20 SKS	3,89	> 20 SKS	3,73	> 20 SKS	...	3,85	<= 20 SKS	>148	3,71	TEPAT WAKTU
10	1D01510	PEREMPUAN	SBMPTN	2,99	<= 20 SKS	3,60	> 20 SKS	3,16	> 20 SKS	3,30	> 20 SKS	...	0,00	<= 20 SKS	<=148	3,18	TIDAK TEPAT WAKTU
11	1D01511	LAKI-LAKI	SBMPTN	3,38	<= 20 SKS	3,33	> 20 SKS	3,51	> 20 SKS	3,49	> 20 SKS	...	3,85	<= 20 SKS	<=148	3,26	TEPAT WAKTU
12	1D01512	PEREMPUAN	SBMPTN	3,79	<= 20 SKS	3,69	> 20 SKS	3,87	> 20 SKS	3,79	> 20 SKS	...	3,90	<= 20 SKS	>148	3,79	TEPAT WAKTU
13	1D01513	PEREMPUAN	SBMPTN	3,16	<= 20 SKS	3,36	> 20 SKS	3,13	> 20 SKS	3,46	> 20 SKS	...	2,00	<= 20 SKS	<=148	3,29	TIDAK TEPAT WAKTU
14	1D01514	LAKI-LAKI	SBMPTN	3,55	<= 20 SKS	3,14	> 20 SKS	3,02	> 20 SKS	3,56	> 20 SKS	...	3,85	<= 20 SKS	<=148	3,30	TIDAK TEPAT WAKTU
15	1D01515	LAKI-LAKI	SBMPTN	3,75	<= 20 SKS	3,56	> 20 SKS	3,77	> 20 SKS	3,21	> 20 SKS	...	2,00	<= 20 SKS	<=148	3,36	TIDAK TEPAT WAKTU
16	1D01516	LAKI-LAKI	SBMPTN	3,83	<= 20 SKS	3,53	> 20 SKS	3,81	> 20 SKS	3,77	> 20 SKS	...	3,90	<= 20 SKS	>148	3,69	TEPAT WAKTU
17	1D01517	PEREMPUAN	MANDIRI	3,81	<= 20 SKS	3,49	> 20 SKS	3,70	> 20 SKS	3,64	> 20 SKS	...	3,90	<= 20 SKS	<=148	3,66	TEPAT WAKTU
18	1D01518	LAKI-LAKI	MANDIRI	3,64	<= 20 SKS	3,63	> 20 SKS	3,90	> 20 SKS	4,00	> 20 SKS	...	4,00	<= 20 SKS	>148	3,84	TEPAT WAKTU
19	1D01519	LAKI-LAKI	MANDIRI	3,43	<= 20 SKS	3,04	> 20 SKS	3,39	> 20 SKS	3,34	> 20 SKS	...	3,90	<= 20 SKS	<=148	3,25	TEPAT WAKTU
20	1D01520	PEREMPUAN	MANDIRI	3,53	<= 20 SKS	3,53	> 20 SKS	3,54	> 20 SKS	3,70	> 20 SKS	...	3,85	<= 20 SKS	<=148	3,55	TEPAT WAKTU
21	1D01521	LAKI-LAKI	MANDIRI	3,94	<= 20 SKS	3,93	> 20 SKS	3,92	> 20 SKS	3,77	> 20 SKS	...	3,90	<= 20 SKS	>148	3,93	TEPAT WAKTU
22	1D01522	LAKI-LAKI	SBMPTN	3,59	<= 20 SKS	3,60	> 20 SKS	3,74	> 20 SKS	3,47	> 20 SKS	...	2,00	<= 20 SKS	<=148	3,43	TIDAK TEPAT WAKTU
23	1D01523	LAKI-LAKI	MANDIRI	3,40	<= 20 SKS	3,09	> 20 SKS	3,30	> 20 SKS	3,46	> 20 SKS	...	3,95	<= 20 SKS	<=148	3,26	TEPAT WAKTU
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
40	1D01616	LAKI-LAKI	SBMPTN	3,55	<= 20 SKS	3,11	> 20 SKS	3,34	> 20 SKS	2,95	> 20 SKS	...	2,00	<= 20 SKS	<=148	3,02	TIDAK TEPAT WAKTU
41	1D01617	PEREMPUAN	SBMPTN	3,50	<= 20 SKS	3,50	> 20 SKS	3,34	> 20 SKS	3,54	> 20 SKS	...	3,80	<= 20 SKS	>148	3,55	TEPAT WAKTU
42	1D01618	LAKI-LAKI	SBMPTN	3,79	<= 20 SKS	3,87	> 20 SKS	3,94	> 20 SKS	3,93	> 20 SKS	...	3,95	<= 20 SKS	>148	3,89	TEPAT WAKTU
43	1D01619	PEREMPUAN	SBMPTN	3,84	<= 20 SKS	3,50	> 20 SKS	3,61	> 20 SKS	3,63	> 20 SKS	...	3,95	<= 20 SKS	>148	3,71	TEPAT WAKTU
44	1D01620	LAKI-LAKI	SBMPTN	3,11	<= 20 SKS	2,90	> 20 SKS	2,90	> 20 SKS	2,95	> 20 SKS	...	2,00	<= 20 SKS	<=148	2,83	TIDAK TEPAT WAKTU
45	1D01621	PEREMPUAN	MANDIRI	3,76	<= 20 SKS	3,86	> 20 SKS	3,74	> 20 SKS	3,65	> 20 SKS	...	3,95	<= 20 SKS	>148	3,75	TEPAT WAKTU
46	1D01622	PEREMPUAN	MANDIRI	3,30	<= 20 SKS	3,31	> 20 SKS	3,40	> 20 SKS	3,24	> 20 SKS	...	2,00	<= 20 SKS	<=148	3,09	TIDAK TEPAT WAKTU
47	1D01623	PEREMPUAN	MANDIRI	3,80	<= 20 SKS	3,84	> 20 SKS	3,64	> 20 SKS	3,64	> 20 SKS	...	4,00	<= 20 SKS	>148	3,73	TEPAT WAKTU
48	1D01624	LAKI-LAKI	MANDIRI	2,99	<= 20 SKS	3,43	> 20 SKS	3,47	> 20 SKS	3,20	> 20 SKS	...	3,85	<= 20 SKS	>148	3,30	TEPAT WAKTU
49	1D01625	LAKI-LAKI	MANDIRI	3,23	<= 20 SKS	2,97	> 20 SKS	3,52	> 20 SKS	2,99	> 20 SKS	...	2,00	<= 20 SKS	<=148	2,86	TIDAK TEPAT WAKTU

Tahap selanjutnya adalah proses *encoding* atau konversi dimana dalam tahap ini dilakukan perubahan data ke dalam format data yang siap untuk diaplikasikan. Data yang diperoleh selanjutnya diidentifikasi mengenai jenis atribut dan rentang nilainya untuk mempermudah dalam proses mining selanjutnya. Proses identifikasi atau penentuan atribut yang akan digunakan dalam penelitian ini adalah sebagai berikut,

a. Atribut Keterangan Kelulusan

Interpretasi data dari atribut Keterangan tersaji dalam gambar dibawah ini,



Gambar 4.1 *Pie Chart* untuk Atribut Keterangan

Berdasarkan Gambar 4.1 diperoleh informasi bahwa mahasiswa Program Studi Matematika dalam kategori lulus tepat waktu berjumlah 32 mahasiswa dan

kategori tidak lulus tepat waktu berjumlah 17 mahasiswa.

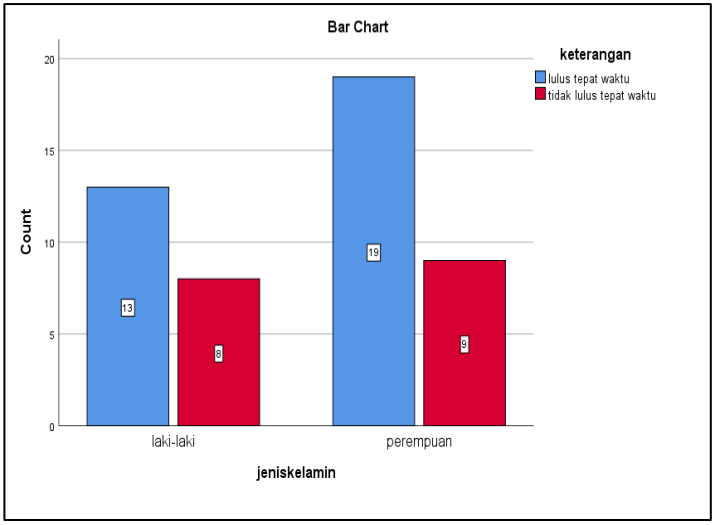
b. Atribut Jenis Kelamin

Dengan menggunakan *Cross Tabulation* didapatkan hasil analisis deskriptif atribut Jenis Kelamin sebagai berikut :

Tabel 4.3 Tabel Silang Atribut Jenis Kelamin Terhadap Keterangan Kelulusan

Keterangan Kelulusan	Jenis Kelamin		Jumlah
	Laki - Laki	Perempuan	
Lulus Tepat Waktu	13 (26,5%)	19 (38,8%)	32 (65,3%)
Tidak Lulus Tepat Waktu	8 (16,3%)	9 (18,4%)	17 (34,7%)
Jumlah	21 (42,8%)	28 (57,1%)	49 (100%)

Dari Tabel 4.3 menunjukkan bahwa sampel yang digunakan dalam proses mining adalah sebesar 26,5% mahasiswa berjenis kelamin laki-laki dalam kategori lulus tepat waktu, 38,8% mahasiswa berjenis kelamin perempuan dalam kategori lulus tepat waktu, 16,3% mahasiswa berjenis kelamin laki-laki dalam kategori tidak lulus tepat waktu dan 18,4% mahasiswa berjenis kelamin perempuan dalam kategori tidak lulus tepat waktu. Data pada Tabel 4.3 disajikan pula dalam bentuk diagram batang sebagai berikut ,



