TESIS - PM 147501

PERANCANGAN SISTEM PERINGATAN DINI DROPOUT DI MMT ITS MENGGUNAKAN METODE KLASIFIKASI NAIVE BAYES

Maks Agustinus
9116 205335

DOSEN PEMBIMBING
Dr.Tech, Ir. R. V. Hari Ginardi, MSc

DEPARTEMEN MANAJEMEN TEKNOLOGI
BIDANG KEAHLIAN MANAJEMEN TEKNOLOGI INFORMASI
FAKULTAS BISNIS DAN MANAJEMEN TEKNOLOGI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

LEMBAR PENGESAHAN

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Manajemen Teknologi (M.MT)
di
Institut Teknologi Sepuluh Nopember

Oleh:
MAKS AGUSTINUS
NRP. 9116205335

Tanggal Ujian : 3 Januari 2018
Periode Wisuda : Maret 2018

Disetujui oleh:

1. **Dr.Tech, Ir. R. V. Hari Ginardi, M.Sc** **(Pembimbing)**
   NIP. 196505181992031003

2. **Dr.Eng. Febriliyan Samopa, S.Kom, M.Kom** **(Penguji)**
   NIP. 197302191998021001

3. **Faizal Mahananto., S.Kom, M.Eng, Phd** **(Penguji)**
   NIP. 5200201301010

Dekan Fakultas Bisnis dan Manajemen Teknologi,

Prof. Dr. Ir. Udisubakti Ciptomulyono, M.Eng.Sc
NIP. 19590318 198701 1 001

(Halaman Sengaja Dikosongkan)

Perancangan Sistem Peringatan Dini Drop Out di MMT ITS Menggunakan Metode Klasifikasi Naive Bayes

Nama	: Maks Agustinus
NRP	: 9116205335
Pembimbing	: Dr.Tech, Ir. R. V. Hari Ginardi, MSc

ABSTRAK

Kejadian *dropout* bukan hanya merugikan bagi mahasiswa yang mengalaminya, tetapi terutama bagi negara dan institusinya. Karena singkatnya masa studi normal, kejadian *dropout* seolah tiba-tiba. Sehingga dibutuhkan sistem peringatan dini yang mampu memberikan peringatan sedini mungkin guna keberhasilan tindakan pencegahan. Metode-metode klasifikasi dalam data mining, menjadi alternatif solusi dalam proses klasifikasi drop out pada sistem peringatan dini ini.

Dari beberapa metode klasifikasi yang baik menurut beberapa penelitian terbaru, Naive Bayes dipilih sebagai alternatif. Metode klasifikasi Naive Bayes digunakan untuk menghasilkan model dalam sebuah sistem peringatan dini *dropout* untuk kemudian digunakan dalam penentuan klasifikasi *dropout* seorang mahasiswa. Output sistem peringatan dini ini adalah daftar mahasiswa yang terklasifikasi *dropout*. Daftar ini dapat digunakan oleh akademik untuk tindakan pencegahan.

Proses krusial pembuatan sistem peringatan dini ini ada pada pemilihan variabel prediktor yang cocok digunakan dalam metode klasifikasi Naive Bayes. Rancangan yang sesuai dibutuhkan untuk mewujudkan implementasi metode klasifikasi Naive Bayes dalam sistem peringatan dini drop out.

***Kata Kunci*: Sistem Peringatan Dini, *Dropout, Data Mining*, Naive Bayes**

Design of Dropout Early Warning System at MMT ITS Using Naive Bayes as A Classification Method

Name	: Maks Agustinus
Student ID	: 9116205335
Supervisor	: Dr.Tech, Ir. R. V. Hari Ginardi, MSc

ABSTRACT

Impact of dropout event is not only affect the student, but also affect the country and the institution. Because of short of study time at magister program, dropout event looks appear suddently. In this situation, early warning system is needed here so that it gives an alert as early as possible for prevention action. Data mining classification methods become alternative solution in dropout classification process.

From some good classification methods from some newest research, Naive Bayes is chosen as an alternative method that will be implemented. Naive Bayes is used to generate a classification model for drop out early warning system. The output of this system will be a list of student that are classified as dropout. This list may be used by academic unit for prevention.

Crucial process in building a dropout early warning system is in choosing the suitable predictor variables for Naive Bayes method. Proper design is needed to implement Naive Bayes classification method in a dropout early warning system

***Keywords*: Early Warning System, Dropout, Data Mining, Naive Bayes**

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa, karena berkat rahmat-Nya, penulis dapat menyelesaikan tesis ini dengan tepat waktu. Tesis ini berjudul "Perancangan Sistem Peringatan Dini Drop Out di MMT ITS Menggunakan Naive Bayes".

Dalam penulisan proposal tesis ini penulis mendapatkan banyak bantuan dari berbagai pihak baik secara moral maupun materi. Dalam kesempatan ini penulis ingin mengucapkan terima kasih kepada :

1. Istri saya tercinta, Esty Andar Mutiasari, dan anak saya terkasih, Chrissoniya Zefanya Maesy, sebagai penyemangat yang telah mendukung dan mendoakan saya dalam menulis tesis ini.
2. Bapak Dr.Tech, Ir. R. V. Hari Ginardi, MSc, selaku dosen pembimbing dan dosen wali yang telah memberikan banyak bimbingan, motivasi dan nasehat kepada penulis.
3. Bapak Dr.Eng. Febriliyan Samopa, S.Kom, M.Kom, selaku Direktur DPTSI ITS dan Bapak Radityo Prasetianto Wibowo, S.Kom, M.Kom, selaku Kabid Data DPTSI ITS, yang telah membantu dalam pengumpulan data awal untuk tesis ini.
4. Bapak Mudji Syukur, S.Kom, Selaku Kasubbag Pemantauan dan Evaluasi Pembelajaran BAPKM ITS, yang telah membantu melengkapi data awal untuk tesis ini.
5. Segenap staf administrasi MMT ITS, khususnya Mas Reval yang dengan sabar membantu penulis baik dalam penyelesaian kuliah maupun tesis.
6. Seluruh civitas akademik Magister Manajemen Teknologi Institut Teknologi Sepuluh Nopember Surabaya, yang telah membantu penulis dalam menyelesaikan studi.
7. Teman-teman MTI MMT ITS angkatan 2016 yang banyak memberi motivasi dan bantuan kepada penulis, khususnya Gilvy sebagai kawan seperjuangan baik dalam sidang proposal dan sidang tesis.

Penulis menyadari bahwa dalam penyusunan tesis ini masih banyak kekurangan baik format laporan maupun isinya. Untuk itu penulis sangat

mengharapkan kritik dan saran yang membangun. Semoga tesis ini dapat bermanfaat baik bagi pembaca maupun penulis, amin.

Surabaya, Januari 2018

Maks Agustinus

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

(Halaman Sengaja Dikosongkan)

Bab I

Pendahuluan

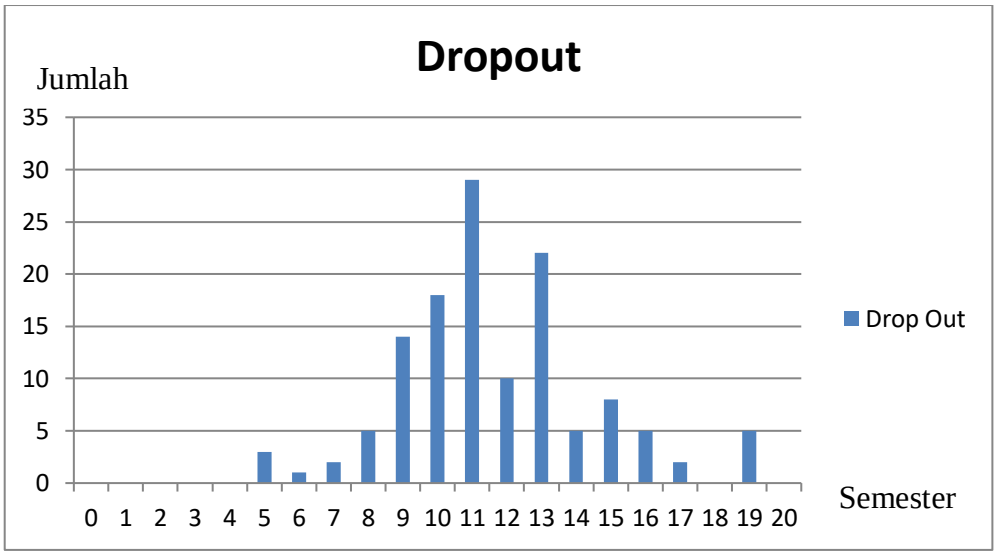
Pada bab 1 ini dijelaskan tentang pendahuluan yang memiliki kaitan dengan penelitian yang terdiri dari latar belakang, perumusan masalah, batasan masalah, tujuan penelitian dan manfaat penelitian.

1.1 Latar Belakang

Kebutuhan akan adanya sistem peringatan dini untuk berbagai macam keperluan semakin dirasa perlu pada awal milenium ini di Indonesia. Kebutuhan ini semakin terasa urgensinya di Indonesia sejak terjadinya tsunami dahsyat yang menimpa pesisir Aceh pada akhir tahun 2006 .

Salah satu keperluan akan sistem ini adalah peringatan dini akademik khususnya untuk menghasilkan daftar siapa saja mahasiswa yang berpotensi akan mengalami dropout. Dengan adanya daftar tersebut, jurusan dapat memberi perhatian lebih kepada mahasiswa yang membutuhkan. Pada akhirnya, universitas dapat mengoptimalkan anggaran serta nilai akreditasi terkait hal ini.

Sebagaimana jurusan pada universitas lain, MMT ITS pun perlu mewaspadai tingkat dropout yang terjadi sedemikian sehingga sedapat mungkin memperkecil tingkat dropout pada tahun-tahun mendatang. Berikut ini sebaran dropout yang terjadi di MMT dari pertama kali dibuka sampai tahun ajaran 2016/2017.



Gambar 1.1 Sebaran dropout di tiap semester.

Selain drop out, ada kondisi lain yang artinya sama yaitu bahwa mahasiswa yang bersangkutan tidak berhasil menyelesaikan studi. Kondisi tersebut antara lain "Mengundurkan Diri", "Non Aktif" dan lain-lain. Untuk kondisi "Drop Out" adalah jelas bahwa mahasiswa yang bersangkutan telah memenuhi kriteria drop out yang telah ditetapkan oleh bagian akademik ITS. Untuk kondisi "Mengundurkan Diri" adalah berbeda dengan kondisi "Dropout", yaitu mahasiswa yang bersangkutan yang memutuskan berhenti studi. Untuk kondisi "Non Aktif" dan lain-lain, tidak terlalu jelas dan homogen mengenai alasan mengapa mahasiswa yang bersangkutan ditetapkan dalam kondisi tersebut.

Banyak metode telah diterapkan untuk menghasilkan sistem peringatan dini yang memadai. Teknik yang umum digunakan dalam sistem ini antara lain mulai dari yang perhitungan sederhana sampai statistik lanjut seperti forecasting dan lain-lain. Teknik lain yang digunakan sebagai alternatif yang mulai disukai adalah data mining.

Data mining adalah teknik penambangan data untuk menemukan suatu pola/model dari data yang berjumlah banyak (Berry). Semakin banyak data yang digunakan akan semakin baik pola/model yang ditemukan. Ukuran baik tidaknya model yang dihasilkan akan diukur berdasarkan tingkat akurasi, tingkat sensitifitas serta tingkat spesifisitas.

Dalam (Halim, 2015), Algoritma C4.5 (Decision Tree) digunakan sebagai metode klasifikasi untuk mengidentifikasi mahasiswa MMT ITS yang lulus tidak tepat waktu. Kejadian lulus tidak tepat waktu dengan kejadian drop out adalah dua kejadian yang sama-sama tidak sering terjadi. Sehingga metode tersebut sudah mestinya cocok untuk data drop out.

Namun dalam (Knowles, 2015), pada penelitian sistem peringatan dini drop out, dihasilkan kesimpulan untuk tidak bergantung pada satu metode agar mendapatkan tingkat prediksi yang lebih baik. Di tambah lagi, dalam (Phyu, 2009), disimpulkan bahwa Naive Bayes dan Decision Tree saling mengisi. Artinya, ketika Naive Bayes memiliki tingkat prediksi yang bagus pada suatu kasus analisis, maka Decision Tree memiliki tingkat prediksi yang kurang bagus, demikian pula sebaliknya. Dan yang semakin menguatkan jatuhnya pilihan pada Naive Bayes untuk diterapkan pada sistem peringatan dini drop out adalah hasil penelitian yang menyatakan bahwa Naive Bayes memiliki performa lebih baik dari Decision Tree dan Neural Networks (Xhemali, 2009).

Pada penelitian ini akan digunakan metode klasifikasi Naive Bayes pada sistem peringatan dini drop out di MMT ITS. Untuk keperluan klasifikasi, digunakan 2 indikator yaitu dropout dan lulus.

1.2 Perumusan Masalah

Untuk membangun sistem peringatan dini drop out, berikut ini permasalahan yang akan dibahas :

1. Variabel apa saja yang tersedia yang dapat digunakan sebagai variabel prediktor pada input data pada sistem peringatan dini drop out ?
2. Seberapa efektifkah penggunaan metode klasifikasi Naive Bayes untuk sistem peringatan dini drop out berdasarkan ukuran tingkat akurasi, tingkat sensitifitas dan tingkat spesifisitas?
3. Seperti apakah kebutuhan dan rancangan sistem peringatan dini drop out di program Magister Managemen Teknologi ITS ?

1.3 Batasan Masalah

Berikut ini batasan yang ditetapkan dalam penelitian ini :

1. Data yang digunakan dalam penelitian ini adalah data mahasiswa MMT ITS mulai dari angkatan awal sampai angkatan 2016.
2. Pengolahan dan simulasi menggunakan software Weka.
3. Dalam proses perancangan menggunakan use case dan mockup untuk menggambarkan sistem yang ingin dibangun.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah untuk membuat rancangan sistem peringatan dini drop out menggunakan metode klasifikasi Naive Bayes. Di dalam pengerjaannya, hal-hal berikut ini diharapkan terlaksana :

1. Menemukan variabel yang tersedia di data akademik MMT ITS yang dapat digunakan sebagai variabel prediktor.

2. Menemukan efektifitas metode klasifikasi Naive Bayes pada sistem seringatan dini drop out di MMT ITS.
3. Memaparkan kebutuhan dan rancangan sistem untuk sistem peringatan dini drop out di MMT ITS.

1.5 Manfaat Penelitian

Bagi mahasiswa yang masuk ke dalam kategori klasifikasi rawan drop out, kemungkinan besar akan drop out bila tidak dilakukan intervensi terhadap mereka. Dengan adanya informasi tersebut, jurusan dapat melakukan intervensi sedemikian sehingga sedapat mungkin mereka terhindar dari drop out.

Bagi jurusan dan universitas, output dari sistem peringatan dini drop out dapat dimanfaatkan untuk optimalisasi anggaran dan nilai akreditasi jurusan maupun institusi terkait tingkat dropout, terutama dalam pengambilan keputusan penentuan mahasiswa yang masuk dalam klasifikasi rawan dropout untuk kemudian dilakukan penanganan yang sesuai.

Manfaat lain dari penelitian ini adalah bertambahnya khasana sistem peringatan dini untuk mendeteksi terduga drop out sedini mungkin. Dari sederetan sistem peringatan dini yang ada, sistem peringatan dini berbasis *data mining* dapat menjadi salah satu pilihan terbaik.

1.6 Sistematika Penulisan

Sistematika penulisan yang digunakan dalam penelitian ini adalah seperti berikut ini :

BAB 1 membahas tentang pendahuluan terkait penelitian yang terdiri dari latar belakang, perumusan masalah, batasan masalah, tujuan penelitian dan manfaat penelitian.

BAB 2 membahas mengenai kajian pustaka dan dasar teori yang berkaitan mengenai penelitian ini. Kajian pustaka dan dasar teori berfungsi sebagai sumber untuk memahami permasalahan dan menyelesaikan permasalahan yang berkaitan dengan penelitian.

BAB 3 membahas mengenai proses-proses atau tahapan-tahapan penelitian yang digunakan untuk mencapai tujuan dalam penelitian.

BAB 4 membahas mengenai pengerjaan penelitian sebagaimana yang telah ditetapkan dalam metodologi penelitian.

BAB 5 memberikan kesimpulan dari pembahasan yang telah dilakukan dalam penelitian ini.

Bab 2

Kajian Pustaka dan Dasar Teori

Bab 2 menjelaskan mengenai kajian pustaka dan dasar teori yang berkaitan mengenai penelitian ini. Kajian pustaka dan dasar teori berfungsi sebagai sumber untuk memahami permasalahan dan menyelesaikan permasalahan yang berkaitan dengan penelitian.

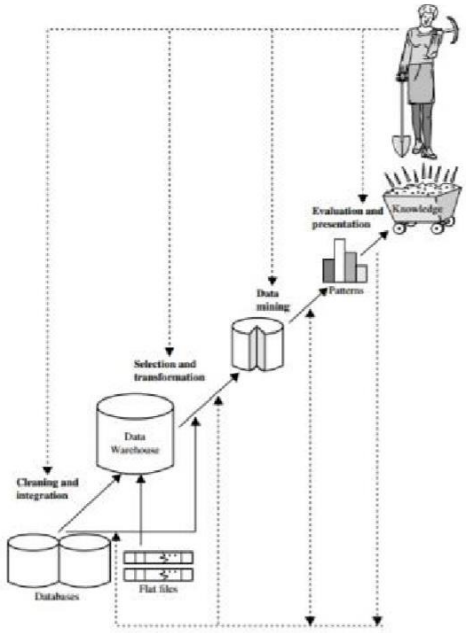
2.1 Penambangan Data (Data Mining)

Data mining merupakan proses penentuan pola dan informasi dari data yang mempunyai jumlah besar. Sementara sumber data yang dapat digunakan dapat berupa database, data warehouse, web, tempat penyimpanan informasi lainnya ataupun data yang mengalir ke dalam sistem yang dinamis (Han, et al, 2012: 8).

Dari apa yang diutarakan oleh Grup Gartner (dalam Larose, 2005: 2) data mining adalah suatu proses menemukan hubungan yang memiliki arti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dengan menggunakan teknik pengenalan pola baik berdasarkan statistik ataupun matematika.

Berdasarkan Turban, dkk (dalam Kusrini & Emha, 2009: 3) Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning guna mengeluarkan sari dan mengindentifikasi informasi yang berguna dan pengetahuan yang terkait dari berbagai database besar.

Data mining merupakan salah satu langkah penting dalam penemuan sebuah pengetahuan pada proses *Knowledge Discovery in Data* (KDD). KDD merupakan proses menentukan informasi yang bermanfaat serta pola-pola yang ada dalam data. Tahapan proses KDD ditunjukkan oleh Gambar 2.1.



Gambar 2.1 Tahap-tahap Knowledge Discovery in Data

(Han, et al, 2006: 6)

Menurut Han, et al (2006: 7) tahapan dalam KDD dapat dijelaskan sebagai berikut :

1. Pembersihan Data (Data Cleaning)

Pembersihan data merupakan proses penghilangan noise dan data yang tidak konsisten. Pada tahap ini data-data yang memiliki isian tidak sempurna seperti data yang tidak memiliki kelengkapan

atribut yang dibutuhkan dan data yang tidak valid dihapus dari database.

2. Integrasi Data (Data Integration)

Integrasi data merupakan proses penggabungan beberapa sumber data ke dalam database. Pada tahap ini dilakukan penggabungan data dari berbagai sumber untuk dibentuk menjadi penyimpanan data yang koheren.

3. Seleksi Data (Data Selection)

Seleksi data merupakan pemilihan data yang digunakan dalam proses data mining. Data hasil seleksi yang akan dipakaikan untuk proses data mining, disimpan dalam suatu berkas dan terpisah dari basis data operasional.

4. Transformasi Data (Data Transformation)

Transformasi data merupakan proses mentransformasikan dan mengkonsolidasikan data yang dipakaikan pada proses mining. Pada tahap ini dilakukan pengubahan format data ke dalam format yang sesuai dengan teknik data mining yang digunakan.

5. Penambangan Data (Data Mining)

Penambangan data merupakan proses utama mencari pengetahuan dari informasi tersembunyi. Penambangan data adalah proses mencari pola atau informasi yang menarik dalam data terpilih dengan menerapkan teknik atau metode tertentu. Teknik dalam data mining sangat banyak dan bervariasi, pemilihan teknik yang cocok sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

6. Evaluasi Pola (Pattern Evaluation)

Evaluasi pola merupakan proses identifikasi kebenaran pola yang telah didapatkan. Pada tahap ini pola yang diperoleh dari proses data mining dievaluasi untuk melihat apakah pola yang ditemukan kontra atau sejalan dengan fakta atau hipotesis yang ada sebelumnya.

7. Representasi Pengetahuan (Knowledge Presentation)

Representasi pengetahuan adalah visualisasi dan penyajian pengetahuan yang telah diperoleh kepada *user*. Pada tahap terakhir ini disajikan pengetahuan dan metode yang digunakan guna memperoleh pengetahuan agar mampu dipahami oleh pengguna atau semua orang. Data mining memiliki beberapa metode yang dilakukan pengguna untuk meningkatkan proses mining supaya lebih efektif. Oleh karena itu, data mining dibagi menjadi beberapa kelompok berdasarkan metodenya, yaitu (Larose, 2005: 11) :

1. Deskripsi

Deskripsi dipakai untuk memberikan gambaran ringkas berupa pola dan tren bagi data yang jumlahnya sangat besar dan jenisnya beragam. Metode dalam data mining yang dapat dipakai untuk mendeskripsikan suatu data antara lain neural network dan exploratory data analysis.

2. Klasifikasi

Pada klasifikasi terdapat variabel target yang berupa nilai kategori/ indikator. Contoh klasifikasi adalah penggolongan

pendapatan masyarakat ke dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Algoritma klasifikasi yang biasa dipakai di antaranya adalah Naïve Bayes, K-Nearest Neighbor, dan C4.5.

3. Estimasi

Estimasi mirip dengan klasifikasi dengan perbedaan variabel target pada proses estimasi lebih condong ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi, kemudian nilai estimasi dari variabel target dibuat berdasarkan nilai prediksi. Contoh algoritma estimasi antara lain linear regression dan neural network.

4. Prediksi

Prediksi hampir mirip dengan klasifikasi dan estimasi, namun pada prediksi data yang digunakan merupakan data runtun waktu (data time series) dan nilai pada hasil akhir dipakai pada beberapa waktu mendatang.

Contoh prediksi dalam bisnis dan penelitian adalah prediksi harga saham dalam beberapa bulan ke depan.

5. Pengelompokan

Pengelompokan data atau penyusunan data ke dalam jenis yang sama. Pengelompokan tidak untuk mengklasifikasi, mengestimasi, atau memprediksi suatu nilai, tetapi membagi

seluruh data menjadi kelompok-kelompok yang relatif sama (homogen). Perbedaan algoritma pengelompokan dengan algoritma klasifikasi adalah pengelompokan tidak memiliki target/ class/ label/ indikator. Contoh algoritma untuk pengelompokan antara lain K Means dan Fuzzy C-Means.

6. Asosiasi

Asosiasi dipakai guna menemukan atribut yang muncul pada waktu yang bersamaan dan guna mencari hubungan antara dua atau lebih data pada sekumpulan data. Contoh penggunaan aturan asosiasi yaitu analisis kemungkinan seorang pelanggan membeli roti dan selai dalam waktu yang bersamaan di suatu pasar swalayan. Contoh algoritma aturan asosiasi yang sering digunakan antara lain Apriori dan FP-Growth.

2.2 Klasifikasi

Klasifikasi merupakan proses penemuan model yang mampu membedakan kelas data atau konsep yang bertujuan agar dapat digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Model ditemukan berdasarkan analisis data training (objek data dengan kelas yang sudah diketahui) (Han, et al, 2006: 24).

Algoritma yang biasa digunakan untuk proses klasifikasi sangat banyak, di antaranya adalah K-Nearest Neighbor, Rough Set, Algoritma Genetika, metode Rule Based, C4.5, Naive Bayes, analisis statistik, Memory Based Reasoning, dan Support Vector Machines (SVM).

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah fase *training/learning*, di mana algoritma klasifikasi dibuat untuk menganalisa data training lalu dinyatakankan dalam bentuk aturan klasifikasi/ model. Proses kedua adalah klasifikasi, di mana aturan klasifikasi/ model digunakan untuk mengklasifikasi data tes guna memperkirakan tingkat akurasinya (Han, et al, 2006: 286).

Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011: 15):

1. Kelas

   Variabel dependen berupa kategori yang merepresentasikan "label" yang terdapat pada objek. Contohnya: risiko dropout, risiko kredit, jenis gempa dan lain-lain.

2. Predictor

   Variabel independen yang diwakili penyataannya oleh karakteristik (atribut) data. Contohnya: merokok atau tidak, minum alkohol atau tidak, nilai ujian, jumlah tabungan, indeks prestasi, jumlah gaji.

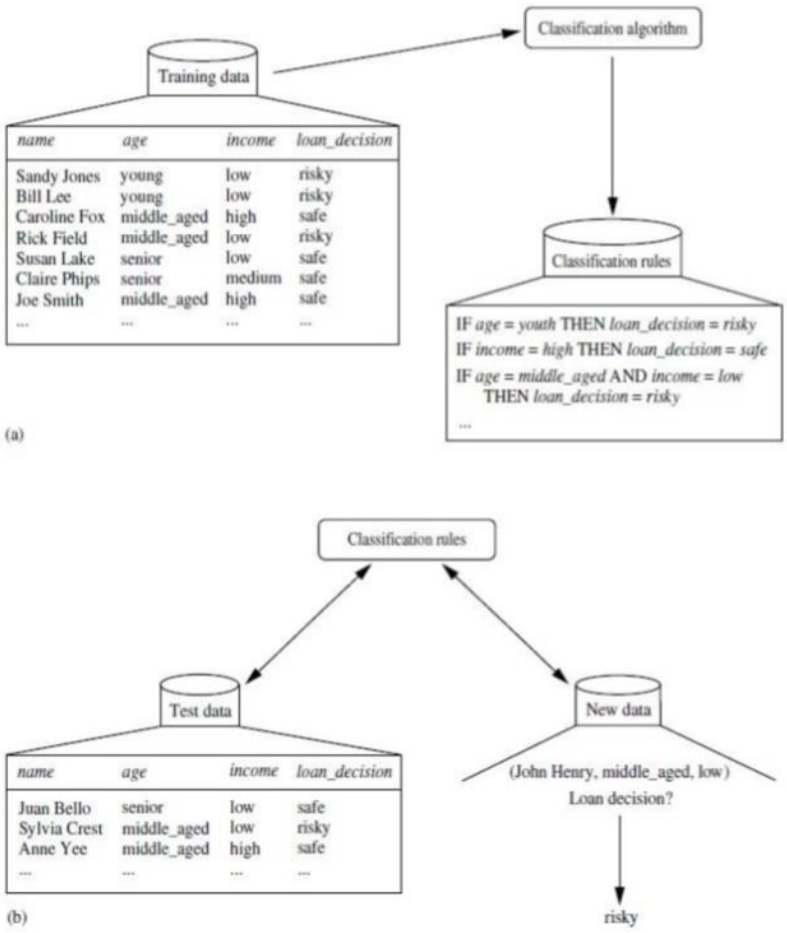
3. Dataset Latih

   Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan variabel predictor/ atribut.

4. Dataset Tes

Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi yang dihasilkan dievaluasi. Proses klasifikasi dapat dicontohkan seperti yang ditunjukkan oleh Gambar 2.2.

Gambar 2.2 poin (a) adalah proses pembelajaran dimana data latih dianalisis menggunakan algoritma klasifikasi. Atribut keputusan pinjaman sebagai label kelas, dan model pembelajaran atau pengklasifikasian dinyatakan dalam bentuk aturan klasifikasi (classification rule). Gambar 2.2 poin (b) merupakan proses klasifikasi. Proses klasifikasi dipakai guna mengestimasi tingkat akurasi dari classification rule yang dihasilkan. Apabila tingkat akurasi dapat diterima maka aturan yang dihasilkan dapat dipakai pada klasifikasi data baru (Han, et al, 2006: 287).



Gambar 2.2 Proses Klasifikasi (Han, et al, 2006: 287)

2.3 Naive Bayes

Sebelum masuk ke dalam pembahasan tentang Naive Bayes, diperlukan pengetahuan terkait peluang bersyarat. Peluang bersyarat merupakan peluang bahwa kejadian X terjadi bila diketahui bahwa suatu kejadian H telah terjadi. Peluang bersyarat dilambangkan dengan $(X|H)$ dibaca "peluang X bila H terjadi". Persamaan untuk peluang bersyarat sebagai berikut (Walpole, 1995: 97-98).

$(X|H) = (X \cap H)\ (H), \qquad jika\ (A) > 0 \qquad (2.1)$

Sama halnya dengan peluang terjadinya kejadian H bila diketahui bahwa

suatu kejadian X telah terjadi.

$$(H|X) = (X \cap H)\ (X), \quad jika\ (H) > 0 \qquad (2.2)$$

Dengan mengkombinasikan persamaan (2.1) dan (2.2) maka diperoleh,

$$P(H|X)P(X) = P(X \cap H) = P(X|H)P(H)$$

sehingga persamaan (2.2) menjadi:

$$(H|X) = P(X \cap H)\ P(X)$$

$$P(H|X) = P(X|H)P(H)\ P(X)$$

Teorema Bayes menebak peluang di masa yang akan datang menurut pengalaman di masa sebelumnya. Pada teorema Bayes, X dijabarkan oleh kumpulan n atribut dengan H adalah beberapa hipotesis, sehingga data X termasuk sebuah kelas C (Han, et al, 2012: 350). Persamaan teorema Bayes adalah

$$(H|X) = (X|H)(H)\ P(X) \qquad (2.3)$$

Keterangan :

X : Data dengan kelas yang belum diketahui

H : Hipotesis data X merupakan suatu kelas khusus

$(H|X)$: Probabilitas hipotesis H berdasarkan kondisi X (posterior probability)

(H) : Probabilitas hipotesis H (prior probability)

$(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H

(X) : Probabilitas X

Naïve Bayes merupakan pengklasifikasian statistik yang dapat dipakai untuk memperkirakan probabilitas keanggotaan suatu class. Bayes adalah teknik berbasis probabilistik sederhana yang didasarkan pada penerapan teorema Bayes dengan asumsi independensi yang kuat.

Klasifikasi Naïve Bayes yang mengacu pada teorema Bayes di atas mempunyai persamaan sebagai berikut :

$$(Ci|X) = (X|Ci)(Ci)\ P(X) \qquad (2.4)$$

Keterangan :

$(Ci|X)$: Probabilitas hipotesis Ci jika diberikan fakta atau record X (posterior probability)

$(X|Ci)$: Nilai parameter yang memberikan kemungkinan yang paling besar (likelihood)

(Ci) : Probabilitas kelas Ci (Prior probability)

(X) : Probabilitas X

Menurut Han, et al (2012: 351) proses dari pengklasifikasian Naïve Bayes adalah sebagai berikut:

a) Variabel D adalah kumpulan dari data dan label yang terkait dengan kelas. Setiap data diwakili oleh vektor atribut dengan n-dimensi, X = ($x1,2,\ldots,xn$) dengan n dibuat dari data n atribut, berturut-turut, $A1,A2,\ldots,An$.
b) Misalkan ada i kelas, $C1,2,\ldots,Ci$. Diberikan sebuah data yaitu X, lalu proses klasifikasi memperkirakan X ke dalam kelompok yang mempunyai probabilitas posterior tertinggi menurut kondisi X. Artinya,

klasifikasi Naïve Bayes memprediksi bahwa X tergolong kelas Ci jika dan hanya jika:

$$(Ci|X) > (Cj|X) \text{ untuk } 1 \leq j \leq m, \neq i \qquad (2.5)$$

Maka nilai $(Ci|X)$ harus lebih dari $(Cj|X)$ yaitu supaya diperoleh hasil akhir $(Ci|X)$.

c) Pada saat (X) konstan bagi semua kelas maka hanya $(X|Ci)(Ci)$ yang perlu dihitung. Jika probabilitas kelas prior sebelumnya belum diketahui, maka diasumsikan bahwa kelasnya sama, yaitu $(C1) = (C2) = \cdots = P(Cm)$, untuk menghitung $P(X|Ci)$ dan $P(X|Ci)P(Ci)$. Perhatikan bahwa probabilitas kelas prior dapat diperkirakan oleh persamaan berikut,

$$(Ci) = (|Ci,D|)\ |D| \qquad (2.6)$$

dimana $|Ci,D|$ adalah jumlah data training dari kelas Ci dan D adalah jumlah total data training yang digunakan.

d) Ketika diberikan kumpulan data yang mempunyai banyak atribut, maka perhitungan $P(X|Ci)$ dengan penjabaran lebih lanjut rumus Bayes tersebut yaitu menjabarkan $P(x1,...,xn|Ci)$ menggunakan aturan perkalian, menjadi sebagai berikut (Samuel Natalius: 2010):

$P(x1,\ldots,xn|Ci) = P(x1|Ci)P(x2,\ldots,xn|Ci,x1) = P(x1|Ci)P(x2|Ci,x1)P(x3,\ldots,xn|Ci,x1,x2)$

$P(x1,\ldots,xn|Ci) = P(x1|Ci)P(x2|Ci,x1)\ldots P(xn|Ci,x1,x2,\ldots,xn-1)$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan sangat kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang sangat mustahil untuk dianalisa

satu-persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Oleh karena itu digunakan asumsi independensi yang sangat kuat (naïve), bahwa masing-masing petunjuk ($x1,2,\ldots,xn$) saling bebas (independen) satu sama lain, maka berlaku suatu persamaan sebagai berikut (Samuel Natalius: 2010):

$$P(x_i|x_j) = \frac{P(x_i \cap x_j)}{P(x_j)} = \frac{P(x_i)P(x_j)}{P(x_j)} = P(x_i) \text{ untuk } i \neq j, \text{sehingga}$$

$$P(x_i|C, x_j) = P(x_i|C_i)$$

Dari persamaan di atas terlihat jelas bahwa asumsi independensi naïve tersebut menjadikan syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran ($x1,\ldots,|Ci$) dapat disederhanakan lagi menjadi seperti berikut:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \quad (2.7)$$

Perhitungan ($X|Ci$) di setiap atribut mengikuti jabaran berikut:

1) jika Ak adalah kategori, maka $P(xk|Ci)$ adalah jumlah data dari kelas Ci di D yang memiliki nilai xk untuk atribut Ak dibagi dengan $|Ci,D|$ yaitu jumlah data dari kelas Ci di D,

2) jika Ak adalah numerik, biasanya diasumsikan memiliki distribusi Gauss dengan rata-rata μ dan standar deviasi σ, didefinisikan oleh:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.8)$$

sehingga diperoleh:

$$(xk|Ci) = (xk,,Ci) \qquad (2.9)$$

Setelah itu akan dihitung μCi dan σCi yang merupakan rata-rata dan standar deviasi dari masing-masing nilai atribut Ak pada tupel training kelas Ci.

e) Untuk $(X|Ci)(Ci)$ dievaluasi pada setiap kelas Ci guna memperkirakan pengklasifikasian label kelas data X dengan menggunakan persamaan,

$$(X|Ci)(Ci) > (X|Cj)(Cj) \text{ untuk } 1 \leq j \leq m, j \neq i \qquad (2.10)$$

label kelas untuk data X yang diprediksi adalah kelas Ci jika nilai $P(X|Ci)P(Ci)$ lebih dari nilai $P(X|Cj)P(Cj)$.