Gambar 4.2 Diagram Batang untuk Atribut jenis Kelamin berdasarkan Keterangan Kelulusan

c. Atribut Jalur Masuk Yang Diikuti

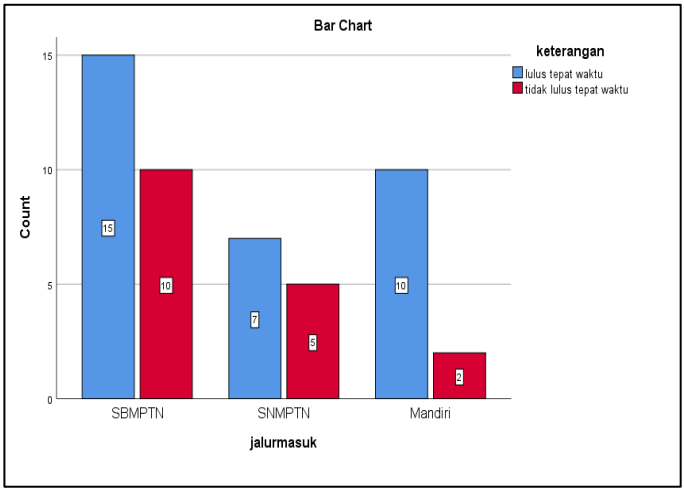
Hasil analisis deskriptif atribut Jalur Masuk yang Diikuti menggunakan *Cross Tabulation* sebagai berikut:

Tabel 4.4 Tabel Silang Atribut Jalur Masuk yang Diikuti Terhadap Keterangan Kelulusan

Keterangan Kelulusan	Jalur Masuk yang Diikuti			Jumlah
	SBMPTN	SNMPTN	Mandiri	
Lulus Tepat Waktu	15 (30,6%)	7 (14,3%)	10 (20,4%)	32 (65,3%)
Tidak Lulus Tepat Waktu	10 (20,4%)	5 (10,2%)	2 (4,1%)	17 (34,7%)

Jumlah	25 (51%)	12 (24,5%)	12 (24,5%)	49 (100%)

Berdasarkan Tabel 4.4 diperoleh informasi bahwa mahasiswa yang masuk melalui jalur SBMPTN sebesar 30,6% dalam kategori lulus tepat waktu dan 20,4% dalam kategori tidak lulus tepat waktu, mahasiswa yang masuk melalui jalur SNMPTN sebesar 14,3% dalam kategori lulus tepat waktu dan 10,2% dalam kategori tidak lulus tepat waktu, mahasiswa yang masuk melalui jalur Mandiri sebesar 20,4% dalam kategori lulus tepat waktu dan 4,1% dalam kategori tidak lulus tepat waktu. Data pada Tabel 4.4 disajikan pula dalam bentuk diagram batang sebagai berikut ,



Gambar 4.3 Diagram Batang untuk Atribut Jalur Masuk yang Diikuti berdasarkan Keterangan Kelulusan

d. Atribut Jumlah SKS yang diambil setiap Semester

Berdasarkan tahapan transformasi pada sampel data diketahui bahwa untuk atribut jumlah SKS yang diambil pada semester 1 hingga semester 8 bertipe kategorik. Jumlah SKS yang diambil tiap semester menggunakan sistem paket yang ditawarkan, sehingga jumlah SKS pada semester 1 hingga semester 8 adalah sama untuk seluruh mahasiswa kecuali pada semester 6 dengan rincian sebagai berikut,

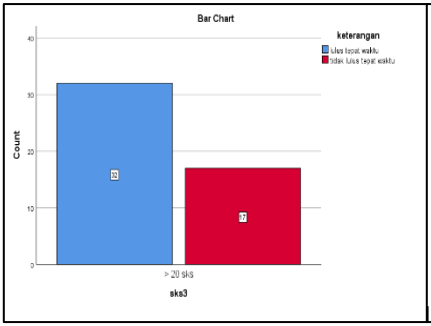
Tabel 4.5 Tabel Silang Atribut Jumlah SKS tiap semester yang diambil Terhadap Keterangan Kelulusan

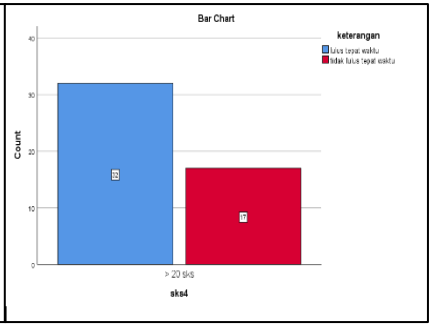
Semester	**Keterangan Kelulusan**	**Jumlah SKS yang diambil**		**Jumlah**
		<= 20 sks	**> 20 sks**	
Semester 1	**Lulus Tepat Waktu**	32 (65,3%)	0 (0 %)	32 (65,3%)
	Tidak Lulus Tepat Waktu	17 (34,7%)	0 (0%)	17 (34,7%)
	Jumlah	49 (100%)	0 (0 %)	49 (100%)
Semester 2	**Lulus Tepat Waktu**	0 (0 %)	32 (65,3%)	32 (65,3%)
	Tidak Lulus Tepat Waktu	0 (0%)	17 (34,7%)	17 (34,7%)
	Jumlah	0 (0 %)	49 (100%)	49 (100%)

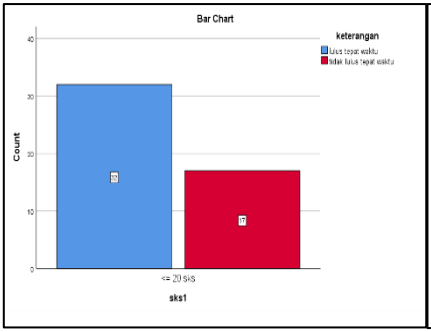
Semester 3	**Lulus Tepat Waktu**	0 (0 %)	32 (65,3%)	32 (65,3%)
	Tidak Lulus Tepat Waktu	0 (0%)	17 (34,7%)	17 (34,7%)
	Jumlah	0 (0 %)	49 (100%)	49 (100%)
Semester 4	**Lulus Tepat Waktu**	0 (0 %)	32 (65,3%)	32 (65,3%)
	Tidak Lulus Tepat Waktu	0 (0%)	17 (34,7%)	17 (34,7%)
	Jumlah	0 (0 %)	49 (100%)	49 (100%)
Semester 5	**Lulus Tepat Waktu**	0 (0 %)	32 (65,3%)	32 (65,3%)
	Tidak Lulus Tepat Waktu	0 (0%)	17 (34,7%)	17 (34,7%)
	Jumlah	0 (0 %)	49 (100%)	49 (100%)
Semester 6	**Lulus Tepat Waktu**	8 (16,3%)	24 (49%)	32 (65,3%)
	Tidak Lulus Tepat Waktu	1 (2%)	16 (32,7)	17 (34,7%)
	Jumlah	9 (18,3%)	40 (81,7%)	49 (100%)
Semester 7	**Lulus Tepat Waktu**	32 (65,3%)	0 (0 %)	32 (65,3%)

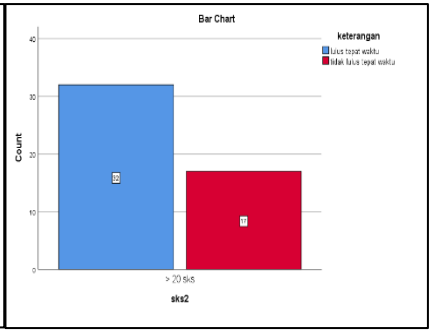
	Tidak Lulus Tepat Waktu	17 (34,7%)	0 (0%)	17 (34,7%)
	Jumlah	49 (100%)	0 (0 %)	49 (100%)
Semester 8	**Lulus Tepat Waktu**	32 (65,3%)	0 (0 %)	32 (65,3%)
	Tidak Lulus Tepat Waktu	17 (34,7%)	0 (0%)	17 (34,7%)
	Jumlah	49 (100%)	0 (0 %)	49 (100%)

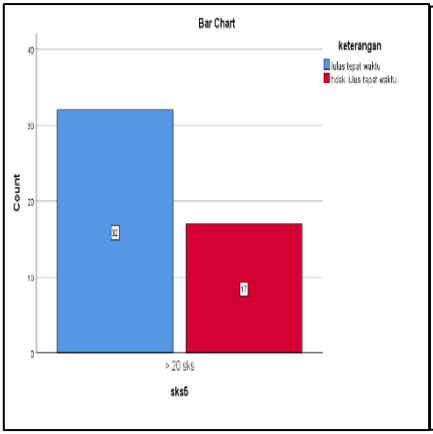
Data pada Tabel 4.5 disajikan pula dalam bentuk diagram batang sebagai berikut,

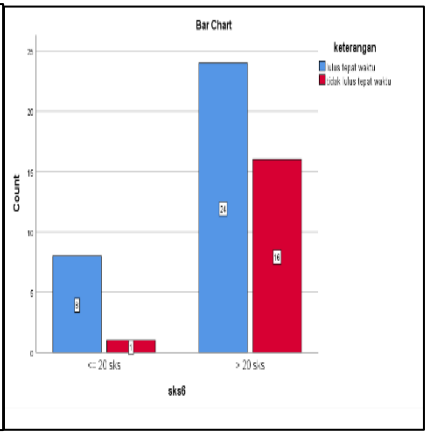

Bar Chart
keterangan
lulus tepat waktu
tidak lulus tepat waktu
Count
32
17
> 20 sks
sks3

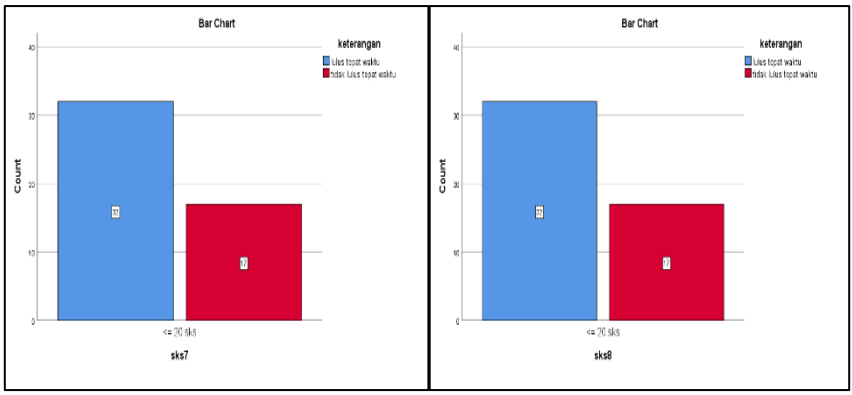

Bar Chart
keterangan
lulus tepat waktu
tidak lulus tepat waktu
Count
32
17
> 20 sks
sks4