2.4 Laplacian Correction

Untuk menghindari agar hasil probabilitas pada perhitungan Naïve Bayes tidak bernilai nol karena tidak adanya data untuk suatu kategori tertentu dari suatu variabel prediktor dalam kelasnya, dapat digunakan teknik estimasi yang biasa disebut Laplace Estimator atau Laplacian Correction (Han and Kamber, 2006). Dalam teknik ini digunakan penambahan nilai 1 pada data untuk masing-masing kategori ketika ada kategori yang memiliki nilai 0 (tidak ada).

Akan lebih mudah untuk memahami dengan contoh berikut. Misalkan variabel prediktor Nilai_Ujian setelah dikategorikan akan memiliki nilai "Baik", "Sedang" atau "Kurang". Kelas yang digunakan adalah Kuliah dan Kerja. Data latih yang digunakan berjumlah 10 data dengan perincian 4 data pada kelas Kuliah dan 6 data pada kelas Kerja. Bila Nilai_Ujian "Baik" dan "Sedang" di kelas Kuliah sama-sama ada 2 kejadian, maka probabilitas Nilai_Ujian "Baik" bernilai 2/4 dan probabilitas Nilai_Ujian "Sedang" bernilai 2/4 juga. Sehingga probabilitas Nilai_Ujian "Kurang" akan bernilai 0/4. Dengan Laplacian Correction, semua kategori kejadian ditambah 1 kejadian sehingga seolah-olah jumlah data pada kelas Kuliah bertambah 3

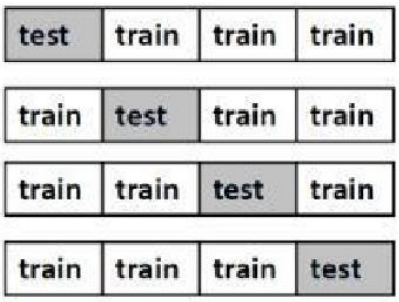
menjadi 7 data. Dengan demikian probabilitas Nilai_Ujian Kurang terhindar dari nilai 0. Sehingga probabilitas selengkapnya menurut contoh tersebut adalah sebagai berikut :

- Probabilitas Nilai_Ujian "Baik" menjadi 3/7
- Probabilitas Nilai_Ujian "Kurang" menjadi 3/7
- Probabilitas Nilai_Ujian "Sedang" menjadi 1/7

2.5 Pengujian dan Evaluasi Model

Model yang didapatkan dari metode klasifikasi Naive Bayes kemudian diuji menggunakan k-fold cross validation. Cross validation merupakan bentuk sederhana dari teknik statistik. Jumlah fold standar untuk memprediksi tingkat error pada data adalah dengan menggunakan 10-fold cross validation (Witten, et al, 2011: 153).

Data yang akan dipakai dibagi secara acak ke dalam k subset yaitu $D1,2,\ldots,Dk$ dengan masing-masing subset memiliki ukuran yang sama. Dataset dibagi menjadi data latih dan data tes/ uji. Proses latih dan uji dilakukan sebanyak k kali secara berulang-ulang. Pada iterasi ke-i, partisi Di digunakan sebagai data tes dan partisi sisanya secara bersamaan dan berurutan berperan sebagai data latih. Iterasi kedua, subset $D1,2,\ldots,Dk$ akan dites pada $D2$, dan selanjutnya hingga Dk (Han, et al, 2012: 364). Gambar 2.3 berikut adalah gambar untuk contoh ilustrasi 4-fold cross validation. Penggunaan K=4 untuk ilustrasi demi kemudahan penjelasan.



Gambar 2.3 Ilustrasi 4-Fold Cross Validation

Berdasarkan Gambar 2.3 ditunjukkan bahwa nilai fold yang digunakan adalah 4-fold cross validation. Berikut diberikan langkah-langkah pengujian data dengan 4-fold cross validation :

a. Dataset yang digunakan dibagi menjadi 4 bagian, yaitu $D1$,$D2$,$D3$,dan $D4$. Dt, = (1,2,3,4) digunakan sebagai data tes dan dataset lainnya sebagai data latih.
b. Tingkat akurasi dihitung pada setiap iterasi (iterasi-1, iterasi-2, iterasi-3, iterasi-4), berikutnya dihitung rata-rata tingkat akurasi dari seluruh iterasi untuk mendapatkan tingkat akurasi dari data keseluruhan.

Evaluasi hasil klasifikasi dilakukan dengan metode confusion matrix. Confusion matrix adalah tool yang digunakan sebagai alat evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas sebenarnya atau dengan kata lain berisi informasi nilai sebenarnya dan prediksi pada klasifikasi (Gorunescu, 2011: 319).

Tabel 2.1 Tabel Confusion Matrix Dua Kelas

Clasification	*Predicted class*	
	Class=Yes	*Class=No*
Class=Yes	a (*true positive*)	b (*false negative*)
Class=No	c (*false positive*)	d (*true negative*)

Pada tabel confusion matrix di atas, true positive (TP) merupakan jumlah record positif yang diklasifikasikan sebagai positif, false positive (FP) adalah jumlah record negatif yang diklasifikasikan sebagai positif, false negatives (FN) merupakan jumlah record positif yang diklasifikasikan sebagai negatif, true negatives (TN) adalah jumlah record negatif yang diklasifikasi sebagai negatif. Setelah data uji diklasifikasikan maka didapatkan confusion matrix sehingga dapat diperkirakan tingkat sensitifitas, spesifisitas, dan akurasi dari model yang telah dibuat.

Sensitifitas merupakan proporsi dari class=yes yang terprediksi dengan benar. Sedangkan Spesifisitas adalah proporsi dari class=no yang teridentifikasi dengan benar. Sebagai contoh, dalam klasifikasi pelanggan butik dimana class=yes adalah pelanggan yang membeli baju sedangkan class=no adalah pelanggan yang tidak melakukan pembelian baju. Bila dihasilkan tingkat Sensitivitas sebesar 95%, artinya ketika dilakukan uji klasifikasi pada pelanggan yang membeli, maka pelanggan tersebut berpeluang 95% dinyatakan positive (membeli). Jika tingkat Spesifisitas sebesar 85%, artinya ketika dilakukan uji klasifikasi pada

pelanggan yang tidak membeli, maka pelanggan tersebut berpeluang 95% dinyatakan negative (tidak membeli).

Rumus untuk menghitung akurasi, spesifisitas, dan sensitivitas pada confusion matrix adalah sebagai berikut (Gorunescu, 2011: 319)

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{a + d}{a + b + c + d} \quad (2.11)$$

$$Sensitivitas = \frac{TP}{TP + FN} = \frac{a}{a + b} \quad (2.12)$$

$$Spesifisitas = \frac{TN}{TN + FP} = \frac{d}{d + c} \quad (2.13)$$

2.5 Waikato Environment for Knowledge (WEKA)

The Waikato Environment for Knowledge Analysis (WEKA) merupakan sebuah sistem open source yang berbasis java yang dipakai untuk data mining. Sistem ini dikembangkan oleh Universitas Waikato di Selandia Baru dan merupakan perangkat lunak gratis yang tersedia di bawah lisensi GNU (General Public License). WEKA menyediakan dukungan yang cukup luas untuk seluruh proses data mining mulai dari menyiapkan data masukkan, evaluasi pembelajaran, skema statistik, visualisasi data input sampai pada hasil pembelajaran.

Metode atau teknik yang digunakan pada WEKA adalah *Predictive* dan Descriptive karena sistem ini mendukung teknik-teknik data *preprocessing, clustering, classification, regression, visualization,* dan *feature Reduction*. (Witten, et all, 2011: 403-404).

Gambar 2.9 Tampilan Awal GUI WEKA

WEKA mulai dikembangkan sejak tahun 1994 dan telah menjadi software data mining open source yang terpopular. WEKA memiliki kelebihan seperti mempunyai banyak algoritma data mining dan machine learning, kemudahan dalam penggunaannya, selalu up-to-date terhadap algoritma-algoritma yang baru.

Software WEKA tidak hanya digunakan untuk akademik saja namun juga banyak dipakai oleh perusahaan untuk meramalkan bisnis suatu perusahaan. WEKA memiliki dukungan format file untuk inputnya, antara lain yaitu:

1. Comma Separated Values (CSV): Merupakan file teks dengan pemisah tanda koma (,) yang cukup umum digunakan. File ini dapat dibuat menggunakan Microsoft Excel atau dibuat sendiri menggunakan perangkat lunak notepad.
2. Format C45: Adalah format file yang dapat diakses menggunakan aplikasi WEKA.

3. Attribute-Relation File Format (ARFF): Merupakan tipe file teks yang berisi berbagai instance data yang berhubungan dengan suatu set atribut data yang dideskripsikan serta di dalam file tersebut.
4. SQL Server/ MySql Server: Dapat mengakses database dengan menggunakan SQL Server/MySql Server.

Beberapa menu dalam tampilan WEKA, di antaranya yaitu :

1. Explorer, menu ini memberikan akses untuk semua fasilitas yang menggunakan pilihan menu dan pengisian data. Pada menu ini terdapat enam sub-menu pada bagian atas window, sub-menu tersebut yaitu:
    a) Preprocess, proses pemilihan dataset yang akan diolah pemilihan filter,
    b) Classify, terdapat berbagai macam teknik klasifikasi dan evaluasinya yang digunakan untuk mengolah data,
    c) Cluster, terdapat berbagai macam teknik cluster yang dapat digunakan untuk mengolah data, Associate, terdapat berbagai macam teknik association rules yang dapat digunakan untuk mengolah data,
    d) Select Atribut, proses pemilihan aspek yang mempunyai hubungan paling relevan pada data,
    e) Visualize, proses menampilan berbagai plot dua dimensi yang dibentuk dari proses pengolahan data.

2. Experimenter, menu ini digunakan untuk mengatur percobaan dalam skala besar, dimulai dari running, penyelesaian, dan menganalisis data secara statistik.
3. Knowledge Flow, pada tampilan menu ini, pengguna memilih komponen WEKA dari toolbar untuk memproses dan menganalisis data serta memberikan alternatif pada menu Explorer untuk kondisi aliran data yang melewati sistem. Selain itu, Knowledge Flow juga berfungsi untuk memberikan model dan pengaturan untuk pengolahan data yang tidak mungkin dilakukan oleh Explorer.
4. Simple CLI, menu yang menggunakan tampilan command-line. Menu ini tampil dengan tampilan command-line untuk menjalankan class di weka.jar, dimana langkah pertama variabel Classpath dijelaskan di file Readme.

Pada sub-menu klasifikasi WEKA terdapat menu *test options* yang digunakan untuk menguji kinerja model klasifikasi. Ada empat model tes yaitu:

1. Use training set

   Pengujian dilakukan dengan menggunakan data latih itu sendiri. Akurasi akan sangat tinggi, tetapi tidak memberikan estimasi akurasi yang sebenarnya terhadap data yang lain (data yang tidak dipakai dalam latihan).

2. Supplied test set

Pengujian dilakukan dengan menggunakan data lain di mana file latih dan file tes tersedia secara terpisah. Dengan menggunakan opsi ini, dapat dilakukan prediksi pada data tes.

3. Cross–validation

Pada cross-validation, akan ada pilihan banyaknya fold yang akan digunakan. Nilai default-nya yaitu 10.

4. Percentage split

Hasil klasifikasi akan dites menggunakan k% dari data tersebut, di mana k adalah proporsi dari dataset yang digunakan untuk data latih. Persentase di kolom adalah bagian dari data yang dipakai sebagai dataset latih. Pada opsi ini, data latih dan data tes terdapat dalam satu file.

2.6. Penelitian yang Relevan

Penelitian tentang data mining telah banyak dilakukan. Beberapa di antaranya yang mendukung penelitian ini dengan variabel dan metode penelitian yang berkaitan. Beberapa penelitian menggunakan metode klasifikasi Decision Tree dan Naive Bayes. Hasil penelitian terkait data pendidikan yang menggunakan data mining untuk klasifikasi dikemukakan terlebih dahulu. Penelitian tersebut sama dengan penelitian ini yaitu fokus pada kasus drop out dan sejenisnya.

Penelitian yang dilakukan oleh Jared E. Knowles memaparkan bahwa usaha untuk menghasilkan sistem peringatan dini drop out yang akurat yang sesuai dengan daerah Wisconsin adalah bagai jarum dalam jerami.

Knowles memaparkan mengenai software DEWS yang menggunakan engine software statistik R. Hasil penelitian ini salah satunya merekomendasikan untuk tidak tergantung pada satu metode agar mendapatkan tingkat prediksi yang paling akurat. Penelitian ini berdasarkan data pendidikan tingkat SMA.

Penelitian yang dilakukan oleh Amelia Halim menggunakan data pendidikan dari MMT ITS untuk angkatan tertentu yaitu dari 2009 sampai 2012. Metode yang digunakan adalah Decision Tree dengan algoritma C4.5. Tujuan penelitian adalah untuk mengidentifikasi kecenderungan lulus tidak tepat waktu. Ujicoba dilakukan dengan 3 skenario (Maba, Semester 1 dan Semester 2) di mana masing-masing dibagi lagi dalam 2 kondisi (non-pruning dan pre-pruning) serta dilakukan pengulangan sebanyak 5 kali untuk setiap kondisi. Sehingga ada 10 percobaan untuk masing-masing skenario. Hasil penelitiannya antara lain tingkat prediksi akan lebih baik bila data latih semakin banyak. Demikian pula bila menambahkan variabel IPS semester 1 dan IPS semester 2 sebagai prediktor, tingkat prediksi dapat ditingkatkan.

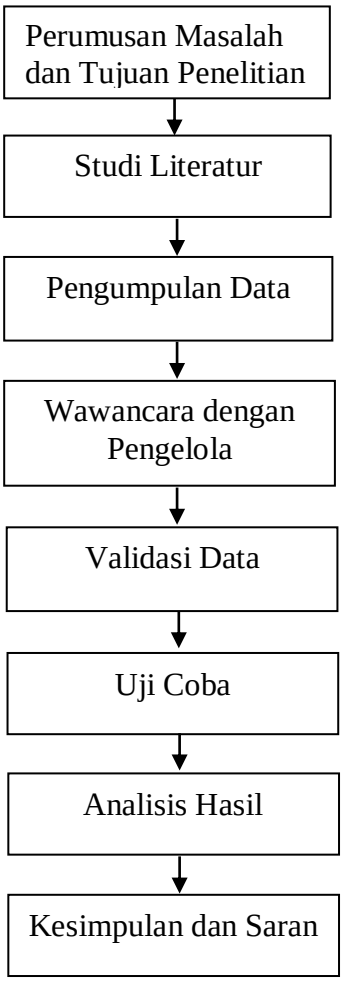
Penelitian yang dilakukan oleh Daniela Xhemali dan kawan-kawan menunjukkan bahwa *Naive Bayes* lebih baik dari *Decision Tree* maupun *Neural Networks* untuk data latih halaman web. Dalam penelitiannya, mereka mengumpulkan halaman web secara otomatis dengan membuat perangkat lunak mesin crawler yang mampu memperhatikan ketepatan dan duplikasi *link*. Data tersebut digunakan sebagai data latih untuk

menghitung probabilitas pasangan fitur dan kategori yang ditemukan. Setelah setiap sesi latihan, fitur-fitur tersebut menjadi semakin kuat terhubung dengan kategori berbeda. Setelah tahap ini, *Classifier* yang dibuat dapat digunakan untuk pengambilan keputusan dalam pengklasifikasian halaman web yang belum pernah diamati sebelumnya.