Bar Chart
keterangan
lulus tepat waktu
tidak lulus tepat waktu
Count
33
17
<= 20 sks
sks1


Bar Chart
keterangan
lulus tepat waktu
tidak lulus tepat waktu
Count
32
17
> 20 sks
sks2


Bar Chart
keterangan
lulus tepat waktu
tidak lulus tepat waktu
Count
32
17
> 20 sks
sks5


Bar Chart
keterangan
lulus tepat waktu
tidak lulus tepat waktu
Count
8
1
24
16
<= 20 sks
> 20 sks
sks6



Gambar 4.4 Diagram Batang untuk Atribut Jumlah SKS tiap Semester berdasarkan Keterangan Kelulusan

e. Atribut Jumlah Total SKS yang diambil

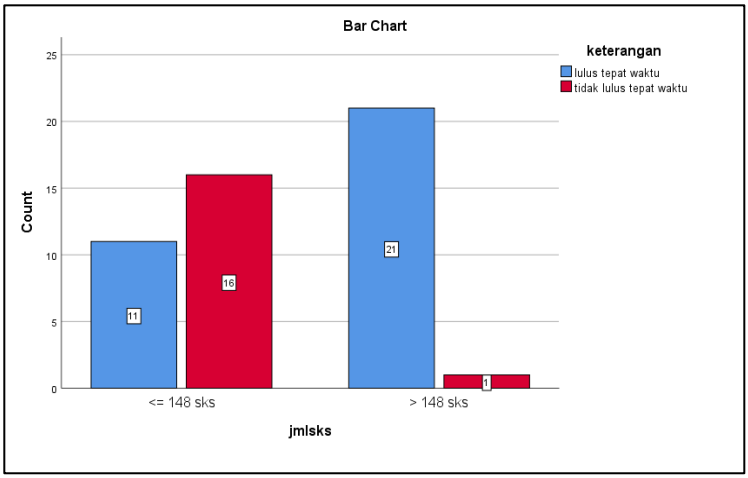
Atribut jumlah SKS menunjukkan jumlah total SKS yang diambil oleh mahasiswa dari semester 1 hingga semester 8. Dengan menggunakan *Cross Tabulation* didapatkan hasil analisis deskriptif atribut Jumlah SKS total sebagai berikut :

Tabel 4.6 Tabel Silang Atribut Jumlah Total SKS yang diambil Terhadap Keterangan Kelulusan

Keterangan Kelulusan	**Jalur Masuk yang Diikuti**		**Jumlah**
	<= 148 SKS	**> 148 SKS**	
Lulus Tepat Waktu	11 (22,4%)	21 (42,9%)	32 (65,3%)
Tidak Lulus Tepat Waktu	16 (32,7%)	1 (2%)	17 (34,7%)

Jumlah	27 (55,1%)	22 (44,9%)	49 (100%)

Data pada Tabel 4.6 disajikan pula dalam bentuk diagram batang sebagai berikut,



Gambar 4.5 Diagram Batang untuk Atribut Jumlah Total SKS yang diambil berdasarkan Keterangan Kelulusan

f. Indeks Prestasi Semester dan Indeks Prestasi Kumulatif

Tipe data pada atribut Indeks Prestasi tiap Semester (IPS) adalah numerik. Dalam proses mining, untuk atribut bertipe numerik diperlukan beberapa ukuran pemusatan dan persebaran data yang tersaji dalam tabel berikut,

Tabel 4.7 Tabel Analisis Deskriptif Untuk Atribut IP Semester 1 – 8 dan IP Kumulatif

Atribut	**Ukuran Deskriptif**

IP Semester 1	Rata-Rata	3,57
	Median	3,64
	Standar Deviasi	0,2526
IP Semester 2	Rata-Rata	3,45
	Median	3,50
	Standar Deviasi	0,252
IP Semester 3	Rata-Rata	3,54
	Median	3,56
	Standar Deviasi	0,278
IP Semester 4	Rata-Rata	3,47
	Median	3,49
	Standar Deviasi	0,2671
IP Semester 5	Rata-Rata	3,36
	Median	3,38
	Standar Deviasi	0,328
IP Semester 6	Rata-Rata	3,43
	Median	3,39
	Standar Deviasi	0,3915
IP Semester 7	Rata-Rata	3,37
	Median	3,53
	Standar Deviasi	0,5633
IP Semester 8	Rata-Rata	3,20
	Median	3,85

	Standar Deviasi	1,0875
IP Kumulatif	Rata-Rata	3,44
	Median	3,41
	Standar Deviasi	0,2821

4.2 Analisis Algoritma Naïve Bayes

Tahapan analisis metode data mining yang dilakukan adalah pemodelan data yang bertujuan untuk memprediksi pola lulusan mahasiswa menggunakan Algoritma Naïve Bayes. Data yang telah dikumpulkan kemudian dilakukan pembersihan, penyeleksian dan transformasi untuk keperluan proses data mining selanjutnya.

Langkah – Langkah perhitungan implementasi Algoritma Naïve Bayes adalah sebagai berikut :

1) Persiapan data Training seperti yang tercantum pada tabel 4.2.
2) Perhitungan nilai probabilitas prior $P(C_i)$ setiap kelas.

Sebelum menentukan nilai probabilitas prior setiap kelas, akan ditentukan terlebih dahulu jumlah data dari setiap kelas. Probabilitas prior diperoleh dari hasil pembagian jumlah data setiap kelas dengan jumlah keseluruhan data.

Tabel 4.8 Data Keterangan Kelulusan

No	Kategori	Jumlah
1	Tidak Lulus Tepat Waktu	17
2	Lulus Tepat Waktu	32
Total		49

Pencarian kelayakan metode Naïve Bayes dilakukan dengan mencari probabilitas dari masing-masing kelas. Prediksi pola lulusan mahasiswa dikategorikan dua jenis yaitu "Lulus Tepat Waktu" dan "Tidak Lulus Tepat Waktu". Probabilitas dari masing-masing kelas adalah,

$$P(Lulus\ Tepat\ Waktu) = \frac{32}{49}$$

$$P(Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{49}$$

3) Perhitungan nilai probabilitas posterior $P(x_k|C_i)$ dari masing – masing atribut

Probabilitas posterior dilakukan pada data *training* yang berjumlah 49 data dengan menghitung probabilitas bersyarat untuk setiap kelas yaitu,

a. Atribut Jenis Kelamin

Berdasarkan Tabel 4.3, nilai probabilitas dari masing-masing kelas berdasarkan atribut jenis kelamin adalah

$$P(JK = P\ |class = Lulus\ Tepat\ Waktu) = \frac{19}{32}$$

$$P(JK = L\ |class = Lulus\ Tepat\ Waktu) = \frac{13}{32}$$

$$P(JK = P\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{9}{17}$$

$$P(JK = L\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{8}{17}$$

b. Atribut Jalur Seleksi Masuk yang diikuti

Berdasarkan Tabel 4.4, nilai probabilitas dari masing-masing kelas berdasarkan atribut jalur seleksi masuk yang diikuti adalah

$$P(JM = SBMPTN \mid class = Lulus\ Tepat\ Waktu) = \frac{15}{32}$$

$$P(JM = SNMPTN \mid class = Lulus\ Tepat\ Waktu) = \frac{7}{32}$$

$$P(JM = Mandiri \mid class = Lulus\ Tepat\ Waktu) = \frac{10}{32}$$

$$P(JM = SBMPTN \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{10}{17}$$

$$P(JM = SNMPTN \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{5}{17}$$

$$P(JM = Mandiri \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{2}{17}$$

c. Atribut Jumlah SKS total yang diambil

Berdasarkan Tabel 4.6, nilai probabilitas dari masing-masing kelas berdasarkan atribut jumlah SKS total adalah

$$P(jmlsks \leq 148\ sks\ |class = Lulus\ Tepat\ Waktu) = \frac{11}{32}$$

$$P(jmlsks > 148\ sks\ |class = Lulus\ Tepat\ Waktu) = \frac{21}{32}$$

$$P(jmlsks > 148\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{1}{17}$$

$$P(jmlsks \leq 148\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{16}{17}$$

d. Atribut Jumlah SKS yang diambil setiap semester

Berdasarkan Tabel 4.6, nilai probabilitas dari masing-masing kelas berdasarkan atribut jumlah SKS yang diambil setiap semester (semester 1 – semester 8) adalah

Semester 1

$$P(jmlsks\ \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks\ \leq 20\ sks\ |class = \ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

Semester 2

$$P(jmlsks\ \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks \leq 20\ sks\,|class = \ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

$$P(jmlsks > 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks > 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

Semester 3

$$P(jmlsks \leq 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks \leq 20\ sks\,|class = \ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

$$P(jmlsks > 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks > 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

Semester 4

$$P(jmlsks \leq 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks \leq 20\ sks\,|class = \ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

$$P(jmlsks > 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks > 20\ sks\,|class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

Semester 5

$$P(jmlsks \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks \leq 20\ sks\ |class = \ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

Semester 6

$$P(jmlsks \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{1}{17}$$

$$= 0{,}0588$$

$$P(jmlsks \leq 20\ sks\ |class = \ Lulus\ Tepat\ Waktu) = \frac{8}{32} = 0{,}25$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{16}{17}$$

$$= 0{,}9412$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{24}{32}$$

$$= 0{,}75$$

Semester 7

$$P(jmlsks \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks \leq 20\ sks\ |class = Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

Semester 8

$$P(jmlsks \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks \leq 20\ sks\ |class = Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{17} = 0$$

$$P(jmlsks > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{0}{32} = 0$$

e. Atribut IP Semester 1 hingga IP Semester 8 dan IP Kumulatif

Atribut IP Semester 1 hingga IP Semester 8 dan IP Kumulatif adalah atribut dengan tipe data numerik sehingga perhitungan probabilitas posterior menggunakan perhitungan Distribusi Gaussian. Dalam perhitungan ini terlebih dahulu akan ditentukan nilai rata-rata μ dan standar deviasi σ sesuai dengan persamaan berikut ,

$$P(x_k|C_i) = \frac{1}{\sigma_{C_i}\sqrt{2\pi}}\, e^{-\frac{(x-\mu_{C_i})}{2\,\sigma_{C_i}^{\,2}}}$$