BAB 3

METODOLOGI PENELITIAN

Dalam bab 3 ini dibahas mengenai proses-proses atau tahapan-tahapan terkait penelitian yang digunakan untuk mencapai tujuan dalam penelitian. Untuk mempermudah pemahaman metode penelitian, berikut ini disajikan tahapan penelitian secara umum.



Gambar 3.1 Kerangka Tahapan Penelitian

3.1 Metode Penelitian

Untuk menghasilkan rancangan sistem peringatan dini dropout dengan metode Naive Bayes, langkah-langkah yang digunakan sebagai berikut :

1. Menentukan variabel indikator dan variabel prediktor dari data yang ada.

   Untuk menentukan variabel prediktor, didasarkan pada penelitian sebelumnya antara lain :

   a. Dalam (Halim, 2013) diperoleh variabel : Nama Perguruan Tinggi S1, Jurusan S1, IPK S1, Tahun Lulus S1, Tahun Masuk S2, Nama Perusahaan Tempat Bekerja, Pekerjaan, Nilai GMAT, Nilai Toefl, Nilai Materi Bidang, Nilai Wawancara, Nilai Skor Akhir, Program Studi S2, Bidang Keahlian S2, IPS Semester 1, IPS Semester 2, Sumber dana S2, dan Lama Studi.
   b. Dalam (Hidayat, 2011) : intelegensia mahasiswa dan penghasilan orang tua serta kualitas interaksi dengan teman dan hubungan keluarga.
   c. Dalam (Khoirunnisak, 2010) : usia, perbedaan asal daerah, perbedaan penghasilan orang tua dan jalur masuk.

   Naive Bayes bersifat menghitung probabilitas setiap variabel prediktor terhadap kejadian/ kelas pada variabel indikator. Selain itu, setiap variabel prediktor dianggap sangat tidak tergantung satu sama lain. Untuk itu maka perlu dilakukan verifikasi keterkaitan setiap variabel prediktor dengan variabel indikator. Harapannya adalah bias dari perhitungan statistik dapat dihindari sedemikian sehingga tingkat SSA (Sensitifitas, Spesifisitas dan

Akurasi) model yang dihasilkan dapat optimal. Berikut ini verifikasi untuk setiap variabel prediktor :

1. NRP

   NRP adalah nomor unik yang diberikan untuk setiap mahasiswa. Bila digunakan dalam Naive Bayes, maka akan memiliki probabilitas yang sama untuk setiap kejadian NRP yang berbeda. Oleh karena itu, tidak ada gunanya menggunakan NRP dalam Naive Bayes kecuali hanya sebagai pengenal (id).

2. Nama Perguruan Tinggi S1

   Nama Perguruan Tinggi S1 adalah perguruan tinggi asal dari mahasiswa yang bersangkutan. Sebuah perguruan tinggi mempunyai nilai akreditasi institusi. Artinya kualitas institusi mungkin memiliki hubungan dengan variabel indikator.

3. Jurusan S1

   Jurusan S1 adalah nama jurusan dari mahasiswa terkait. Sebagaimana variabel Nama Perguruan Tinggi S1, Jurusan S1 juga memiliki kualitas yang mungkin mempengaruhi mahasiswa yang bersangkutan sehingga mungkin memiliki pengaruh terhadap variabel indikator.

4. IPK S1

   Sebagai nilai capaian prestasi akademik bagi seorang mahasiswa, variabel IPK S1 jelas merupakan variabel yang mungkin memiliki pengaruh terhadap variabel indikator.

5. Tahun Lulus S1 dan Tahun Masuk S2

Dua variabel ini digabungkan untuk dibentuk menjadi variabel Waktu Tunggu. Variabel Waktu Tunggu ini diperoleh dengan mengurangkan variabel Tahun Masuk S2 dengan variabel Tahun Masuk S1. Hasil pengurangan tersebut dikelompokkan dalam 3 kategori yaitu Dekat, Sedang dan Jauh. Variabel Waktu Tunggu akan bernilai "Dekat" bila hasil pengurangan tersebut bernilai lebih kecil atau sama dengan 2. Variabel Waktu Tunggu akan mempunyai nilai "Sedang" bila hasil pengurangan tersebut bernilai lebih besar dari 2 dan lebih kecil atau sama dengan 10. Variabel Waktu Tunggu akan bernilai "Jauh" bila hasil pengurangan tersebut bernilai lebih dari 10.

Variabel Tahun Lulus S1 dan variabel Tahun Masuk S2 jelas tidak ada kaitan yang kuat dengan klasifikasi drop out. Namun bila digabungkan menjadi variabel Waktu Tunggu, variabel ini mungkin dapat digunakan sebagai salah satu penentu seorang mahasiswa akan masuk klasifikasi drop out atau tidak.

6. Nama Perusahaan Tempat Bekerja dan Pekerjaan

Variabel Nama Perusahaan Tempat Bekerja dan Variabel Pekerjaan mungkin tidak terlalu terkait dengan penentuan klasifikasi dropout. Tetapi bila dari salah satu atau keduanya digunakan untuk menurunkan variabel baru yaitu Status Pekerjaan yang berisi Ya atau Tidak, variabel baru tersebut kemungkinan besar memiliki pengaruh terhadap penentuan klasifikasi drop out.

Salah satu kemungkinannya adalah waktu yang digunakan untuk mengerjakan tugas (termasuk tesis) akan terbatas, yang pada gilirannya akan mungkin menentukan klasifikasi drop out mahasiswa.

7. Nilai GMAT

   Ujian GMAT merupakan saringan awal bagi calon mahasiswa. Jika dalam ujian tidak terjadi kecurangan, maka Nilai GMAT dapat digunakan sebagai variabel prediktor.

8. Nilai TOEFL

   Ujian TOEFL juga merupakan saringan awal bagi calon mahasiswa. Sehingga Nilai TOEFL dapat dipakai sebagai variabel prediktor.

9. Nilai Materi Bidang

   Ujian Materi Bidang pun merupakan saringan awal bagi calon mahasiswa. Maka dari itu, Nilai Materi Bidang pun dapat digunakan sebagai variabel prediktor.

10. Nilai Wawancara

    Saringan terakhir adalah ujian Wawancara. Dalam ujian ini calon mahasiswa akan diwawancarai berdasarkan isian form Wawancara dan form rekomendasi. Dengan proses tatap muka ini diharapkan dapat diketahui secara langsung mengenai kesiapan dan kemauan calon mahasiswa untuk menyelesaikan studi di MMT ITS. Sehingga jelas bahwa Nilai Wawancara dapat digunakan sebagai variabel prediktor.

11. Nilai Skor Akhir

Skor Akhir didapatkan dari total nilai ujian GMAT, TOEFL, Materi Bidang dan Wawancara. Hal ini berarti Skor Akhir jelas tidak independen terhadap 4 variabel lain. Padahal Naive Bayes mengasumsikan bahwa antar varibel prediktor sangat independen satu sama lain. Untuk itu Skor Akhir tidak dijadikan sebagai variabel prediktor.

12. Program Studi S2

Bila tidak memperhatikan data yang digunakan oleh Amelia Halim (Halim 2015), maka Program Studi S2 pasti akan dikira berisi data monoton yaitu Magister Manajemen Teknologi (MMT). Namun ternyata isinya adalah Bidang Keahlian yaitu :

- Managemen Teknologi Informasi
- Managemen Industri
- Managemen Proyek
- Managemen Maritim
- dll

Dalam setiap bidang keahlian, ada yang dropout, lulus tepat waktu atau lebih cepat, lulus tidak tepat waktu dan lulus dengan predikat pujian (cumlaude). Sehingga Program Studi S2 (mungkin sebaiknya bernama Bidang Keahlian S2) mungkin dapat digunakan untuk memprediksi dropout.

13. Bidang Keahlian S2

Bidang Keahlian S2 di sini juga memiliki ketidak jelasan nama sebagaimana variabel nomor 12. Karena isinya ternyata adalah jenis kelas yang diadakan untuk setiap bidang studi. Ada kelas Profesional yang diperuntukkan bagi non pejabat. Ada juga kelas Eksekutif yang dirancang khusus untuk pejabat (pengambil keputusan) dengan biaya sedikit lebih mahal dari kelas Profesional. Sebagaimana variabel nomor 12, variabel Bidang Keahlian S2 (mungkin seharusnya bernama Jenis Kelas S2) juga mungkin dapat digunakan sebagai variabel prediktor.

14. IPS Semester 1

Sistem Peringatan Dini yang akan dirancang memiliki kemampuan peringatan dini sejak calon mahasiswa memperoleh nilai ujian pertama kali yaitu pada saat tes masuk. Karena penentuan dropout mulai dilakukan pada saat akhir semester 3, maka IPS Semester 1 dapat digunakan terutama untuk mempertajam hasil peringatan dini pada tahap ke-2.

15. IPS Semester 2

Tahap pertama Sistem Peringatan Dini dilakukan pada saat tes masuk. Tahap kedua dilakukan pada akhir semester 1. Sedangkan tahap terakhir dilakukan pada akhir semester 2. Karena itu, maka IPS Semester 2 dapat digunakan terutama untuk mempertajam hasil peringatan dini pada tahap ke-3 (terakhir).

16. Sumber Dana S2

Secara umum, sumber dana untuk studi di MMT berasal dari biaya sendiri atau beasiswa. Umumnya beasiswa diberikan kepada mahasiswa yang lulus tes untuk memastikan agar beasiswa tersebut tidak sia-sia. Sedangkan sumber dana studi dari biaya sendiri oleh anggapan umum sering dikatakan sebagai penjamin bahwa studi akan dilaksanakan dengan baik agar dana tersebut tidak sia-sia. Untuk itu maka variabel Sumber Dana S2 dapat dimasukkan sebagai variabel prediktor dalam Sistem Peringatan Dini Dropout.

17. Lama Studi

Variabel Lama Studi adalah variabel yang berisi angka jumlah semester yang ditempuh oleh seorang mahasiswa untuk menyelesaikan studinya. Tentunya angka ini diketahui setelah mahasiswa menyelesaikan studinya, sehingga variabel ini tidak cocok digunakan untuk Sistem Peringatan Dini.

18. Intelegensia Mahasiswa

Tidak ada pengukuran *intelegent quotion* (IQ) pada saat tes masuk MMT. Namun ada tes GMAT dan ujian lainnya yang mungkin sudah mewakili variabel prediktor ini.

19. Penghasilan Orang Tua

Penghasilan orang tua terkait dengan kelengkapan fasilitas belajar dan kelegaan tentang biaya pendidikan. Mungkin variabel

prediktor ini memiliki pengaruh terhadap klasifikasi drop out mahasiswa.

20. Kualitas Interaksi Dengan Teman

    Mungkin yang dapat mewakili variabel ini dari variabel yang ada adalah Persentase Kehadiran Perkuliahan (PKP). Namun variabel PKP belum mendapatkan tempat di dalam sistem informasi akademik universitas. Penggunaannya masih manual untuk menentukan layak tidaknya mahasiswa mengikuti ujian.

    Variabel prediktor ini dapat juga di peroleh dari kuisioner dan wawancara. Namun daya yang dibutuhkan untuk melengkapi isi variabel ini sangat besar dan butuh waktu yang tidak sedikit. Terlebih akan sangat sulit untuk menemui seorang alumni, apalagi yang sudah ditetapkan drop out.

21. Hubungan Keluarga

    Untuk melengkapi variabel prediktor ini, mungkin satu-satunya cara adalah dengan kuisioner dan wawancara. Sebagaimana dengan variabel prediktor nomor 20, variabel prediktor ini sangat sulit untuk dilengkapi.

22. Usia

    Variabel prediktor ini mungkin memiliki hubungan dengan variabel indikator. Karena meskipun belajar dapat dilakukan hampir di semua usia, namun ada usia tertentu yang merupakan masa keemasan sebagai usia ideal untuk menimba ilmu.

23. Asal Daerah

Kecerdasan, kerajinan dan daya juang seorang mahasiswa, mungkin dipengaruhi oleh lingkungan tempat dia dibesarkan. Sehingga mungkin variabel Asal Daerah dapat menjadi salah satu penentu klasifikasi drop out.

24. Jalur Masuk

Untuk MMT ITS jalur masuknya sama, yaitu melalui ujian yang sama. Lagi pula bila variabel ini dimaksudkan terkait dengan sumber biaya studi, maka sudah terwakili oleh variabel prediktor Sumber Dana S2.

Berdasarkan studi pustaka tersebut di atas, akan dipetakan terhadap data ITS. Bila ada data yang bersesuaian, akan digunakan dalam penelitian ini.

2. Mengolah data yang telah diperoleh dan uji coba dengan metode klasifikasi Naive Bayes menggunakan software WEKA.

Data yang akan diperoleh terlebih dahulu dikenakan proses pembersihan dan validasi agar dapat diolah ke dalam software WEKA. Setelah data diolah, akan didapatkan nilai-nilai tingkat prediksi (akurasi, spesifisitas dan sensitifitas). Nilai-nilai tingkat prediksi tersebut kemudian dianalisis untuk kemudian diambil kesimpulan dari padanya.

Uji coba dilakukan dalam 3 tahapan. Yang membedakan masing-masing tahapan adalah penggunaan atribut yang dapat digunakan pada suatu tahapan uji coba. Pada tahapan pertama, semua atribut digunakan kecuali atribut IPS Semester 1 dan IPS Semester 2. Pada tahapan kedua,

semua atribut digunakan kecuali atribut IPS Semester 2. Pada tahapan ketiga, semua atribut digunakan.

Untuk setiap tahapan uji coba, dilakukan 4 jenis percobaan. Yang membedakan dari keempatnya adalah atribut yang digunakan dalam proses klasifikasi. Pertama, percobaan dengan pengelompokan atribut numerik berdasarkan pembulatan 1 angka dibelakang koma. Kedua, percobaan dengan pengelompokan atribut numerik berdasarkan pengelompokan dalam 5 kategori. Ketiga, percobaan dengan pemilihan menggunakan atribut yang dikenakan pengelompokan pada percobaan pertama atau percobaan kedua. Pilihan dijatuhkan pada atribut yang memiliki tingkat korelasi atribut yang lebih tinggi. Keempat, percobaan dengan menggabungkan atribut numerik yang digunakan pada percobaan pertama dan percobaan kedua.

Dari hasil percobaan tersebut dilakukan pengamatan untuk mengetahui performa metode klasifikasi Naive Bayes pada data drop out MMT ITS.

3. Penerapan hasil analisis pada perancangan sistem peringatan dini dropout.

Dalam perancangan sistem peringatan dini dropout, dibutuhkan temu wawancara dan hasil analisa untuk menentukan dan membuat bagian-bagian berikut ini :

1. Diagram use case

Diagram ini digunakan untuk menggambarkan aktor sistem dan fungsi yang harus ada dalam sistem.

2. Mockup sistem

Untuk penggambaran sedekat mungkin dengan sistem yang akan dikembangkan, digunakan penggambaran antarmuka sistem yang biasa disebut mockup sistem.

3.2 Pengumpulan Data

Dalam penelitian ini, data diambil dari Sistem Informasi Penerimaan Mahasiswa Baru ITS dan Sistem Informasi Akademik ITS dan dipadukan dengan data hasil tes masuk dari Sistem Pasca Sarjana ITS. Informasi tentang detail data didapatkan dari pihak ahli dalam ini BAPKM ITS dan bagian Akademik MMT ITS. Data yang dikumpulkan adalah data yang sesuai dengan yang telah diuraikan pada bagian metode penelitian di atas yang tersedia di database ITS dan dapat digunakan .

Selain data primer, dalam penelitian ini juga digunakan data sekunder. Yaitu berupa hasil penelitian dari penelitian sejenis yang akan digunakan dalam komparasi hasil penelitian. Ada dua penelitian sejenis yang dapat digunakan sebagai pembanding yaitu penelitian yang sama-sama menggunakan data mining untuk klasifikasi. Pertama, penelitian yang menggunakan metode klasifikasi *Decision Tree* dengan algoritma C4.5 dengan studi kasus pada mahasiswa MMT ITS (Halim, 2015). Kedua, penelitian yang menggunakan metode klasifikasi *Support Vector Machine* dengan studi kasus pada mahasiswa program magister Statistika ITS (Hilmiyah, 2017). Meskipun kedua penelitian tersebut tidak sama persis, namun ada beberapa kesamaan terutama penekanan pada salah satu ukuran

kinerja prediksi yaitu tingkat *recall* atau sensitifitas dari metode klasifikasi yang digunakan.