Dengan

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 1 adalah,

Tabel 4.9 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 1

IP Semester 1	Lulus Tepat Waktu	Rata – Rata	3,608
		Standar Deviasi	0,2454
	Tidak Lulus Tepat Waktu	Rata – Rata	3,496
		Standar Deviasi	0,2569

$$P(ips1 = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2454} e^{\frac{-(x_1-3{,}608)^2}{2(0{,}2454)^2}}$$

$$P(ips1 = x_1|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2569} e^{\frac{-(x_1-3{,}496)^2}{2(0{,}2569)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 2 adalah,

Tabel 4.10 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 2

IP Semester 2	Lulus Tepat Waktu	Rata – Rata	3,483
		Standar Deviasi	0,2471
	Tidak Lulus Tepat Waktu	Rata – Rata	3,372
		Standar Deviasi	0,2528

$$P(ips2 = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2472} e^{\frac{-(x_1-3{,}483)^2}{2(0{,}2472)^2}}$$

$$P(ips2 = x_1|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2528} e^{\frac{-(x_1-3{,}372)^2}{2(0{,}2528)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 3 adalah,

Tabel 4.11 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 3

IP Semester 3	Lulus Tepat Waktu	Rata – Rata	3,483
		Standar Deviasi	0,2471
	Tidak Lulus Tepat Waktu	Rata – Rata	3,414
		Standar Deviasi	0,2857

$$P(ips3 = x_1 | class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2471} e^{\frac{-(x_1-3{,}483)^2}{2(0{,}2471)^2}}$$

$$P(ips3 = x_1 | class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2857} e^{\frac{-(x_1-3{,}414)^2}{2(0{,}2857)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 4 adalah,

Tabel 4.12 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 4

IP Semester 4	Lulus Tepat Waktu	Rata – Rata	3,554
		Standar Deviasi	0,2561

	Tidak Lulus Tepat Waktu	Rata – Rata	3,309
		Standar Deviasi	0,2145

$$P(ips4 = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2561} e^{\frac{-(x_1-3{,}554)^2}{2(0{,}2561)^2}}$$

$$P(ips4 = x_1|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2145} e^{\frac{-(x_1-3{,}309)^2}{2(0{,}2145)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 5 adalah,

Tabel 4.13 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 5

IP Semester 5	Lulus Tepat Waktu	Rata – Rata	3,413
		Standar Deviasi	0,3255
	Tidak Lulus Tepat Waktu	Rata – Rata	3,261
		Standar Deviasi	0,3182

$$P(ips5 = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}3255} e^{\frac{-(x_1-3{,}413)^2}{2(0{,}3255)^2}}$$

$$P(ips5 = x_1 | class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}3182} e^{\frac{-(x_1-3{,}261)^2}{2(0{,}3182)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 6 adalah,

Tabel 4.14 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 6

IP Semester 6	Lulus Tepat Waktu	Rata – Rata	3,369
		Standar Deviasi	0,2692
	Tidak Lulus Tepat Waktu	Rata – Rata	3,547
		Standar Deviasi	0,5459

$$P(ips6 = x_1 | class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}2692} e^{\frac{-(x_1-3{,}369)^2}{2(0{,}2692)^2}}$$

$$P(ips6 = x_1 | class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.0{,}5459} e^{\frac{-(x_1-3{,}547)^2}{2(0{,}5459)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 7 adalah,

Tabel 4.15 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 7

IP Semester 7	Lulus Tepat Waktu	Rata – Rata	3,719
		Standar Deviasi	0,2435
	Tidak Lulus Tepat Waktu	Rata – Rata	2,706
		Standar Deviasi	0,3499

$$P(ips7 = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.\,0{,}2435} e^{\frac{-(x_1-3{,}729)^2}{2(0{,}2435)^2}}$$

$$P(ips7 = x_1|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.\,0{,}3499} e^{\frac{-(x_1-2{,}706)^2}{2(0{,}3499)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Semester 8 adalah,

Tabel 4.16 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Semester 8

IP Semester 8	Lulus Tepat Waktu	Rata – Rata	3,389
		Standar Deviasi	0,5156
	Tidak Lulus Tepat Waktu	Rata – Rata	1,874
		Standar Deviasi	0,8348

$$P(ips8 = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.\,0{,}5156} e^{\frac{-(x_1-3{,}389)^2}{2(0{,}5156)^2}}$$

$$P(ips8 = x_1|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.\,0{,}8348} e^{\frac{-(x_1-3{,}389)^2}{2(0{,}8348)^2}}$$

Probabilitas dari masing-masing kelas berdasarkan atribut IP Kumulatif adalah

Tabel 4.17 Nilai Rata – Rata dan Standar Deviasi untuk Atribut IP Kumulatif

IP Kumulatif	Lulus Tepat Waktu	Rata – Rata	3,559
		Standar Deviasi	0,2243
	Tidak Lulus Tepat Waktu	Rata – Rata	3,215
		Standar Deviasi	0,2444

$$P(ipk = x_1|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.\,0{,}2243} e^{\frac{-(x_1-3{,}559)^2}{2(0{,}2243)^2}}$$

$$P(ipk = x_1|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}.\,0{,}2444} e^{\frac{-(x_1-3{,}215)^2}{2(0{,}2444^2}}$$

Prediksi data testing pada Algoritma Naïve Bayes menggunakan data mahasiswa Program Studi Matematika Angkatan Tahun 2017 yang telah lulus dengan mensubstitusikan nilai dari masing – masing atribut ke fungsi distribusi Gaussian sebagai berikut ,

Tabel 4.18 Data *Testing*

MAHASISWA	1	2	3	4
JENIS KELAMIN	PEREMPUAN	PEREMPUAN	PEREMPUAN	PEREMPUAN
JALUR MASUK	MANDIRI	SNMPTN	SBMPTN	SNMPTN
IPS1	3,76	3,90	3,84	3,30
SKS1	<= 20 SKS	<= 20 SKS	<= 20 SKS	<= 20 SKS
IPS2	3,86	3,79	3,39	3,03
SKS2	> 20 SKS	> 20 SKS	> 20 SKS	> 20 SKS
IPS3	3,74	3,90	3,46	3,41
SKS3	> 20 SKS	> 20 SKS	> 20 SKS	> 20 SKS
IPS4	3,65	3,69	3,53	3,03
SKS4	> 20 SKS	> 20 SKS	> 20 SKS	> 20 SKS
IPS5	3,20	3,56	3,56	2,89
SKS5	> 20 SKS	> 20 SKS	> 20 SKS	> 20 SKS
IPS6	2,42	2,14	2,06	2,55
SKS6	<= 20 sks	<= 20 sks	<= 20 sks	> 20 sks
IPS7	2,38	2,30	2,24	2,18
SKS7	<= 20 sks	<= 20 sks	<= 20 sks	<= 20 sks
IPS8	3,95	3,95	3,90	3,75
SKS8	<= 20 sks	<= 20 sks	<= 20 sks	<= 20 sks
IPK	3,37	3,40	3,25	3,02
JUMLAH SKS TOTAL	> 148 sks	> 148 sks	> 148 sks	> 148 sks
KETERANGAN	?	?	?	?

Pengujian data *testing* mahasiswa ke-1 :

$$P(JK = P|Class = Tidak\ Tepat\ Waktu) = \frac{9}{17}$$

$$P(JK = P|Class = Tepat\ Waktu) = \frac{8}{17}$$

$$P(JM = Mandiri\ |Class = Tepat\ Waktu) = \frac{10}{32}$$

$$P(JM = Mandiri\ |Class = Tidak\ Tepat\ Waktu) = \frac{2}{17}$$

$$P(jmlsks > 148\ sks\ |Class = Tidak\ Lulus\ \ Tepat\ Waktu) = \frac{1}{17}$$

$$P(jmlsks > 148\ sks\ |Class = Lulus\ \ Tepat\ Waktu) = \frac{21}{32}$$

$$P(jmlsks1\ \leq 20\ sks\ |class = \ Tidak\ Lulus\ Tepat\ waktu = \frac{17}{17} = 1$$

$$P(jmlsks1\ \leq 20\ sks\ |class = \ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks2 > 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks2 > 20\ sks\ |class = \ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks3 > 20\ sks \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks3 > 20\ sks \mid class = Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks4 > 20\ sks \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks4 > 20\ sks \mid class = \ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks5 > 20\ sks \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$$

$$P(jmlsks5 > 20\ sks \mid class = Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$$

$$P(jmlsks6 \leq 20\ sks \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{1}{17}$$

$$P(jmlsks6\ \leq 20\ sks \mid class = \ Lulus\ Tepat\ Waktu) = \frac{8}{32}$$

$$P(jmlsks6 > 20\ sks \mid class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{16}{17}$$

$P(jmlsks6 > 20\ sks\ |class = Lulus\ Tepat\ Waktu) = \frac{24}{32}$

$P(jmlsks7 \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$

$P(jmlsks7\ \leq 20\ sks\ |class =\ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$

$P(jmlsks8 \leq 20\ sks\ |class = Tidak\ Lulus\ Tepat\ Waktu) = \frac{17}{17} = 1$