(Halaman Sengaja Dikosongkan)

BAB 4

PEMBAHASAN

Dari penelusuran data mahasiswa MMT ITS, diperoleh dataset dari 100 mahasiswa dengan komposisi 85 % berstatus “Lulus” dan 15 % berstatus “DO”. Didapatkan jumlah dalam dataset yang tidak terlalu banyak karena banyak item data yang bernilai kosong di mana seharusnya memiliki nilai tertentu. Dari jumlah calon dataset sebanyak 1446 data mahasiswa, hanya 100 data mahasiswa yang layak digunakan.

Pada Bab 3 telah dijabarkan 24 calon variabel yang didapatkan dari studi pustaka. Dari 24 variabel yang teridentifikasi tersebut, variabel Waktu Tunggu merupakan turunan dari variabel Tahun Masuk S2 dan Tahun Lulus S1. Karena Waktu Tunggu berasal dari 2 variabel maka jumlah calon variabel pokok sesungguhnya adalah 25 variabel. Dari 25 calon variabel tersebut, hanya 14 variabel yang tersedia dari database dan dapat digunakan. Terpilihnya 14 variabel ini sesuai dengan pemilihan subyektif yang telah dilakukan pada Bab 3 serta ketersediaan data dari database. Ada 4 variabel yang dapat digunakan untuk membentuk 2 variabel baru yaitu Waktu Tunggu dan Kesebidangan. Sehingga total variabel yang digunakan menjadi 16 variabel yang terdiri dari 15 atribut dan 1 kelas/ variabel indikator.

Variabel tersebut antara lain :

1. Indeks Prestasi Kumulatif S1

   Atribut ini biasa disingkat menjadi “IPK S1”. IPK S1 merupakan rerata nilai hasil belajar pada jenjang pendidikan strata 1 dengan rentang nilai dari 0 – 4. Atribut ini hasil entrian dari calon mahasiswa yang divalidasi otomatis oleh program komputer yang digunakan.

   Meskipun rentangnya hanya dari 0 – 4, namun karena hasil proses rerata yang menghasilkan nilai dalam rentang kontinyu dan

disederhanakan dengan hanya menampilkan nilai dengan dua angka dibelakang koma, maka akan dilakukan pengelompokan nilai untuk atribut seperti ini.

2. Indeks Prestasi Sementara Semester 1

   Atribut ini sering disebut sebagai “IPS Semester 1”. Variabel ini adalah rerata nilai hasil belajar mahasiswa pada semester 1 di MMT ITS dengan rentang nilai dari 0 - 4. Variabel ini hasil entrian dari dosen mata kuliah yang divalidasi otomatis oleh sistem yang digunakan.

3. Indeks Prestasi Sementara Semester 2

   Atribut ini dalam penyebutannya menjadi “IPS Semester 2”. Variabel ini adalah rerata nilai hasil belajar mahasiswa di semester 2 di MMT ITS dengan rentang nilai dari 0 - 4. Atribut ini hasil entrian dari dosen mata kuliah yang divalidasi otomatis oleh sistem yang digunakan.

4. Sumber Dana

   Atribut ini didapatkan dari sistem informasi ITS. Validasi sistem dilakukan dengan memberikan pilihan yang dapat dipilih oleh mahasiswa.

5. Nilai Tes GMAT

   Atribut ini merupakan hasil tes masuk dengan materi uji GMAT. Nilai yang digunakan atribut ini dientri oleh petugas MMT. Validasi entrian dilakukan pada saat diproses menjadi nilai dengan range normal dari 0-100 untuk digunakan dalam penentuan nilai akhir sebagai acuan penerimaan mahasiswa.

6. Nilai Tes TOEFL

   Atribut ini merupakan hasil tes masuk dengan materi uji TOEFL. Nilai yang digunakan atribut ini dientri oleh petugas CLC ITS.

Validasi entrian dilakukan pada saat diproses menjadi nilai dengan range normal dari 0-100 untuk digunakan dalam penentuan nilai akhir sebagai acuan penerimaan mahasiswa.

7. Nilai Tes Materi Bidang Minat

   Atribut ini merupakan hasil tes masuk dengan materi uji Materi Bidang Minat. Nilai yang digunakan atribut ini dientri oleh petugas MMT ITS. Validasi entrian dilakukan pada saat diproses untuk digunakan dalam penentuan nilai akhir sebagai acuan penerimaan mahasiswa.

8. Nilai Tes Wawancara

   Atribut ini merupakan hasil tes masuk yang dilakukan dengan wawancara. Nilai yang digunakan atribut ini dientri oleh pewawancara MMT ITS. Validasi entrian dilakukan pada saat diproses untuk digunakan dalam penentuan nilai akhir sebagai acuan penerimaan mahasiswa.

9. Skor Akhir Nilai Tes

   Atribut ini merupakan skor akhir yang terdiri dari 20 % nilai GMAT, 20 % nilai TOEFL, 30 % nilai Materi Bidang Minat, 30% nilai Wawancara. Nilai atribut ini digunakan sebagai acuan penentu apakah pemilik nilai akan diterima atau tidak di MMT ITS. Sehingga petugas MMT akan melakukan usaha lebih agar nilai ini valid, selain bantuan terotomasi dari program aplikasi yang digunakan.

10. Tahun Lulus S1

    Atribut ini menunjukkan tahun lulus yang bersangkutan pada strata S1. Data ini dientri oleh yang bersangkutan pada saat masih sebagai calon mahasiswa.

11. Tahun Masuk S2

Atribut ini menunjukkan tahun masuknya yang bersangkutan di MMT ITS. Data ini didapatkan dari ekstraksi NRP yang menunjukkan tahun masuk yang bersangkutan di MMT ITS.

12. Waktu Tunggu

Nilai atribut ini didapatkan dari pengurangan dari atribut Tahun Masuk S2 dengan Tahun Lulus S1. Pengurangan dilakukan otomatis menggunakan program aplikasi Microsoft Excell.

13. Bidang Minat

Atribut ini merupakan bidang minat yang diambil oleh yang bersangkutan di MMT ITS. Nilai atribut ini didapatkan dari komponen NRP yang menunjukkan bidang minat mahasiswa.

14. Jurusan S1

Atribut ini seharusnya berisi nama jurusan strata 1 yang telah ditempuh oleh yang bersangkutan. Nilai atribut ini didapatkan dari entrian calon mahasiswa yang mana bukan dalam bentuk pilihan tetapi dalam bentuk isian bebas. Sehingga pada atribut ini dibutuhkan validasi manual untuk memastikan bahwa nama jurusan tertulis konsisten sehingga dapat dilakukan pra proses pengelompokan.

15. Kesebidangan

Atribut ini merupakan bentukan yang mengacu pada atribut Bidang Minat dan atribut Jurusan S1. Kemungkinan nilainya adalah 1,2 dan 3. Penentuannya adalah sebagai berikut :

a. Sebidang (nilai = 1) bila nilai dari atribut Jurusan S1 termasuk ke dalam salah satu dari sekumpulan jurusan yang dianggap sebidang terhadap nilai dari atribut Bidang Minat. Misal untuk Bidang Minat = “Manajemen Teknologi Informasi”, maka

sekumpulan jurusan yang sebidang antara lain : Teknik Informatika, Sistem Informasi, Teknik Komputer dan lain-lain.

b. Sebidang sebagian (nilai = 2) bila atribut Jurusan S1 termasuk ke dalam salah satu dari sekumpulan jurusan yang dianggap sebidang sebagian terhadap atribut Bidang Minat. Misal Bidang Minat =”Manajemen Industri”, maka sekumpulan jurusan yang mungkin dianggap sebidang sebagian antara lain : Manajemen, Manajemen Transportasi, Teknik Elektronika Industri dan lain-lain.
c. Tidak sebidang (nilai = 3) yaitu bila atribut Jurusan S1 tidak termasuk a dan b.

16. Status Kelulusan (class)

Atribut ini adalah status kelulusan dari mahasiswa yang bersangkutan dari MMT ITS. Nilai yang dipakai hanya 2 yaitu “Lulus” dan tidak lulus (“DO”).

4.1. Praproses Data

Sudah menjadi hal yang umum di dalam penerapan proses penambangan data adalah melakukan praproses terhadap dataset yang diperoleh agar sesuai dengan proses metode yang digunakan. Dalam Naive Bayes, yang terjadi adalah dilakukan perhitungan probabilitas kemunculan dari setiap kejadian dari sebuah atribut untuk seluruh atribut yang ada. Hitungan probabilitas inilah yang akan digunakan sebagai model prediksi. Untuk itu, setiap item data sebaiknya dikenakan pengelompokan data pada setiap atribut numerik. Semua atribut yang diterima oleh Naive Bayes akan dianggap sebagai data nominal. Pengelompokan dibutuhkan untuk membuat data yang berdekatan dan memiliki arti yang kurang lebih mirip agar terkumpul ke dalam satu kategori.

Pada penelitian ini, digunakan 2 cara pengelompokan pada beberapa atribut yang perlu diubah menjadi nominal. Atribut tersebut antara lain : IPK S1, IPS Semester 1, IPS Semester 2, TOEFL, GMAT, Materi Bidang, Wawancara,

Skor Akhir, dan Waktu Tunggu. Proses pengelompokan tersebut dijabarkan sebagai berikut :

1. Pengelompokan ke dalam 5 kategori.

   Hasil proses ini akan diletakkan ke dalam variabel yang akan diperlakukan sebagai atribut baru. Ke-9 atribut di atas di kenakan pra proses ini yang dilakukan sebagai berikut :

   - Hitung nilai minimum ,
   - Hitung nilai maksimum,
   - Hitung panjang interval tiap kelas,
   - Tranformasi setiap datum ke dalam kelas yang sesuai.

2. Pengelompokan melalui pembulatan.

   Hasil proses ini akan menggantikan data yang lama pada atribut tersebut. Dari ke-9 atribut di atas, hanya atribut Waktu Tunggu yang tidak ikut diproses karena sudah mengelompok dalam bilangan bulat. Pembulatan yang dilakukan adalah sebagai berikut :

   - Pembulatan 0 angka dibelakang koma menjadi bilangan bulat untuk hasil tes masuk,
   - Pembulatan satu angka dibelakang koma untuk IPK/IPS.

Di awal terdapat 14 variabel terpilih. Kemudian muncul 2 variabel baru turunan dari 4 variabel awal. Lalu pada proses Pengelompokan Pembulatan, proses ini tidak memunculkan variabel baru melainkan hanya *update* item data dari 8 variabel numerik selain Waktu Tunggu. Berikutnya pada proses Pengelompokan 5 Kategori, proses ini memunculkan 9 variabel baru yang merupakan turunan dari 9 variabel numerik. Sehingga total 25 variabel di mana 1 variabel sebagai atribut indikator dan 24 variabel sebagai atribut prediktor.

4.2. Sebaran Data

Seperti yang telah dijelaskan di atas, dataset terdiri dari 85 % mahasiswa yang lulus dan 15 % mahasiswa yang dropout. Kebetulan jumlah mahasiswa yang digunakan atributnya dalam dataset sebanyak 100 orang. Sehingga satuan dari jumlah yang ada dapat dianggap dalam persen.

Jumlah atribut awal yang terpilih sebanyak 16 atribut. Berikut ini sebaran data dari ke-16 atribut tersebut :

a. Indeks Prestasi Kumulatif Strata 1 (IPK S1)

Sebaran data IPK S1 dinyatakan dalam tabel dibawah ini :

Tabel 4.1 Sebaran Data Atribut IPK S1

IPK S1	LULUS	DO	Jumlah
2.3	1	0	1
2.7	3	0	3
2.8	7	2	9
2.9	4	0	4
3.0	7	3	10
3.1	11	2	13
3.2	12	3	15
3.3	10	4	14
3.4	9	0	9
3.5	8	1	9
3.6	5	0	5
3.7	4	0	4
3.8	4	0	4
Total	85	15	100

b. Indeks Prestasi Sementara Semester 1 (IPS Sem 1)

Tabel 4.2 Sebaran Data Atribut IPS Sem 1

IPS Sem1	LULUS	DO	Jumlah
0.0	0	2	2
1.2	0	1	1
1.3	0	1	1

1.5	0	1	1
1.6	0	1	1
1.8	0	2	2
1.9	0	2	2
2.0	0	1	1
2.1	0	1	1
2.2	0	1	1
2.5	2	0	2
2.6	1	0	1
2.7	2	0	2
2.8	1	1	2
3.0	4	0	4
3.1	3	0	3
3.2	4	0	4
3.3	6	0	6
3.4	9	1	10
3.5	22	0	22
3.6	12	0	12
3.7	9	0	9
3.8	6	0	6
3.9	4	0	4
Total	85	15	100

c. Indeks Prestasi Sementara Semester 2 (IPS Sem 2)

Tabel 4.3 Sebaran Data Atribut IPS Sem 2

IPS Sem 2	LULUS	DO	Jumlah
0.0	0	3	3
1.0	0	1	1
1.2	0	1	1
1.4	0	1	1
1.5	0	1	1
1.8	0	1	1
1.9	0	1	1
2.0	0	2	2
2.1	0	1	1
2.3	0	1	1
2.4	1	1	2
2.6	1	0	1

2.7	2	0	2
2.8	3	0	3
2.9	5	0	5
3.1	6	0	6
3.2	2	0	2
3.3	8	0	8
3.4	8	0	8
3.5	22	0	22
3.6	11	0	11
3.7	7	0	7
3.8	9	1	10
Total	85	15	100

d. Nilai Tes GMAT

Tabel 4.4 Sebaran Data Atribut GMAT

GMAT	LULUS	DO	Jumlah
40	1	0	1
41	1	0	1
45	1	0	1
47	2	0	2
50	1	0	1
52	1	2	3
53	1	0	1
54	1	0	1
56	2	0	2
57	2	0	2
58	2	0	2
60	12	1	13
61	2	1	3
62	2	1	3
64	2	1	3
65	2	0	2
66	5	1	6
67	3	1	4
68	0	1	1
69	2	0	2
70	2	0	2
71	1	1	2

72	3	0	3
73	4	0	4
74	1	0	1
75	2	1	3
76	1	0	1
77	2	1	3
79	2	0	2
80	2	0	2
82	3	1	4
83	2	0	2
85	2	1	3
87	3	0	3
88	1	0	1
89	3	0	3
91	1	0	1
92	1	0	1
93	1	0	1
95	1	0	1
96	1	0	1
97	0	1	1
99	1	0	1
Total	85.0	15.0	100.0

e. Nilai Tes TOEFL

Tabel 4.5 Sebaran Data Atribut TOEFL

TOEFL	LULUS	DO	Jumlah
48	4	0	4
49	3	0	3
50	8	2	10
51	2	0	2
52	3	0	3
54	2	1	3
55	2	0	2
56	6	1	7
57	6	1	7
58	1	0	1
59	2	0	2
60	9	2	11

61	5	0	5
62	2	0	2
63	1	1	2
64	1	0	1
65	2	0	2
66	2	3	5
67	4	0	4
68	1	0	1
69	4	0	4
70	1	0	1
72	1	2	3
73	1	1	2
74	1	0	1
75	1	0	1
76	2	0	2
78	1	0	1
81	1	0	1
86	1	0	1
87	2	0	2
89	1	0	1
90	1	0	1
95	1	0	1
98	0	1	1
Total	85	15	100

f. Nilai Tes Materi Bidang Minat

Tabel 4.6 Sebaran Data Atribut Materi Bidang

Materi Bidang	LULUS	DO	Jumlah
35	0	1	1
40	0	1	1
44	0	1	1
47	1	0	1
48	1	0	1
49	1	0	1
50	1	0	1
51	1	0	1
52	3	0	3
53	5	0	5

55	1	0	1
56	4	1	5
57	1	1	2
58	3	1	4
59	3	0	3
60	6	0	6
61	3	0	3
62	4	0	4
63	8	1	9
64	0	1	1
65	4	1	5
66	3	0	3
67	3	0	3
68	3	0	3
69	4	0	4
70	4	1	5
71	3	0	3
72	3	1	4
73	3	0	3
74	1	0	1
75	1	1	2
78	1	2	3
80	0	1	1
85	1	0	1
89	1	0	1
91	2	0	2
95	1	0	1
99	1	0	1
Total	85	15	100

g. Nilai Tes Wawancara

Tabel 4.7 Sebaran Data Atribut Wawancara

Wawancara	LULUS	DO	Jumlah
50	1	0	1
72	1	0	1
74	1	1	2
76	1	1	2
77	1	0	1

78	2	1	3
79	3	0	3
80	2	1	3
81	5	1	6
82	5	1	6
83	6	2	8
84	10	1	11
85	7	1	8
86	9	1	10
87	6	2	8
88	4	2	6
89	8	0	8
90	5	0	5
91	3	0	3
92	4	0	4
94	1	0	1
Total	85	15	100

h. Skor Akhir

Tabel 4.8 Sebaran Data Atribut Skor Akhir

Skor Akhir	LULUS	DO	Jumlah
61	1	1	2
62	4	0	4
63	3	1	4
64	4	0	4
65	2	2	4
66	5	0	5
67	5	1	6
68	8	1	9
69	3	2	5
70	6	1	7
71	8	0	8
72	8	0	8
73	0	2	2
74	4	1	5
75	7	0	7
76	2	0	2

77	0	1	1
78	2	1	3
79	4	0	4
80	5	0	5
82	2	1	3
84	2	0	2
Total	85	15	100

i. Waktu Tunggu

Tabel 4.9 Sebaran Data Atribut Waktu Tunggu

Waktu Tunggu	LULUS	DO	Jumlah
0	16	4	20
1	21	1	22
2	5	3	8
3	10	3	13
4	4	2	6
5	4	1	5
6	1	0	1
7	3	0	3
8	3	0	3
9	4	0	4
10	2	0	2
11	1	0	1
12	1	1	2
13	3	0	3
14	2	0	2
15	1	0	1
16	1	0	1
20	1	0	1
21	1	0	1
22	1	0	1
Total	85	15	100

j. Tahun Lulus S1

Tabel 4.10 Sebaran Data Atribut Tahun Lulus S1

Tahun Lulus S1	LULUS	DO	Jumlah
1991	1	0	1

1993	1	0	1
1994	1	0	1
1997	1	0	1
1998	1	0	1
1999	2	0	2
2000	2	0	2
2001	1	0	1
2002	2	1	3
2003	1	0	1
2004	4	0	4
2005	2	0	2
2006	5	0	5
2007	1	0	1
2008	3	1	4
2009	3	1	4
2010	3	3	6
2011	12	0	12
2012	10	2	12
2013	20	5	25
2014	9	2	11
Total	85	15	100

k. Tahun Masuk S2

Tabel 4.11 Sebaran Data Atribut Tahun Masuk S2

Tahun Masuk S2	LULUS	DO	Jumlah
2013	41	7	48
2014	44	6	50
2016	0	2	2
Total	85	15	100

Dari sebaran data tersebut di atas, terlihat bahwa penelitian ini hanya menggunakan data untuk tahun masuk 2013, 2014 dan 2016. Padahal diawal proposal penelitian, direncanakan akan mengunakan data dari awal MMT ITS berdiri sampai tahun 2016. Hal ini sebagaimana telah dijelaskan di awal bab ini, disebabkan karena banyaknya item data yang kosong atau tidak valid, sehingga yang terpilih hanya yang ada di dalam dataset tersebut.

l. Sumber Dana

Tabel 4.12 Sebaran Data Atribut Sumber Dana

Sumber Dana	LULUS	DO	Jumlah
Sendiri/Instansi	56	13	69
Beasiswa_Calon_Dosen	4	0	4
Beasiswa_Fresh_Grad	11	1	12
Kerjasama	10	0	10
Beasiswa_BPKLN	1	0	1
Beasiswa_LPDP	3	1	4
Total	85	15	100

m. Bidang Minat

Tabel 4.13 Sebaran Data Atribut Bidang Minat

Bidang Minat	LULUS	DO	Jumlah
S2 MANAJEMEN INDUSTRI	43	5	48
S2 MANAJEMEN PROYEK	15	3	18
S2 MANAJEMEN TEKNOLOGI INFORMASI	26	7	33
S2 MANAJEMEN BISNIS MARITIM	1	0	1
Total	85	15	100

n. Jurusan S1

Tabel 4.14 Sebaran Data Atribut Jurusan S1

Jurusan S1	LULUS	DO	Jumlah
Desain_Interior	2	1	3
Desain_Komunikasi_Visual	1	1	2
Desain_Produk_Industri	1	0	1
Fisika	1	0	1
Gas_Dan_Petrokimia	0	1	1
Ilmu_Komputer	2	1	3
Kimia	3	0	3
Manajemen	4	1	5
Pendidikan_Teknik_Mesin	1	0	1
Sistem_Informasi	3	3	6
Statistika	4	0	4

Teknik_Arsitektur	1	1	2
Teknik_Elektro	6	0	6
Teknik_Fisika	2	0	2
Teknik_Industri	12	1	13
Teknik_Informatika	19	2	21
Teknik_Kelautan	1	0	1
Teknik_Keselamatan_Kesehatan_Kerja	2	0	2
Teknik_Kimia	2	0	2
Teknik_Komputer	1	0	1
Teknik_Mekatronika	1	0	1
Teknik_Mesin	4	0	4
Teknik_Perminyakan	1	0	1
Teknik_Sipil	4	1	5
Teknik_Sistem_Perkapalan	3	0	3
Teknik_Telekomunikasi	0	1	1
Teknologi_Industri	1	0	1
Teknologi_Informasi	1	0	1
Teknik_Arsitektur	1	0	1
Teknik_Elektro_Industri	0	1	1
Teknik_Elektronika	1	0	1
Total	85	15	100

o. Kesebidangan

Tabel 4.15 Sebaran Data Atribut Kesebidangan

Kesebidangan	**LULUS**	**DO**	**Jumlah**
1	37	8	45
2	9	1	10
3	39	6	45
Total	85	15	100

p. Status Kelulusan

Status Kelulusan digunakan sebagai variabel indikator. Komposisi dalam dataset sebagaimana telah disebutkan di atas sebanyak 85 % untuk Lulus dan 15 % untuk Dropout.

Selain itu juga terdapat 9 atribut baru yang berasal dari pengelompokan item dalam atribut ke dalam 5 kategori yaitu kategori A,

kategori B, kategori C, kategori D dan kategori E. Berikut ini sebaran datanya :

a. Indeks Prestasi Kumulatif Strata 1 dalam 5 Kategori (IPK S1_5)

Tabel 4.16 Sebaran Data Atribut IPK S1 dalam 5 Kategori

IPK S1	LULUS	DO	Jumlah
A	13	0	13
B	27	5	32
D	14	2	16
C	30	8	38
E	1	0	1
Total	**85**	**15**	**100**

b. Indeks Prestasi Sementara Semester 1 dalam 5 Kategori (IPS Sem 1_5)

Tabel 4.17 Sebaran Data Atribut IPS Sem 1 dalam 5 Kategori

IPS Sem1	LULUS	DO	Jumlah
A	31	0	31
B	37	1	38
C	17	9	26
D	0	3	3
E	0	2	2
Total	**85**	**15**	**100**

c. Indeks Prestasi Sementara Semester 2 dalam 5 Kategori (IPS Sem 2_5)

Tabel 4.18 Sebaran Data Atribut IPS Sem 2 dalam 5 Kategori

IPS Sem 2	LULUS	DO	Jumlah
A	27	1	28
B	38	0	38
C	20	7	27
D	0	4	4
E	0	3	3
Total	**85**	**15**	**100**

d. Waktu Tunggu dalam 5 Kategori

Tabel 4.19 Sebaran Data Atribut Waktu Tunggu dalam 5 Kategori

Waktu Tunggu	LULUS	DO	Jumlah
A	56	13	69
B	11	1	12
C	11	1	12
D	4	0	4
E	3	0	3
Total	**85**	**15**	**100**

e. Nilai Test GMAT dalam 5 Kategori

Tabel 4.20 Sebaran Data Atribut GMAT dalam 5 Kategori

GMAT	LULUS	DO	Jumlah
A	10	1	11
B	17	3	20
C	27	6	33
D	25	5	30
E	6	0	6
Total	**85**	**15**	**100**

f. Nilai Test TOEFL dalam 5 Kategori

Tabel 4.21 Sebaran Data Atribut TOEFL dalam 5 Kategori

TOEFL	LULUS	DO	Jumlah
A	3	1	4
B	4	0	4
C	12	3	15
D	29	6	35
E	37	5	42
Total	**85**	**15**	**100**

g. Nilai Tes Materi Bidang Minat dalam 5 Kategori

Tabel 4.22 Sebaran Data Atribut Materi Bidang Minat dalam 5 Kategori

Materi Bidang	LULUS	DO	Jumlah

A	5	0	5
B	4	4	8
C	45	5	50
D	30	3	33
E	1	3	4
Total	**85**	**15**	**100**

h. Nilai Tes Wawancara dalam 5 Kategori

Tabel 4.23 Sebaran Data Atribut Wawancara dalam 5 Kategori

Wawancara	LULUS	DO	Jumlah
A	40	5	45
B	41	8	49
C	3	2	5
D	0	0	0
E	1	0	1
Total	**85**	**15**	**100**

i. Skor Akhir dalam 5 Kategori

Tabel 4.24 Sebaran Data Atribut Skor Akhir dalam 5 Kategori

Skor Akhir	LULUS	DO	Jumlah
A	9	1	10
B	15	2	17
C	20	3	23
D	27	5	32
E	14	4	18
Total	**85**	**15**	**100**

4.3. Uji Signifikansi dan Multikolinieritas

Uji Signifikansi dan Uji Multikolinieritas dilakukan terhadap variabel yang akan digunakan dalam proses data mining dilakukan dengan alat bantu software SPSS. Berikut tabel hasil uji yang dilakukan.

Tabel 4.25 Tabel Uji Signifikansi dan Uji Multikolinieritas

Atribut	t	sig.	Tolerance	VIF
IPKS1	0.607	0.546	0.067	14.883

IPSSem1	4.863	0.000	0.080	12.527
IPSSem2	3.005	0.004	0.064	15.586
SumberDana	0.737	0.464	0.634	1.577
GMAT	-1.320	0.191	0.008	130.138
TOEFL	-1.328	0.188	0.011	93.391
Materi	-0.855	0.395	0.006	162.478
Wawancara	-0.775	0.441	0.023	43.537
SkorAkhir	0.915	0.363	0.002	522.998
TahunMasukS2	-2.259	0.027	0.598	1.671
WaktuTunggu	-0.688	0.494	0.055	18.228
IPSSem1_5	2.419	0.018	0.161	6.202
IPKS1_5	0.711	0.479	0.064	15.604
IPSSem2_5	2.060	0.043	0.109	9.139
WaktuTunggu_5	0.567	0.573	0.055	18.242
GMAT_5	-0.716	0.476	0.053	18.859
TOEFL_5	-0.250	0.803	0.064	15.632
MateriBid_5	1.540	0.128	0.131	7.606
Wawancara_5	0.034	0.973	0.149	6.717
MateriBid	1.541	0.124	0.131	7.606
SkorAkhir_5	-0.997	0.322	0.051	19.549
BidangMinat	-1.359	0.178	0.403	2.480
JurusanS1	0.038	0.970	0.678	1.475
Sebidang	0.272	0.786	0.536	1.865

Untuk menentukan simpulan terkait signifikansi, dibutuhkan nilai t tabel. Nilai t tabel sebesar 1,9921 didapatkan dari tabel distribusi t berdasarkan tingkat keyakinan 5 % dengan pengamatan 2 sisi (2,5%) serta jumlah pengamatan 100 (total dataset yang digunakan) dan jumlah variabel sebanyak 25 buah. Dari nilai t, bila nilai t lebih besar dari 1,9921 berarti variabel terkait berpengaruh terhadap variabel indikator. Dari nilai signifikansi, bila nilainya lebih kecil dari 0,05 berarti variabel terkait berpengaruh signifikan terhadap variabel indikator.

Dari nilai t bila dibandingkan dengan nilai t tabel, dari 24 variabel prediktor, hanya terdapat 4 variabel yang dinyatakan berpengaruh terhadap variabel indikator. Variabel tersebut antara lain : IPS Semester 1 dengan pembulatan, IPS Semester 2 dengan pembulatan, IPS Semester 1 dengan 5 kategori dan IPS Semester 2 dengan 5 Kategori. Sedangkan dari nilai signifikansi, terdapat 5 variabel yang dinyatakan berpengaruh signifikan. Yaitu 4 variabel yang telah disebutkan serta variabel Tahun Masuk S2.

Untuk menentukan simpulan terkait multikolinieritas dilakukan dengan 2 cara. Pertama, dengan melihat nilai *Tolerance*, yaitu jika nilainya lebih kecil dari

0,1 berarti terjadi multikolinieritas pada data yang diuji. Kedua, dengan melihat nilai VIF (Variance Inflation Factor), yaitu bila nilainya lebih besar dari 10 berarti ada multikolinieritas pada data yang diuji.

Dari hasil uji multikolinieristas terlihat bahwa 14 variabel yang digunakan dalam uji coba penelitian ini dinyatakan mengalami multikolonieritas. Artinya, benar terjadi adanya hubungan antar variabel. Padahal Naive Bayes mengasumsikan bahwa setiap variabel indikator tidak ada hubungan sama sekali. Hanya 10 variabel yang independen yaitu : Sumber Dana, Tahun Masuk S2, IPS Semester 1 dengan 5 kategori, IPS Semester 2 dengan 5 Kategori, Materi Bidang, Materi Bidang dengan 5 kategori, Wawancara dengan 5 kategori, Bidang Minat, Jurusan S1 dan Kesebidangan.

4.4. Proses Data Mining

Setelah data siap untuk diproses, selanjutnya proses yang dilakukan adalah menentukan Korelasi Atribut dari setiap variabel prediktor terhadap variabel indikator. Kemudian dilakukan pengujian dalam tiga tahapan :

1. Tahap Awal (Tahap I)

   Dalam tahap ini, atribut IPK Semester 1 dan IPK Semester 1 (5 Kategori) serta IPK Semester 2 dan IPK Semester 2 (5 Kategori) tidak diikutsertakan. Karena hanya dua atribut tersebut yang dihasilkan dari proses perkuliahan di MMT ITS.

2. Tahap Semester 1 (Tahap II)

   Dalam tahap ini, semua variabel pada tahap awal diikutkan dan ditambahkan atribut IPK Semester 1 dan IPK Semester 1 (5 Kategori). Sedangkan atribut IPK Semester 2 dan IPK Semester 2 (5 Kategori) belum diikutsertakan.

3. Tahap Semester 2 (Tahap III)

   Dalam tahap ini, semua variabel pada tahap Semester 1 dan atribut IPK Semester 2 dan IPK Semester 2 (5 Kategori) diikutsertakan dalam uji coba.

Dalam masing-masing tahap uji coba tersebut dilakukan 4 percobaan dengan mencatat nilai akurasi, sensitifitas dan spesifisitas yang dihasilkan. Di setiap percobaan tersebut dilakukan dengan metode validasi silang K dengan nilai K = 10. Yang artinya, dataset yang digunakan akan dibagi 10 sub dataset. Kemudian secara bergilir masing-masing sub dataset akan diuji dengan 9 sub

dataset lainnya sebagai data latih. Ke-4 variasi percobaan yang dilakukan adalah sebagai berikut :

1. Percobaan dengan dataset menggunakan atribut pengelompokan pembulatan.

   Percobaan ini menggunakan atribut yang tidak dikenakan proses pengelompokan dan atribut yang dikenakan proses pengelompokan pembulatan.
2. Percobaan dengan dataset menggunakan atribut pengelompokan 5 kategori.

   Percobaan ini menggunakan atribut yang tidak dikenakan proses pengelompokan dan atribut yang dikenakan proses pengelompokan 5 kategori.
3. Percobaan dengan dataset menggunakan atribut campuran pilihan.

   Percobaan ini menggunakan atribut yang tidak dikenakan proses pengelompokan. Selain itu juga menggunakan atribut yang dikenakan proses pengelompokan pembulatan atau atribut yang dikenakan proses pengelompokan 5 kategori. Pemilihannya didasarkan pada nilai Korelasi Atribut yang tertinggi.
4. Percobaan dengan dataset menggunakan atribut gabungan.

   Percobaan ini menggunakan atribut yang tidak dikenakan proses pengelompokan dan atribut yang dikenakan proses pengelompokan baik pembulatan maupun 5 kategori.