$P(jmlsks8\ \leq 20\ sks\ |class =\ Lulus\ Tepat\ Waktu) = \frac{32}{32} = 1$

$$P(ips1 = 3{,}76|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi * 0{,}2454}} e^{\frac{-(3{,}76-3{,}608)^2}{2(0{,}2454)^2}} = 0{,}805888$$

$$P(ips1 = 3{,}76|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi * 0{,}2569}} e^{\frac{-(3{,}76-3{,}496)^2}{2(0{,}2569)^2}} = 0{,}7889088$$

$$P(ips2 = 3{,}86|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi * 0{,}2472}} e^{\frac{-(3{,}86-3{,}483)^2}{2(0{,}2472)^2}} = 0{,}80604286$$

$$P(ips2 = 3{,}86|class = tidak\ lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}2528} e^{\frac{-(3{,}86-3{,}372)^2}{2(0{,}2528)^2}} = 0{,}7995145$$

$$P(ips3 = 3{,}74|class = lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}2471} e^{\frac{-(3{,}74-3{,}483)^2}{2(0{,}2471)^2}} = 0{,}8041728$$

$$P(ips3 = 3{,}74|class = tidak\ lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}2857} e^{\frac{-(3{,}74-3{,}414)^2}{2(0{,}2857)^2}} = 0{,}7496156$$

$$P(ips4 = 3{,}65|class = tidak\ lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}2145} e^{\frac{-(3{,}65-3{,}309)^2}{2(0{,}2145)^2}} = 0{,}863690576$$

$$P(ips4 = 3{,}65|class = lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}2561} e^{\frac{-(3{,}65-3{,}554)^2}{2(0{,}2561)^2}} = 0{,}788563247$$

$$P(ips5 = 3{,}20|class = tidak\ lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}3182} e^{\frac{-(3{,}20-3{,}261)^2}{2(0{,}3182)^2}} = 0{,}707362105$$

$$P(ips5 = 3{,}20|class = lulus\ tepat\ waktu)$$
$$= \frac{1}{\sqrt{2\pi}*0{,}3255} e^{\frac{-(3{,}20-3{,}413)^2}{2(0{,}3255)^2}} = 0{,}700935999$$

$$P(ips6 = 2{,}42|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}5459} e^{\frac{-(2{,}42-3{,}547)^2}{2(0{,}5459)^2}} = 0{,}652447076$$

$$P(ips6 = 2{,}42|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}2692} e^{\frac{-(2{,}42-3{,}269)^2}{2(0{,}2692)^2}} = 0{,}79441004$$

$$P(ips7 = 2{,}38|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}3499} e^{\frac{-(2{,}38-2{,}706)^2}{2(0{,}3499)^2}} = 0{,}678833831$$

$$P(ips7 = 2{,}38|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}2435} e^{\frac{-(2{,}38-3{,}719)^2}{2(0{,}2435)^2}} = 0{,}852598781$$

$$P(ips8 = 3{,}00|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}8348} e^{\frac{-(3{,}00-1{,}875)^2}{2(0{,}8348)^2}} = 0{,}678646015$$

$$P(ips8 = 3{,}00|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}5156} e^{\frac{-(3{,}00-3{,}389)^2}{2(0{,}5156)^2}} = 0{,}566877137$$

$$P(ipk = 3{,}37|class = tidak\ lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}2444} e^{\frac{-(3{,}37-3{,}215)^2}{2(0{,}2444)^2}} = 0{,}807553724$$

$$P(ipk = 3{,}37|class = lulus\ tepat\ waktu) = \frac{1}{\sqrt{2\pi}*0{,}2243} e^{\frac{-(3{,}37-3{,}559)^2}{2(0{,}2243)^2}} = 0{,}843112098$$

Perhitungan selanjutnya adalah dengan menentukan nilai fungsi likelihood untuk masing-masing kriteria keputusan (Lulus Tepat waktu / Tidak Lulus Tepat Waktu) yaitu ,

Likelihood Lulus Tepat Waktu

$$= \frac{8}{17} * \frac{10}{32} * \frac{21}{32} * 1 * 1 * 1 * 1 * 1 * \frac{8}{32} * 1 * 1 * 0{,}80588418 * 0{,}8060428 * 0{,}804172865 * 0{,}7885632 * 0{,}700935999 * 0{,}79441004 * 0{,}8525988 * 0{,}566877137 * 0{,}843112098$$

$$= 0{,}002255065$$

Likelihood Tidak Lulus Tepat Waktu

$$= \frac{9}{17} * \frac{2}{17} * \frac{1}{17} * 1 * 1 * 1 * 1 * 1 * \frac{1}{17} * 1 * 1 * 0{,}788908873 * 0{,}799514516 * 0{,}74961538 * 0{,}863690576 * 0{,}707362105 * 652447076 * 0{,}678833831 * 0{,}678646015 * 0{,}807553724$$

$$= 0{,}0001511$$

Karena nilai *Likelihood* Lulus Tepat Waktu lebih besar dibandingkan nilai *Likelihood* Tidak Lulus Tepat Waktu dapat disimpulkan bahwa hasil prediksi pola lulusan mahasiswa tersebut adalah Lulus Tepat Waktu.

4.3 Analisis Algoritma C5.0

Algoritma C5.0 merupakan algoritma data mining yang digunakan untuk mengklasifikasikan data dengan menghasilkan model pohon keputusan (*decision tree*). Langkah pertama yang dilakukan untuk membentuk model pohon keputusan adalah menghitung jumlah kejadian untuk masing-masing atribut. Algoritma C5.0 bekerja dengan cara memilih atribut berdasarkan nilai gain ratio tertinggi.

Tahapan dalam membuat decision tree menggunakan algoritma C5.0 yaitu :

1. Mempersiapkan data training. Data *training* seringkali diambil dari data histori yang pernah terjadi sebelumnya dan telah dikelompokkan ke dalam kelas-kelas tertentu. Data *training* yang digunakan seperti yang tersaji pada Tabel 4.2.
2. Menghitung nilai entropy total seluruh data berdasarkan atribut. Untuk menentukan nilai entropy, pertama dihitung jumlah kejadian untuk setiap atribut. Atribut yang digunakan dalam penelitian ini terdiri atas atribut kategorik dan atribut numerik. Jumlah kejadian berdasarkan atribut bertipe kategorik dikelompokkan menggunakan teknik *Cross Tabulation* yang dijabarkan pada tabel 4.19 berikut,

Tabel 4.19 Jumlah Kejadian Untuk Atribut Kategorik

Atribut	Kategori	Jumlah	Lulus Tepat Waktu	Tidak Lulus Tepat Waktu
Jenis Kelamin	Laki-Laki	21	13	8
	Perempuan	28	19	9
Jalur Masuk	SBMPTN	25	15	10
	SNMPTN	12	7	5
	Mandiri	12	10	2

Jumlah SKS Semester 1	<= 20 SKS	49	32	17
	> 20 SKS	0	0	0
Jumlah SKS Semester 2	<= 20 SKS	0	0	0
	> 20 SKS	49	32	17
Jumlah SKS Semester 3	<= 20 SKS	0	0	0
	> 20 SKS	49	32	17
Jumlah SKS Semester 4	<= 20 SKS	0	0	0
	> 20 SKS	49	32	17
Jumlah SKS Semester 5	<= 20 SKS	0	0	0
	> 20 SKS	49	15	34
Jumlah SKS Semester 6	<= 20 SKS	9	8	1
	> 20 SKS	40	24	16
Jumlah SKS Semester 7	<= 20 SKS	0	0	0
	> 20 SKS	49	32	17
Jumlah SKS Semester 8	<= 20 SKS	0	0	0
	> 20 SKS	49	32	17
Jumlah SKS Total	<= 148 SKS	27	11	16
	> 148 SKS	22	21	1

Dalam menentukan jumlah kejadian untuk atribut bertipe numerik dikelompokkan berdasarkan nilai rata-rata dan median. Selanjutnya, ditentukan kategori berdasarkan nilai rata-rata dan median dapat dilihat pada Tabel 4.20.

Tabel 4.20 Perhitungan Nilai Rata-Rata dan Median Untuk Atribut Numerik

Atribut	Rata-Rata	Median
IP Semester 1	3,57	3,64
IP Semester 2	3,45	3,50
IP Semester 3	3,54	3,56
IP Semester 4	3,47	3,49
IP Semester 5	3,36	3,38
IP Semester 6	3,43	3,39
IP Semester 7	3,37	3,53
IP Semester 8	3,20	3,85
IP Kumulatif	3,44	3,41

Tahapan selanjutnya menentukan jumlah kejadian dari setiap atribut bertipe numerik yang dikelompokkan berdasarkan kategori nilai rata-rata dan median yang tersaji pada Tabel 4.21.

Tabel 4.21 Jumlah Kejadian Untuk Atribut Numerik

Atribut	Kategori	Jumlah Total	Lulus tepat Waktu	Tidak Lulus Tepat Waktu
IP Semester 1	Berdasarkan Nilai Rata-Rata			
	<= 3,57	23	14	9
	>3,57	26	18	8

	Berdasarkan Nilai Median			
	<=3,64	25	15	10
	> 3,64	24	17	7
IP Semester 2	Berdasarkan Nilai Rata-Rata			
	<=3,45	21	12	9
	> 3,45	28	20	8
	Berdasarkan Nilai Median			
	<=3,50	26	16	10
	> 3,50	23	16	7
IP Semester 3	Berdasarkan Nilai Rata-Rata			
	<= 3,54	23	11	12
	>3,54	26	21	5
	Berdasarkan Nilai Median			
	<=3,56	25	13	12
	> 3,56	24	19	5
IP Semester 4	Berdasarkan Nilai Rata-Rata			
	<= 3,47	22	10	12
	> 3,47	27	22	5
	Berdasarkan Nilai Median			
	<=3,49	26	12	14
	> 3,49	23	20	3
IP Semester 5	Berdasarkan Nilai Rata-Rata			
	<= 3,36	24	13	11
	> 3,36	25	19	6
	Berdasarkan Nilai Median			

	<= 3,38	25	14	11
	> 3,38	24	18	6
IP Semester 6	Berdasarkan Nilai Rata-Rata			
	<= 3,43	29	19	10
	> 3,43	20	13	7
	Berdasarkan Nilai Median			
	<= 3,39	25	17	8
	> 3,39	24	15	9
IP Semester 7	Berdasarkan Nilai Rata-Rata			
	<= 3,37	20	4	16
	> 3,37	29	28	1
	Berdasarkan Nilai Median			
	<= 3,53	24	8	16
	> 3,53	25	24	1
IP Semester 8	Berdasarkan Nilai Rata-Rata			
	<= 3,20	16	0	16
	> 3,20	33	32	1
	Berdasarkan Nilai Median			
	<= 3,85	28	11	17
	> 3,85	21	21	0
IP Kumulatif	Berdasarkan Nilai Rata-Rata			
	<= 3,44	26	11	15
	> 3,44	23	21	2
	Berdasarkan Nilai Median			
	<= 3,41	25	11	14
	> 3,41	24	21	3

Setelah diperoleh jumlah kejadian setiap atribut pada Tabel 4.21, selanjutnya akan dihitung nilai entropy total. Berdasarkan persamaan (2.1) dapat dihitung nilai entropy untuk keseluruhan data sampel sebagai berikut,

$$Entropy(Total) = \sum_i^c - p_i \, log_2 \, p_i$$

$$Entropy(Total) = -\left[\frac{32}{49} log_2 \frac{32}{49} + \frac{17}{49} log_2 \, \frac{17}{49}\right]$$

$$= 0{,}9313$$

Setelah memperoleh entropy dari keseluruhan data sampel, kemudian dilakukan analisis pada setiap atribut dan nilainya. Perhitungan entropy untuk setiap atribut tersaji pada Tabel 4.22.

Tabel 4.22 Perhitungan Entropy Seluruh Atribut

Atribut	**Kategori**	**Jumlah**	**Lulus Tepat Waktu**	**Tidak Lulus Tepat Waktu**	**Entropy**
Jenis Kelamin	Laki-Laki	21	13	8	0,958711883
	Perempuan	28	19	9	0,905928216
Jalur Masuk	SBMPTN	25	15	10	0,970950594
	SNMPTN	12	7	5	0,979868757
	Mandiri	12	10	2	0,650022422
Atribut	**Kategori**	**Jumlah**	**Lulus Tepat Waktu**	**Tidak Lulus Tepat Waktu**	**Entropy**
	<= 20 SKS	49	32	17	0,931304369

Jumlah SKS Semester 1	> 20 SKS	0	0	0	0
Jumlah SKS Semester 2	<= 20 SKS	0	0	0	0
	> 20 SKS	49	32	17	0,931304369
Jumlah SKS Semester 3	<= 20 SKS	0	0	0	0
	> 20 SKS	49	32	17	0,931304369
Jumlah SKS Semester 4	<= 20 SKS	0	0	0	0
	> 20 SKS	49	32	17	0,931304369
Jumlah SKS Semester 5	<= 20 SKS	0	0	0	0
	> 20 SKS	49	15	34	0,931304369
Jumlah SKS Semester 6	<= 20 SKS	9	8	1	0,503258335
	> 20 SKS	40	24	16	1,03161824
Jumlah SKS Semester 7	<= 20 SKS	0	0	0	0
	> 20 SKS	49	32	17	0,931304369
Jumlah SKS Semester 8	<= 20 SKS	0	0	0	0
	> 20 SKS	49	32	17	0,931304369
Jumlah SKS Total	<= 148 SKS	27	11	16	0,975119065
	> 148 SKS	22	21	1	0,266764988
IP Semester 1	<= 3,57	23	14	9	0,965636133
	> 3,57	26	18	8	0,89049164
	<=3,64	25	15	10	0,970950594
	> 3,64	24	17	7	0,870864469

IP Semester 2	<=3,45	21	12	9	0,985228136
	> 3,45	28	20	8	0,863120569
	<=3,50	26	16	10	0,961236605
	> 3,50	23	16	7	0,886540893
IP Semester 3	<= 3,54	23	11	12	0,998635964
	>3,54	26	21	5	0,706274089
	<=3,56	25	13	12	0,998845536
	> 3,56	24	19	5	0,738284866
IP Semester 4	<= 3,47	22	10	12	0,994030211
	> 3,47	27	22	5	0,691289869
	<=3,49	26	12	14	0,995727452
	> 3,49	23	20	3	0,558629373
IP Semester 5	<= 3,36	24	13	11	0,994984828
	> 3,36	25	19	6	0,795040279
	<= 3,38	25	14	11	0,989587521
	> 3,38	24	18	6	0,811278124
IP Semester 6	<= 3,43	29	19	10	0,929363626
	> 3,43	20	13	7	0,934068055

	<= 3,39	25	17	8	0,904381458
	> 3,39	24	15	9	0,954434003
IP Semester 7	<= 3,37	20	4	16	0,721928095
	> 3,37	29	28	1	0,216396932
	<= 3,53	24	8	16	0,918295834
	> 3,53	25	24	1	0,242292189
IP Semester 8	<= 3,20	16	0	16	0
	> 3,20	33	32	1	0,195909271
	<= 3,85	28	11	17	0,966618633
	> 3,85	21	21	0	0
IP Kumulatif	<= 3,44	26	11	15	0,98285869
	> 3,44	23	21	2	0,426228657
	<= 3,41	25	11	14	0,989587521
	> 3,41	24	21	3	0,543564443

3. Menentukan *node* akar (*root*) dari pohon. *Node* akar diambil dari atribut yang terpilih dengan cara menghitung nilai gain dari masing-masing atribut. Langkah selanjutnya adalah menghitung nilai *Information Gain* menggunakan Persamaan (2.2). Nilai *Information Gain* digunakan untuk menentukan *node*

akar (*root*). Sebagai contoh, untuk atribut Jenis Kelamin, maka persamaan (2.2) didefinisikan sebagai,

$$Gain(Total, Jenis\ Kelamin) = Entropy(S) - \sum_{i=1}^{n} \frac{|Jenis\ Kelamin_i|}{|Total|}\ Entropy(Jenis\ Kelamin_i)$$

Perhitungan *information Gain* digunakan untuk menentukan *node* akar (*root*) ke-1 pada Algoritma C5.0. Sehingga hasil perhitungan *Information Gain* untuk seluruh atribut tertampil dalam Tabel 4.23.

Tabel 4.23 Perhitungan Entropy dan *Information Gain Node* Akar (*Root*) ke-1

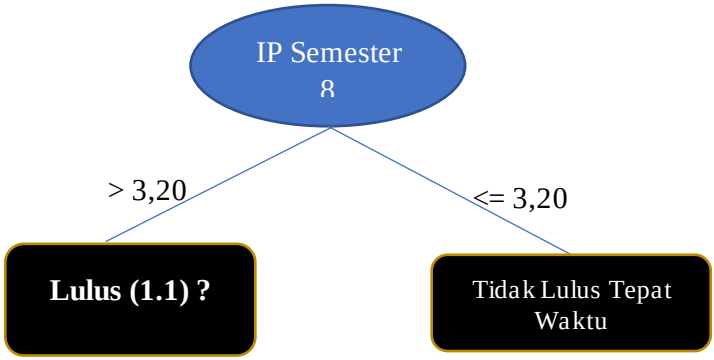
Atribut	**Kategori**	***Entropy***	***Gain***	***Information Gain C5.0***
Jenis Kelamin	Laki-Laki	0,958712	0,00275	0,0014773
	Perempuan	0,905928		
Jalur Masuk	SBMPTN	0,970951	0,0367644	0,0141355
	SNMPTN	0,979869		
	Mandiri	0,650022		
Jumlah SKS Semester 1	<= 20 SKS	0,931304	0,00000	0,00000
	> 20 SKS	0,00000		
Jumlah SKS Semester 2	<= 20 SKS	0,00000	0,00000	0,00000
	> 20 SKS	0,931304		
	<= 20 SKS	0,00000	0,00000	0,00000

Jumlah SKS Semester 3	> 20 SKS	0,931304		
Jumlah SKS Semester 4	<= 20 SKS	0,00000	0,00000	0,00000
	> 20 SKS	0,931304		
Jumlah SKS Semester 5	<= 20 SKS	0,00000	0,00000	0,00000
	> 20 SKS	0,931304		
Jumlah SKS Semester 6	<= 20 SKS	0,503258	-0,19275	-0,1255795
	> 20 SKS	1,03162		
Jumlah SKS Semester 7	<= 20 SKS	0,00000	0,00000	0,00000
	> 20 SKS	0,931304		
Jumlah SKS Semester 8	<= 20 SKS	0,00000	0,00000	0,00000
	> 20 SKS	0,931304		
Jumlah SKS Total	<= 148 SKS	0,975119	0,27422	0,2208111
	> 148 SKS	0,266765		
IP Semester 1	<= 3,57	0,965636133	0,00554	0,002985152
	> 3,57	0,89049164		
	<=3,64	0,970950594	0,00934	0,005090386
	> 3,64	0,870864469		
IP Semester 2	<=3,45	0,985228136	0,01585	0,008576296
	> 3,45	0,863120569		
	<=3,50	0,961236605	0,00513	2,7760E-03
	> 3,50	0,886540893		

IP Semester 3	<= 3,54	0,998635964	0,08780	0,051497846
	>3,54	0,706274089		
	<=3,56	0,998845536	0,06008	0,034585996
	> 3,56	0,738284866		
IP Semester 4	<= 3,47	0,994030211	0,10409	0,061762905
	> 3,47	0,691289869		
	<=3,49	0,995727452	0,14075	0,090548965
	> 3,49	0,558629373		
IP Semester 5	<= 3,36	0,994984828	0,03833	0,021414261
	> 3,36	0,795040279		
	<= 3,38	0,989587521	0,02905	0,016132276
	> 3,38	0,811278124		
IP Semester 6	<= 3,43	0,929363626	0,00002	1,10373E-05
	> 3,43	0,934068055		
	<= 3,39	0,904381458	0,00241	0,001295114
	> 3,39	0,954434003		
IP Semester 7	<= 3,37	0,721928095	0,50857	0,54199576
	> 3,37	0,216396932		
	<= 3,53	0,918295834	0,35791	0,308385358
	> 3,53	0,242292189		
IP Semester 8	<= 3,20	0	0,79937	4,080284043
	> 3,20	0,195909271		
	<= 3,85	0,966618633	0,37895	0,392037616

	> 3,85	0		
IP Kumulatif	<= 3,44	0,98285869	0,20972	0,148834635
	> 3,44	0,426228657		
	<= 3,41	0,989587521	0,16018	0,104475704
	> 3,41	0,543564443		

Pada tahap *node* akar ke-1 akan diambil atribut yang memiliki nilai *Information Gain* tertinggi. Dari Tabel 4.23 dapat dilihat pada atribut bertipe numerik meliputi IP Semester 1 hingga Semester 8 dan IP Kumulatif terdapat masing-masing 2 kelas yang dikelompokkan berdasarkan nilai rata-rata dan nilai median. Nilai *information gain* tertinggi diperoleh dari atribut IP Semester 8 sebesar 4,08028. Pengambilan keputusan setiap cabang dilihat berdasarkan nilai entropy. Jika nilai entropy pada cabang bernilai 0, maka cabang tersebut akan diberikan keputusan.