4.4.1 Penghitungan Korelasi Atribut

Perhitungan Korelasi Atribut dilakukan dengan menggunakan aplikasi WEKA dengan menggunakan fungsi "CorelationAttributEval" dan "Ranker". Dengan menggunakan kedua fungsi tersebut, dapat diperoleh nilai Korelasi Atribut dari masing-masing atribut beserta rerata peringkatnya melalui proses uji coba validasi silang dengan K=10. Berikut hasil yang diperoleh dari proses tersebut :

Tabel 4.26. Peringkat Korelasi Atibut

Nilai Rerata Korelasi	Nilai Rerata Peringkat	Atribut
0.295 +- 0.009	2 +- 0	IPS_Sem1_5
0.264 +- 0.012	3 +- 0	IPS_Sem2_5
0.158 +- 0.006	4.6 +- 0.92	IPS_Sem2
0.154 +- 0.004	5.2 +- 0.6	IPS_Sem1
0.14 +- 0.031	6.6 +- 1.11	Sumber_Dana

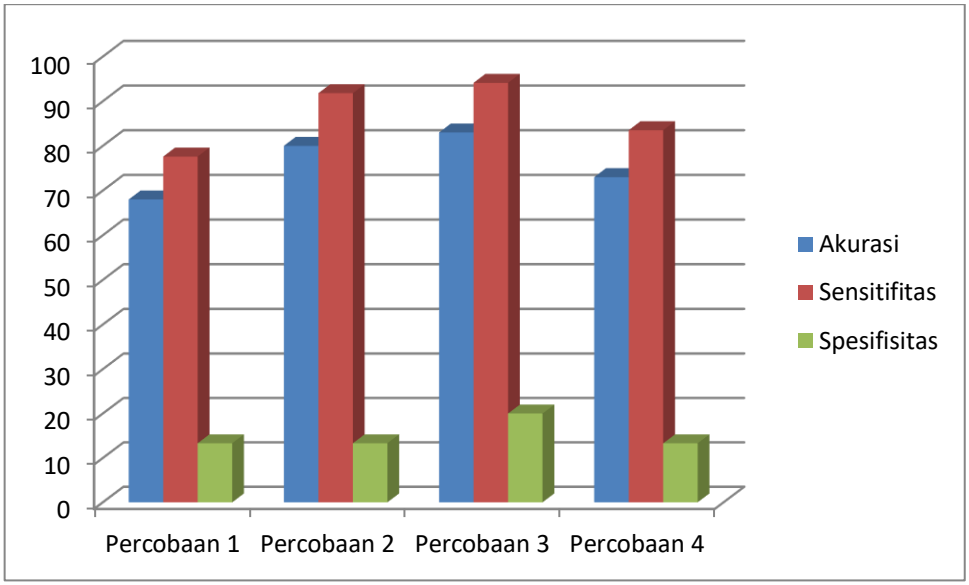
0.132 +- 0.029	7.1 +- 2.17	Waktu_Tunggu5
0.11 +- 0.019	9.5 +- 3.83	Materi_Bid5
0.105 +- 0.014	10 +- 2	Waktu_Tunggu
0.096 +- 0.005	11.3 +- 1.35	Skor_Akhir
0.106 +- 0.032	11.8 +- 5.02	Bidang_Minat
0.092 +- 0.008	12.8 +- 3.46	Jurusan_S1
0.087 +- 0.007	14.2 +- 2.52	Tahun_Lulus_S1
0.087 +- 0.008	14.6 +- 2.58	IPK_S1
0.091 +- 0.02	15.1 +- 5.41	IPK_S1_5
0.084 +- 0.004	15.1 +- 1.76	Materi_Bidang
0.08 +- 0.005	17.2 +- 3.22	TOEFL
0.08 +- 0.004	17.5 +- 1.86	GMAT
0.072 +- 0.023	18 +- 5.48	Wawancara5
0.073 +- 0.009	19.3 +- 2.79	Wawancara
0.062 +- 0.017	21.7 +- 3.9	TOEFL5
0.061 +- 0.013	22.7 +- 2.28	Tahun_Masuk_S2
0.055 +- 0.026	22.7 +- 3.13	Kesebidangan
0.051 +- 0.012	23.5 +- 1.57	GMAT5
0.045 +- 0.013	24.9 +- 1.37	Skor_Akhir5

4.4.2 Tahap Data Sebelum Perkuliahan (Awal)

Tabel 4.27. Rekapitulasi 4 Percobaan pada Tahap Awal

Akurasi	Sensitifitas	Spesifisitas
Percobaan 1 : Pembulatan pada atribut numerik		
68 %	77,6 %	13,3 %
Percobaan 2 : Pembagian 5 kategori pada atribut numerik		
80 %	91,8 %	13,3 %
Percobaan 3 : Campuran pembulatan atau pembagian 5 kategori, pilih korelasi atribut terbaik		
83 %	94,1 %	20 %
Percobaan 4 : Atribut hasil keduanya digunakan semua		
73 %	83,5 %	13,3 %

Dari tabel di atas terlihat bahwa tingkat akurasi paling rendah 68 % sedangkan tingkat sensitifitas paling rendah 77,6 %. Keduanya berasal dari uji coba dengan dataset yang mengandung atribut pembulatan saja (Percobaan 1). Tiga uji coba lainnya menunjukkan tingkat akurasi dan sensitifitas yang lebih tinggi lagi. Namun, hal ini tidak diikuti oleh tingkat spesifisitas yang hanya mencapai 20 % pada uji coba dengan dataset campuran pilihan (Percobaan 3). Meskipun setidaknya pada uji coba Tahap Awal ini dapat menangkap 20 % data dropout. Gambar 4.1 menampilkan perbandingan secara grafis pada tahap ini.



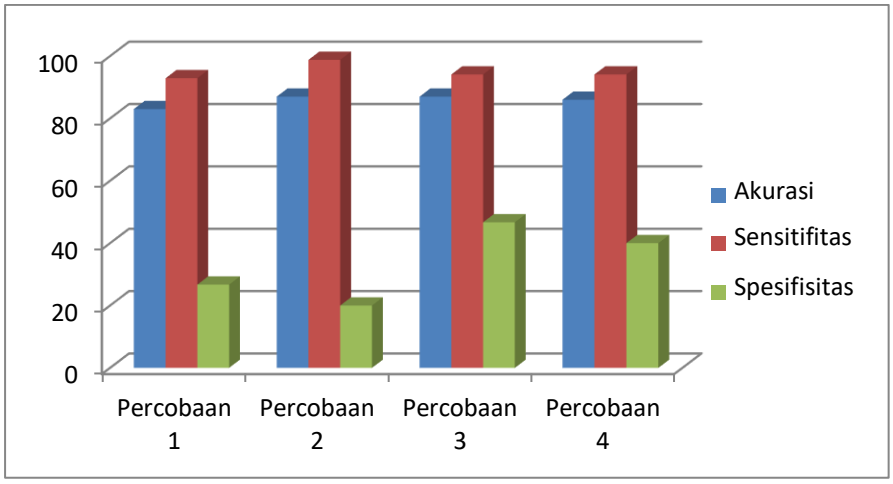
Gambar 4.1 Grafik Perbandingan Tingkat Prediksi pada Uji Coba Tahap Awal

4.4.3 Tahap Semester 1

Tabel 4.28. Rekapitulasi 4 Percobaan pada Tahap Semester I

Akurasi	Sensitifitas	Spesifisitas
Percobaan 1 : Pembulatan pada atribut numerik		
83 %	92,9 %	26,7 %
Percobaan 2 : Pembagian 5 kategori pada atribut numerik		
87 %	98,8 %	20 %
Percobaan 3 : Campuran pembulatan atau pembagian 5 kategori , pilih korelasi atribut terbaik		
87 %	94,1 %	46,7 %
Percobaan 4 : Atribut hasil keduanya digunakan semua		
86 %	94,1 %	40 %

Karena uji coba ini menggunakan data pincang, artinya data dropout jauh lebih kecil dari pada data lulus, serta tingkat akurasi dan sensitifitas tiap uji oba di atas tidak mengecewakan, maka dilakukan penekanan analisis yang berfokus pada tingkat spesifisitas. Spesifisitas terendah sebesar 20 % diperoleh pada uji coba dengan dataset yang mengandung atribut pengelompokan 5 kategori saja (Percobaan 2). Sedangkan spesifisitas tertinggi didapatkan pada uji coba dengan dataset yang mengandung atribut campuran pilihan dengan tingkat spesifisitas sebesar 46,7 % (Percobaan 3). Gambar 4.2 menampilkan perbandingan secara grafis pada uji coba tahap ini.



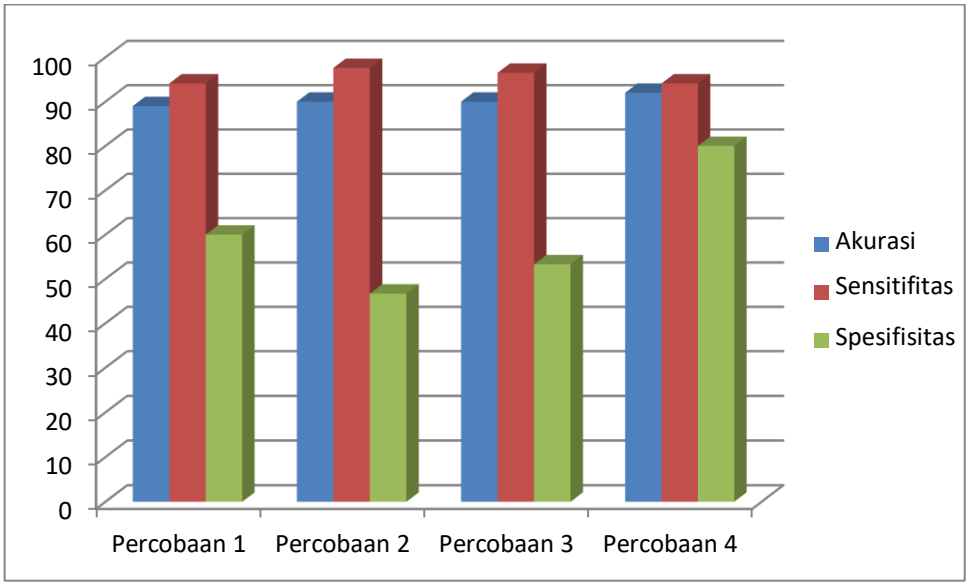
Gambar 4.2. Grafik Perbandingan Tingkat Prediksi pada Uji Coba Tahap Semester 1

4.4.4 Tahap Semester 2

Tabel 4.29. Rekapitulasi 4 Percobaan pada Tahap Semester 2

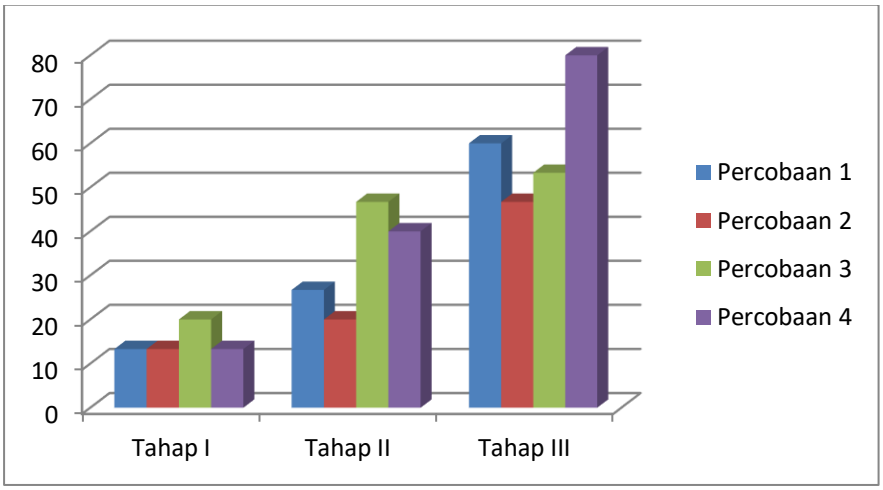
Akurasi	Sensitifitas	Spesifisitas
Percobaan 1 : Pembulatan pada atribut numerik		
89 %	94,1 %	60 %
Percobaan 2 : Pembagian 5 kelas pada atribut numerik		
90 %	97,6 %	46,7 %
Percobaan 3 : Campuran pembulatan atau pembagian 5 kategori, pilih korelasi atribut terbaik		
90 %	96,5%	53,3 %
Percobaan 4 : Atribut hasil keduanya digunakan semua		
92 %	94,1 %	80 %

Angka akurasi dan sensitifitas keempat percobaan pada Tahap III di atas sangat baik. Lebih baik dari percobaan pada 2 tahap sebelumnya. Sedangkan pada tingkat spesifisitas mengalami lonjakan yang sangat signifikan. Tingkat spesifisitas terendah sebesar 53,3 % pada uji coba dengan dataset yang mengandung atribut campuran pilihan. Pencapaian tertinggi pada uji coba dengan dataset yang mengandung atribut gabungan dengan tingkat spesifisitas sebesar 80 %. Gambar 4.3 menampilkan perbandingan secara grafis pada tahap ini.



Gambar 4.3 Grafik Perbandingan Tingkat Prediksi pada Uji Coba Tahap Semester 2

Dari ketiga tahapan uji coba, bila kita fokuskan pada tingkat spesifisitas pada masing-masing percobaan, akan terlihat jelas bahwa nilai tingkat spesifisitas semakin meningkat pada tahap II (semester 1) dan tahap III (semester 2). Kemudian, terkait kandungan atribut yang mengalami multikolinieritas, data hasil percobaan tahap ini menunjukkan bahwa penggabungan suatu atribut (Percobaan 1) dengan atribut turunannya (Percobaan 2) dalam suatu uji coba (Percobaan 4), mampu menghasilkan tingkat spesifisitas yang lebih tinggi yaitu pada uji coba Tahap II dan uji coba Tahap III. Gambar 4.4 menampilkan perbandingan ini secara grafis.



Gambar 4.4 Grafik Perbandingan Tingkat Spesifisitas Tiap Percobaan pada Tiap Tahapan Uji Coba

4.4.5. Perbandingan dengan Hasil Penelitian Lain

Seperti yang telah dijelaskan pada Bab 3, pada bagian ini akan membandingkan hasil penelitian ini dengan hasil penelitian lain. Penelitian pertama adalah penelitian dengan metode Support Vector Machine (Hilmiyah, 2017). Penelitian kedua adalah penelitian dengan metode Decision Tree pada algoritma C4.5 (Halim, 2015). Perbandingan pertama ditunjukkan oleh tabel 4.30.

Tabel 4.30. Perbandingan pada Tahap Awal

Percobaan	Akurasi	Spesifisitas
Percobaan 1	68	13.3
Percobaan 2	80	13.3
Percobaan 3	83	20
Percobaan 4	73	13.3
SVM	41.49	46.68
C4.5	60.35	31.1

Pada perbandingan ini, sebagaimana ditunjukkan oleh tabel di atas, terlihat bahwa SVM lebih unggul. Sedangkan hasil penelitian ini masih dibawah SVM dan C4.5. Hal ini wajar mengingat atribut yang digunakan oleh penelitian lain adalah lebih lengkap.

Pada perbandingan berikutnya, hasil penelitian ini hanya akan dibandingkan dengan hasil penelitian C4.5. Karena kedua penelitian ini sama-sama menggunakan tahapan penelitian 3 tahap. Tabel 4.31 dan tabel 4.32 menunjukkan perbandingan ini.

Tabel 4.31. Perbandingan pada Tingkat Akurasi

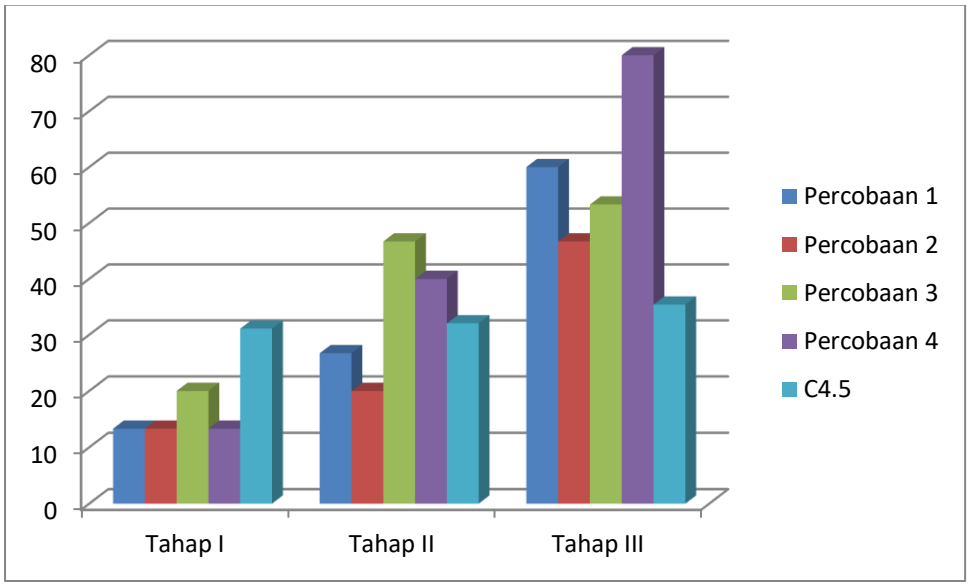
Percobaan	Tahap I	Tahap II	Tahap III
Percobaan 1	68	83	89
Percobaan 2	80	87	90
Percobaan 3	83	87	90
Percobaan 4	73	86	92
C4.5	60.35	61.25	61.5

Pada tabel di atas terlihat bahwa hasil penelitian ini memiliki tingkat akurasi yang unggul pada semua percobaan di tiap tahap uji coba. Namun keunggulan ini belum lengkap bila tidak memperhatikan tingkat spesifisitas.