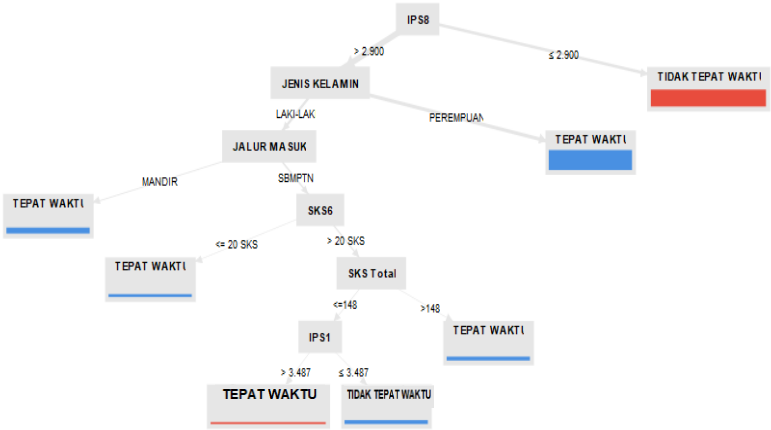


Gambar 4.6 *Node* Akar ke-1 yang dihasilkan dengan Algoritma C5.0

Dari pohon keputusan pada Gambar 4.6 menunjukkan bahwa dalam menentukan keputusan tiap cabang dilihat dari nilai entropy. Pada atribut IP Semester 8 terdapat cabang yang bernilai 0 yaitu kategori IP Semester <= 3,20. Sedangkan untuk kategori IP semester 8 > 3,20 akan dilanjutkan perhitungan *node* akar 1.1. Perhitungan nilai *information gain* akan berhenti jika tidak terdapat lagi atribut yang dapat dipecah menjadi lebih kecil.

4. Mengulangi Langkah ke-2. Kelas dibagi ke dalam cabang dan jika terdapat cabang yang memiliki dua kelas maka yang terpilih adalah kelas yang terbanyak dan proses diulang untuk masing-masing cabang hingga semua kelas pada cabang memiliki kelasnya masing-masing.

Dengan berbantuan *software RapidMiner* , diperoleh model pohon keputusan dari algoritma C5.0 pada Gambar 4.7.



Gambar 4.7 Model Pohon Keputusan

Dari Gambar 4.7 didapatkan kesimpulan :

a. Pohon keputusan yang diperoleh terdiri atas 6 atribut yang mempengaruhi yaitu IP Semester 1, Jumlah SKS total yang diambil, Jumlah SKS yang diambil pada semester 6, jalur masuk / seleksi yang dipilih saat pertama mendaftar, jenis kelamin dan IP Semester 8. Beberapa atribut seperti jumlah sks di semester 1-5 dan semester 7-8, IP Kumulatif dan lainnya tidak terlihat dari pohon keputusan. *Node* akar pada pohon keputusan ini adalah IP Semester 8 karena memiliki *gain* tertinggi.
b. Apabila mahasiswa memiliki IP Semester 8 $>$ 2,90 dengan jenis kelamin laki-laki, selanjutnya jalur masuk / seleksi yang diikuti adalah SBMPTN, jumlah SKS pada semester 6 lebih dari 20 SKS dan IP Semester 1 $\leq$ 3,487 maka diprediksi mahasiswa tersebut tidak lulus tepat waktu.

Setelah pohon keputusan terbentuk, aturan yang berlaku pada data training direpresentasikan dalam bentuk *list*. Aturan tersebut dihasilkan dari lintasan pohon yang diambil dari setiap cabang terakhir. Aturan pada data training kemudian diujikan pada data testing. Pengujian ini bertujuan untuk mencocokan data testing dengan pola lulusan mahasiswa yang terbentuk menggunakan struktur program *decision* "*IF..Then*". Adapun aturan yang terbentuk berdasarkan pohon keputusan Gambar 4.7 adalah

```
IPS8 > 2.900
|   JENIS KELAMIN = LAKI-LAKI
|   |   JALUR MASUK = MANDIRI: TEPAT WAKTU {TEPAT
WAKTU=5, TIDAK TEPAT WAKTU=0}
|   |   JALUR MASUK = SBMPTN
|   |   |   SKS6 = <= 20 SKS: TEPAT WAKTU {TEPAT
WAKTU=2, TIDAK TEPAT WAKTU=0}
|   |   |   SKS6 = > 20 SKS
|   |   |   |   SKS Total = <=148
|   |   |   |   |   IPS1 <= 3.487: TIDAK TEPAT
WAKTU {TEPAT WAKTU=0, TIDAK TEPAT WAKTU=1}
|   |   |   |   |   IPS1 > 3.487: TEPAT WAKTU
{TEPAT WAKTU=3, TIDAK TEPAT WAKTU=0}
```

```
|   |   |   |   SKS Total = >148: TEPAT WAKTU
{TEPAT WAKTU=3, TIDAK TEPAT WAKTU=0}
|   JENIS KELAMIN = PEREMPUAN: TEPAT WAKTU {TEPAT
WAKTU=19, TIDAK TEPAT WAKTU=0}
IPS8 ≤ 2.900: TIDAK TEPAT WAKTU {TEPAT WAKTU=0,
TIDAK TEPAT WAKTU=16}
```

4.4 Pengujian dan Akurasi Data

Dalam proses mining Algoritma Naïve Bayes menggunakan data mahasiswa Angkatan tahun 2015 – 2016 didapatkan pola informasi dan pengetahuan baru yang digunakan untuk memprediksi pola lulusan mahasiswa Program Studi Matematika UIN Walisongo Semarang. Penelitian ini menggunakan data training dan data testing untuk mencari probabilitas setiap atribut dalam upaya mendapatkan pola informasi baru.

Setelah dilakukan analisis klasifikasi Naïve Bayes dan Algoritma C5.0 menggunakan *software RapidMiner* menggunakan 10 *Fold Cross validation* diperoleh akurasi seperti yang tersaji pada Tabel 4.24.

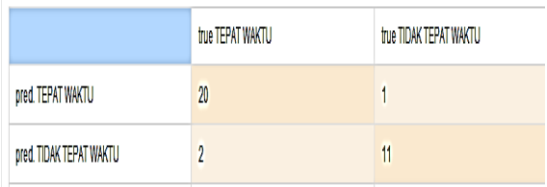
Tabel 4.24 Hasil Evaluasi Algoritma C5.0 dan Naïve Bayes menggunakan 10-*fold cross validation*

Algoritma	No	Spesifikasi Pengukuran	Nilai
Naïve Bayes	1	*Classification Error*	8,82%
	2	*Weighted mean recall*	91,29%
	3	*Weighted mean precision*	89,93%
	4	*Kappa atatistic*	0,810
	5	*Mean absoluter error*	0,078
	6	*Relative error*	7,81%

	7	*Root relative squared error*	38,2%
C5.0	1	*Classification Error*	5,88%
	2	*Weighted mean recall*	91,76%
	3	*Weighted mean precision*	95,83%
	4	*Kappa atatistic*	0,866
	5	*Mean absoluter error*	0,059
	6	*Relative error*	5,88%
	7	*Root relative squared error*	37,5%

Hasil evaluasi perbandingan antara dua algoritma menggunakan metode *Confussion Matrix. Confussion Matrix* digunakan untuk menganalisis kualitas *classifier* dalam mengenali tuple-tuple dari kelas yang ada. Metode ini menunjukkan ketepatan klasifikasi dengan prediksi menggunakan algoritma data mining yang dipilih. Hasil perhitungan *confussion matrix* disajikan pada tabel 4.25 sebagai berikut :

Tabel 4.25 Perhitungan *Confussion Matrix* Algoritma Naïve Bayes dan C5.0

Algoritma Naïve Bayes	
Algoritma C5.0	

Dari Tabel 4.25 dapat diperoleh nilai akurasi ketepatan klasifikasi untuk algoritma Naïve Bayes adalah

$$Akurasi = \frac{20+11}{20+1+2+11} \times 100\% = 91{,}18\%.$$

Sedangkan untuk algoritma C5.0 diperoleh nilai akurasi sebesar,

$$Akurasi = \frac{22+10}{22+2+0+10} \times 100\% = 94{,}12\% .$$

Tabel 4.26 Perbandingan Nilai Akurasi Kedua Algoritma

	Algoritma C5.0	Naïve Bayes
Nilai Akurasi	94,12%	91,18%

Tabel 4.26 menunjukkan perbandingan nilai Akurasi yang didapatkan dari hasil komputasi *software RapidMiner*. Terlihat bahwa nilai akurasi tertinggi pada algoritma C5.0. Hal ini menunjukkan bahwa algoritma C5.0 memberikan model klasifikasi yang lebih baik dibandingkan algoritma Naïve Bayes.

BAB V

PENUTUP

5.1 Kesimpulan

Kesimpulan yang didapatkan dari penelitian ini antara lain :

a. Model algoritma klasifikasi C5.0 membentuk pohon keputusan (*decision tree*) untuk memprediksi pola lulusan mahasiswa jurusan matematika UIN Walisongo Semarang. Pohon keputusan yang terbentuk berdasarkan pada perhitungan nilai *entropy* dan *Information Gain* dari 21 atribut yang digunakan dalam penelitian. Pada model algoritma C5.0, atribut yang paling mempengaruhi pola lulusan mahasiswa yaitu nilai IP Semester 1, Jumlah SKS total yang diambil, Jumlah SKS yang diambil pada semester 6, jalur masuk / seleksi yang dipilih saat pertama mendaftar, jenis kelamin dan IP Semester 8.
b. Model algoritma klasifikasi Naïve Bayes menggunakan nilai probabilitas $P(C_i|X)$ untuk memprediksi pola lulusan mahasiswa jurusan matematika UIN Walisongo Semarang yang didefinisikan sebagai

$$P(C_i|X) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i).P(x_2|C_i).\ldots P(x_n|C_i)$$

Algoritma Naïve Bayes mampu menampilkan informasi prediksi pola lulusan mahasiswa dengan menggunakan data mahasiswa yang dinyatakan lulus. Klasifikasi pola lulusan mahasiswa diperoleh dengan memaksimalkan probabilitas $P(C_i|X)$ untuk setiap kelas pada atribut yang digunakan untuk menentukan kelas baru.

c. *Output* dari penelitian ini adalah pola lulusan mahasiswa yang di prediksi menggunakan teknik data mining Algoritma C5.0 dan Algoritma Naïve Bayes dengan menggunakan *software RapidMiner*. Dari hasil

perhitungan diperoleh bahwa algoritma C5.0 memberikan hasil yang lebih baik daripada Algoritma Naïve Bayes yang ditunjukkan berdasarkan nilai Akurasi sebesar 94,12% dibandingkan hasil yang didapatkan algoritma Naïve Bayes yaitu nilai akurasi 91,18%. Data mining dengan algoritma C5.0 dan Naïve Bayes dapat diimplementasikan untuk memprediksi pola lulusan mahasiswa Program Studi Matematika UIN Walisongo Semarang dengan kategori lulus tepat waktu dan tidak lulus tepat waktu. Secara keseluruhan, variabel yang paling mempengaruhi hasil prediksi pola lulusan adalah nilai IP Semester 1, Jumlah SKS total yang diambil, Jumlah SKS yang diambil pada semester 6, jalur masuk / seleksi yang dipilih saat pertama mendaftar, jenis kelamin dan IP Semester 8.

5.2 Saran

Berdasarkan hasil penelitian yang telah disimpulkan diatas, dalam memprediksi pola lulusan mahasiswa dapat digunakan metode klasifikasi lain untuk mendapatkan metode klasifikasi terbaik. Selain itu, pada penelitian ini terdapat keterbatasan atribut yang digunakan. Penelitian ini masih dapat dilanjutkan dengan menambahkan lebih banyak atribut pengujian dan menggunakan data yang lebih banyak agar hasil prediksi lebih real dan mendekati hasil yang sebenarnya.

DAFTAR PUSTAKA

Amelia, Mongan Winny, dkk. (2017). *"Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes".* Jurnal Teknik Informarika Vol 11 No 1 ISSN : 2301-8364.

Ashraf, Mudasir et all.(2019). "*An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering Approaches*" Science Direct Procedia Computer Science 167 (1471 – 1483). https://doi.org/10.1016/j.procs.2020.03.358

David Hartanto Kamagi, Seng Hansun. (2014). *"Implementasi Data mining dengan Algoritma C4.5 untuk memprediksi Tingkat Kelulusan Mahasiswa"* .Program Studi, Teknik Informatika, Universitas Multimedia Nusantara.

Farida, Ida dan Spits Warnars H.L.H. (2019). "*Prediksi Pola Kelulusan Mahasiswa Menggunakan Teknik Data Mining Classification Emerging Pattern".* Jurnal Petir Vol 12 No 1 P-ISSN : 1978-9262 E-ISSN : 2655-5018.

Gorunescu, Florin. (2011). *"Data Mining Concept ,Model Technique".*Springer-Verlag Berlin Heidelberg.

Guo,Zizheng et all.(2021). "*Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management"* Jurnal Geoscience Frontiers 12 https://doi.org/10.1016/j.gsf.2021.101249.

Han and Kamber. (2012). *"Data Mining Concepts and Techniques Third Edition"*. Elsevier and Morgan Kaufmann (Vol.1). https://doi.org/10.1017/CBO9781107415324.004

Kusrini, dan Emha Taufik Luthfi. (2009). *"Algoritma Data mining"*. Yogyakarta , Penerbit Andi.

Larose, and Daniel T. (2005). "*Discovering knowledge in data: an introduction to data mining"*. USA: John Wiley and Sons.

MacLennan, J., Zhao Hui Tang, Bog, Crivat, (2009). *"Data Mining with Microsoft® SQL Server® 2008"*, Wiley Publishing, Inc., Indianapolis, Indiana.

Nuqson, Masykur Huda. (2010). *"Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (Studi Kasus Di Fakultas Mipa (Universitas Diponegoro)"*. Program Studi, Teknik Informatika, Fakultas Matematika dan Pengetahuan Alam, and Universitas Diponegoro

Prasetyo, Eko. (2012). *"Data Mining Konsep dan Aplikasi Menggunakan MATLAB"*. Yogyakarta:Andi.

Priati. (2016). *"Kajian Perbandingan Teknik Klasifikasi Algoritma C4.5, Naïve Bayes Dan CART Untuk Prediksi Kelulusan Mahasiswa (Studi Kasus : STMIK Rosma Karawang)".*Media Informatika. 15, 1-17.

Pratiwi, Reni. (2019). "*Perbandingan Klasifikasi Algoritma C5.0 dan Classification and Regression Tree".* Program Studi Statistika

Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Mulawarman Samarinda.

Santosa, Budi. (2007). *"Data mining Teknik Pemanfaatan Data untuk keperluan Bisnis"*. Graha Ilmu. Yogyakarta.

Sillueta, C.Y. (2016). *"Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Dengan Metode Klasifikasi Dan Algoritma Knearest Neighbor Berbasis Desktop (Studi Kasus : Fakultas Teknologi Informasi, Program Studi Teknik Informatika",* Tugas Akhir.

Suyatno. (2017*). "Data Mining Untuk Klasifikasi dan Klasterisasi Data".* Bandung: Informatika.

Turban, Efraim, Sharda, R. dan Delen, D. (2011). *"Decision Support Systems and Business Intelligent Systems".* Pearson

UIN Walisongo Semarang. (2015). *"Panduan Pengembangan Kurikulum UIN Walisongo Semarang Mengacu pada KKNI dan SN-Dikti"*. Kementerian Agama Republik Indonesia : UIN Walisongo Semarang.

Witten, I. H and Frank, E. (2005). "*Data Mining : Practical Machine Learning Tools and Techniques Second Edition"*. San Francisco: Elsevier.

Lampiran 1

BIODATA PENELITI

Peneliti I

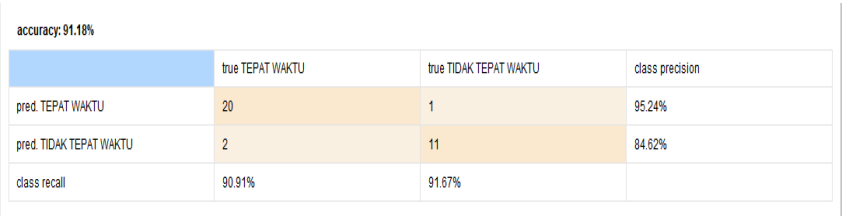
Nama Lengkap dan Gelar	: Ariska Kurnia Rachmawati,M.Sc
NIDN	: 2008118901
Fakultas/Program Studi	: Sains dan Teknologi/ Pendidikan Matematika
Perguruan Tinggi	: UIN Walisongo Semarang
Bidang Keahlian	: Matematika Komputasi
Alamat tinggal	: Jl. Kemantren RT 1 RW 4 Kelurahan Wonosari Kecamatan Ngaliyan, Semarang

Peneliti II

Nama Lengkap dan Gelar	: Riska Ayu Ardani,M.Pd
NIDN	: 2026079302
Fakultas/Program Studi	: Sains dan Teknologi/ Pendidikan Matematika
Perguruan Tinggi	: UIN Walisongo Semarang
Bidang Keahlian	: Pembelajaran Matematika
Alamat tinggal	: Mranggen, Kabupaten Demak, Jawa Tengah

Lampiran II

Hasil prediksi algoritma Naïve Bayes menggunakan software RapidMiner



accuracy: 91.18%

	true TEPAT WAKTU	true TIDAK TEPAT WAKTU	class precision
pred. TEPAT WAKTU	20	1	95.24%
pred. TIDAK TEPAT WAKTU	2	11	84.62%
class recall	90.91%	91.67%	

PerformanceVector

```
PerformanceVector:
accuracy: 91.18%
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    20      1
TIDAK TEPAT WAKTU:      2       11
classification_error: 8.82%
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    20      1
TIDAK TEPAT WAKTU:      2       11
kappa: 0.810
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    20      1
TIDAK TEPAT WAKTU:      2       11
weighted_mean_recall: 91.29%, weights: 1, 1
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    20      1
TIDAK TEPAT WAKTU:      2       11
weighted_mean_precision: 89.93%, weights: 1, 1
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
```

```
TEPAT WAKTU:    20      1
TIDAK TEPAT WAKTU:      2       11
spearman_rho: 0.812
kendall_tau: 0.812
absolute_error: 0.078 +/- 0.234
relative_error: 7.81% +/- 23.44%
relative_error_lenient: 7.81% +/- 23.44%
relative_error_strict: ∞%
normalized_absolute_error: 0.121
root_mean_squared_error: 0.247 +/- 0.000
root_relative_squared_error: 0.382
squared_error: 0.061 +/- 0.204
correlation: 0.812
squared_correlation: 0.659
cross-entropy: 0.726
```

Lampiran III

Hasil prediksi algoritma C5.0 menggunakan software RapidMiner :

accuracy: 94.12%

	true TEPAT WAKTU	true TIDAK TEPAT WAKTU	class precision
pred. TEPAT WAKTU	22	2	91.67%
pred. TIDAK TEPAT WAKTU	0	10	100.00%
class recall	100.00%	83.33%	

PerformanceVector

```
PerformanceVector:
accuracy: 94.12%
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    22      2
TIDAK TEPAT WAKTU:      0       10
classification_error: 5.88%
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    22      2
TIDAK TEPAT WAKTU:      0       10
kappa: 0.866
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    22      2
TIDAK TEPAT WAKTU:      0       10
weighted_mean_recall: 91.67%, weights: 1, 1
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    22      2
TIDAK TEPAT WAKTU:      0       10
weighted_mean_precision: 95.83%, weights: 1, 1
ConfusionMatrix:
True:   TEPAT WAKTU     TIDAK TEPAT WAKTU
TEPAT WAKTU:    22      2
```

```
TIDAK TEPAT WAKTU:     0       10
spearman_rho: 0.874
kendall_tau: 0.874
absolute_error: 0.059 +/- 0.235
relative_error: 5.88% +/- 23.53%
relative_error_lenient: 5.88% +/- 23.53%
relative_error_strict: ∞%
normalized_absolute_error: 0.091
root_mean_squared_error: 0.243 +/- 0.000
root_relative_squared_error: 0.375
squared_error: 0.059 +/- 0.235
correlation: 0.874
squared_correlation: 0.764
cross-entropy: ∞
```