Tabel 4.32. Perbandingan pada Tingkat Spesifisitas

Percobaan	Tahap I	Tahap II	Tahap III
Percobaan 1	13.3	26.7	60
Percobaan 2	13.3	20	46.7
Percobaan 3	20	46.7	53.3
Percobaan 4	13.3	40	80
C4.5	31.1	32.05	35.39

Pada tabel spesifisitas di atas terlihat bahwa pada tahap I, C4.5 lebih unggul. Namun mulai pada tahap II mulai tersaingi. Puncaknya pada perbandingan tahap III yang mana tingkat spesifisitas C4.5 terlewati. Gambar 4.5 Memperjelas hal ini.



Gambar 4.5 Grafik Perbandingan Spesifisitas Tiap Percobaan pada Tiap Tahap dengan C4.5

4.5. Perancangan Sistem

4.5.1. Kebutuhan Sistem

Sistem Peringatan Dini Dropout pada dasarnya adalah sistem yang sederhana. Umumnya, di setiap Perguruan Tinggi yang terakreditasi sudah terdapat sistem informasi terkait akademik. Bahkan sudah ada menu Peringatan Dini untuk pencegahan kejadian dropout dengan cara yang sudah biasa berjalan selama ini.

Untuk mengetahui kebutuhan Sistem Peringatan Dini Dropout, perlu kiranya diketahui mengenai uraian jalannya sistem peringatan dini yang selama ini dilakukan/ berjalan. Kemudian berikutnya akan membahas mengenai rancangan sistem baru yang mungkin dapat berjalan dengan baik berdasarkan bentuk integrasi sistem yang sudah ada.

4.5.1.1. Sistem Peringatan Dini Saat Ini

Yang menjadi pijakan peringatan dini pada sistem saat ini adalah bahwa yang mendapat peringatan adalah hanya mahasiswa yang telah memenuhi

sebagian kriteria dropout. Padahal ada beberapa contoh kasus bahwa mahasiswa yang prestasi akademiknya memenuhi sebagian kriteria dropout, ternyata berhasil lulus dengan nilai memuaskan. Begitu pula sebaliknya, ada contoh kasus bahwa mahasiswa yang awalnya menunjukkan prestasi cemerlang, kemudian ternyata berujung pada kejadian dropout.

Sistem Peringatan Dini yang sudah berjalan selama ini adalah peringatan yang diawali oleh BAPKM dengan melakukan penyaringan terhadap mahasiswa yang prestasi akademiknya dan masa penyelesaian kuliahnya memenuhi sebagian dari kriteria dropout. Hasilnya adalah daftar mahasiswa yang layak diberi peringatan. Daftar tersebut kemudian dimasukkan ke dalam *table* Peringatan dalam *database* yang digunakan oleh Sistem Informasi Akademik (SIA) untuk mendapatkan data peringatan. Ketika Dosen Wali, Pengelola Jurusan dan Mahasiswa mengakses data peringatan terkait peran masing-masing user tersebut, SIA akan mengambilkan data peringatan pada *table* Peringatan.

Demikianlah cara kerja dan penggunaan peringatan dini yang selama ini dilakukan di ITS termasuk di departemen MMT.

4.5.1.2. Rancangan Sistem Peringatan Dini Dropout

Sistem Peringatan Dini Dropout menawarkan hal yang berbeda. Yaitu pendeteksian dropout tidak berdasarkan nilai saja, melainkan pola atau model kejadian dropout yang mana pembentuknya bukan saja berdasarkan nilai mahasiswa saja. Sebab, mahasiswa yang sedang berprestasi akademik tidak jelek tetapi menurut model tersebut ia memiliki kecenderungan untuk dropout, maka namanya akan muncul dalam daftar peringatan. Untuk kali ini, model yang digunakan adalah model klasifikasi Naive Bayes.

Berdasarkan sistem yang sudah berjalan selama ini, Sistem Peringatan Dini Dropout sebaiknya diintegrasikan dengan sistem yang sudah ada. Namun bila proses integrasi membutuhkan waktu dan tenaga yang tidak sedikit, sebaiknya menggunakan cara kerja sistem yang sudah berlaku selama ini oleh BAPKM. Sebenarnya tidak banyak yang diubah, karena Sistem Peringatan Dini Dropout

hanya menambahkan sistem peringatan yang sudah ada yang dilakukan secara semi manual oleh BAPKM.

Proses yang ada dalam Sistem Peringatan Dini Dropout sangat sederhana. Yang dilakukan user antara lain menyiapkan data dan menjalankan fungsi prediksi. Sistem menghasilkan daftar mahasiswa terdeteksi dropout. Agar dapat terintegrasi dengan sistem yang sudah ada, maka harus ada fungsi untuk mengirimkan daftar ini secara elektronik ke *table* Peringatan dalam *database* terintegrasi.

Dari hasil uji coba, dapat diketahui apa saja yang dibutuhkan untuk membangun Sistem Peringatan Dini Dropout. Sistem Peringatan Dini Dropout ini diasumsikan akan dikembangkan untuk MMT ITS dahulu sesuai dengan batasan permasalahan penelitian ini. Berikut akan dibahas mengenai aktor dan fungsi yang harus ada.

4.5.1.2.1. Aktor

Aktor yang akan berinteraksi dengan sistem ini adalah sedikit berbeda dengan sistem yang berlaku selama ini. Berikut uraian dari aktor yang ada dan perannya :

- Mahasiswa

  Mahasiswa tidak menjadi aktor dalam sistem ini. Sebab data yang digunakan sistem ini langsung diambil dari sistem yang sudah ada atau dientri oleh petugas. Sedangkan untuk pelaporan, mahasiswa tidak mengakses langsung sistem ini, melainkan melalui sistem informasi akademik yang sudah ada untuk mengakses data peringatan dini.

- Dosen Wali

  Sama dengan Mahasiswa, Dosen Wali mengakses data peringatan dini hanya melalui sistem informasi akademik. Sedangkan sistem ini mensuplai data untuk sistem akademik.

- Pengelola Jurusan

  Sistem ini secara penggunaan dan manfaat adalah milik dari Pengelola Jurusan dalam hal ini adalah MMT. Sehingga aktor

inilah yang seharusnya menjalankan sistem ini. Tentu dibutuhkan kerjasama yang baik dengan BAPKM.

- Pengelola Fakultas/ Institut

  Pengelola Fakultas/ Institut juga mengakses data peringatan dini melalui sistem informasi akademik.

- Staf BAPKM

  Staf BAPKM, sesuai dengan yang selama ini berlaku, seharusnya menjadi user utama (admin) dalam sistem ini. Namun karena jurusan juga harus bertanggung jawab terhadap datanya masing-masing, maka staf BAPKM dapat bekerja sama dengan Pengelola Jurusan.

4.5.1.2.2. Fungsi

Fungsi utama sistem ini ada 3 :

1. Persiapan data

   Fungsi-fungsi dalam penyiapan data antara lain :

   - Import data, baik melalui transfer data manual maupun otomatis

     Data yang belum ada akan ditambahkan ke dalam basis data sistem. Untuk data yang sudah ada di dalam sistem, data pada atribut Status_Kelulusan akan diperbaharui. Program akan secara otomatis membedakan antara data uji coba dengan data yang akan dikenakan klasifikasi.

   - Pengelompokan data

     Fungsi ini berisi pilihan berbagai macam cara pengelompokan data termasuk pengelompokan pembagian 5 kategori dan pengelompokan pembulatan.

2. Proses klasifikasi

   Fungsi-fungsi dalam proses klasifikasi antara lain :

   - Pembuatan model klasifikasi

Fungsi ini akan memanfaatkan data uji coba untuk menghitung ke-3 tingkat prediksi yang akan diulang prosesnya dalam 3 tahapan berbeda dan dengan masing-masing 4 percobaan berbeda. Program akan otomatis membandingkan semua percobaan dan mengusulkan model klasifikasi terbaik. Namun pengguna dapat memilih model klasifikasi lain yang dipandang lebih baik.

- Pengklasifikasian Dropout

  Menggunakan model klasifikasi yang sudah dibuat, akan dilakukan pengklasifikasian terhadap data yang perlu diklasifikasi. Fungsi ini menghasilkan daftar mahasiswa yang terindikasi dropout.

3. Hasil klasifikasi
   - Kirim ke tabel Peringatan

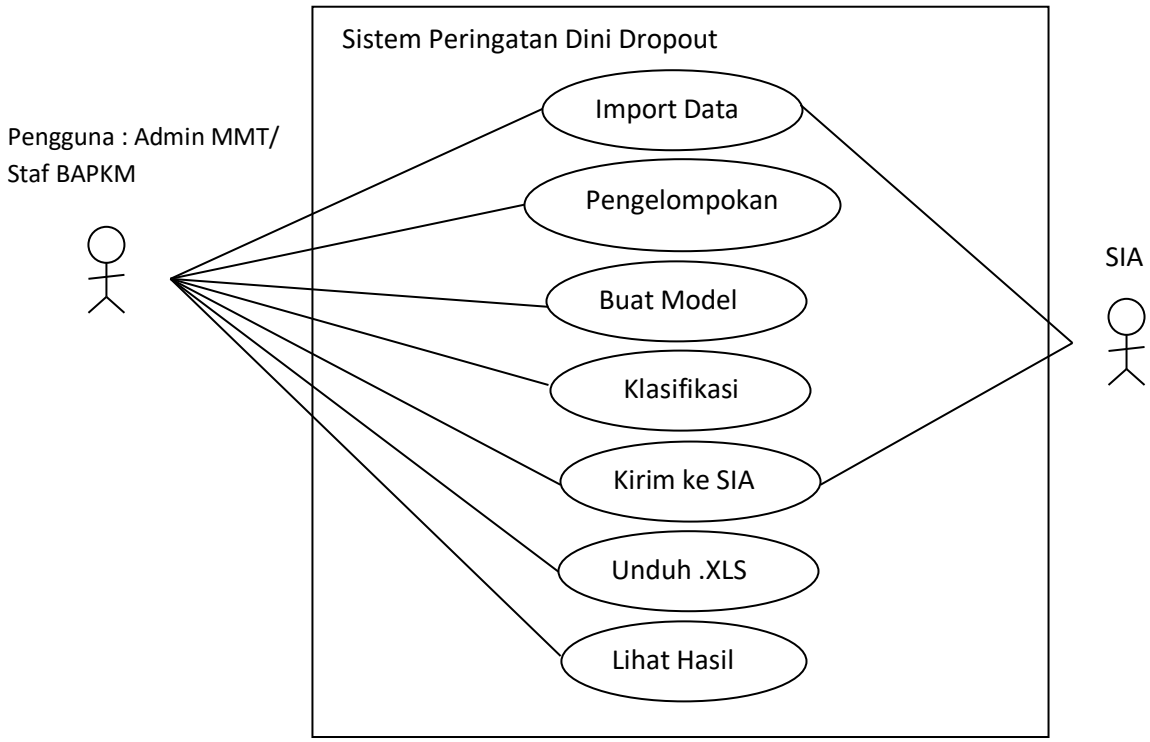
     Fungsi ini akan mengirimkan hasil klasifikasi ke tabel Peringatan agar dapat diakses melalui Sistem Informasi Akademik yang sudah ada sebelumnya.

   - Unduh dalam format Excel

     Fungsi ini akan mengunduh hasil klasifikasi ke dalam format excel terutama untuk keperluan pengolahan lebih lanjut.

4.5.1.2.3. Use Case

Use Case untuk aktor dan fungsi yang telah diuraikan di atas adalah sebagai berikut :



Gambar 4.6 Diagram Use Case Sistem Peringatan Dini Dropout

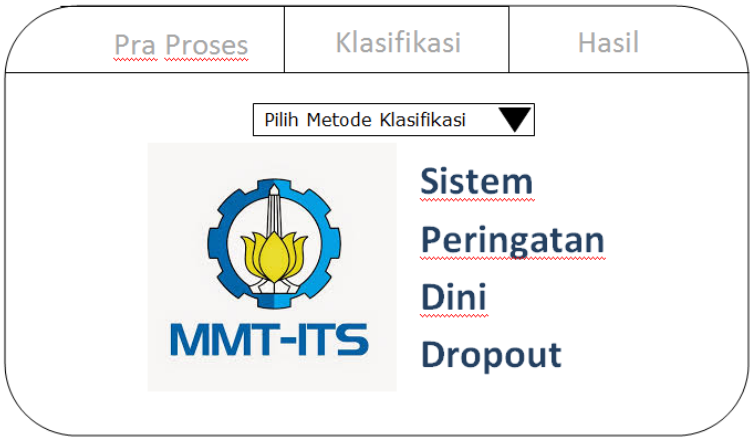
SIA (Sistem Informasi Akademik ITS) berhubungan dengan sistem pada use case Import Data dan use case Kirim ke SIA. Meskipun sebuah sistem, SIA adalah termasuk aktor pada Sistem Peringatan Dini Dropout. Sedangkan pengguna sistem ini adalah admin MMT atau staf BAPKM. Pengguna dapat mengakses semua use case.

4.5.2. Mockup Sistem

Berdasarkan uraian kebutuhan sistem, mockup sistem dirancang sedemikian sehingga nantinya akan dapat digunakan sebagai acuan pengembangan Sistem Peringatan Dini Dropout. Dalam pembahasan fungsi di atas, terdapat 3 fungsi utama. Fungsi utama ini diterjemahkan sebagai 3 tab dalam tampilan antarmuka pengguna. Yaitu tab Pra Proses, tab Klasifikasi dan tab Hasil.

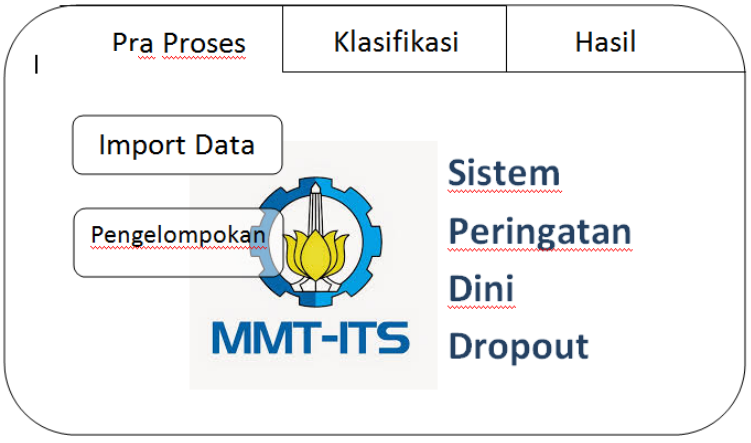
Sebelum memasuki tab-tab tersebut, ada fungsi pilihan metode klasifikasi dalam bentuk *drop down list*. Hal ini didasarkan pada hasil penelitian yang menyatakan bahwa sebaiknya tidak tergantung pada satu metode klasifikasi untuk kasus dropout (Knowles, 2015). Dalam penelitian kali ini, fungsi ini berisi *default* yaitu metode klasifikasi Naive Bayes.

Gambar 4.7 memperlihatkan visualisasi fungsi pemilihan metode klasifikasi.



Gambar 4.7 Fungsi Pemilihan Metode Klasifikasi

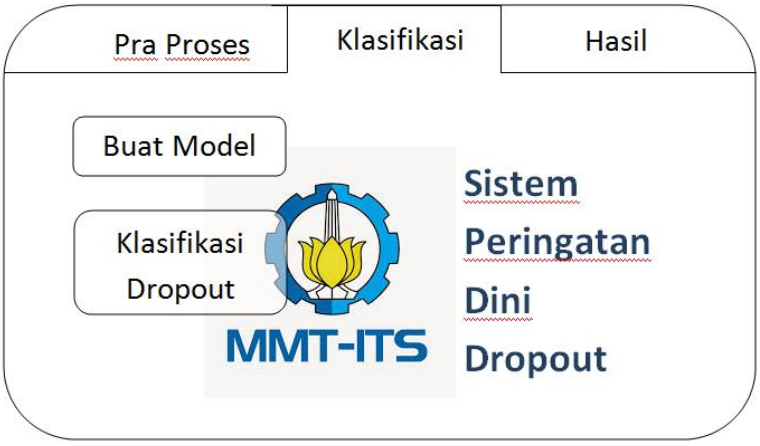
Gambar 4.8 menjelaskan visualisasi tab Pra Proses yang di dalamnya terdapat 2 fungsi. Pertama, fungsi Import Data. Fungsi ini akan mengakomodasi import data baik dari Sistem Informasi Akademik melalui jaringan internet maupun dari penyimpanan fisik. Alternatif import data dari tempat penyimpanan fisik dibuat agar sistem dapat tetap digunakan meskipun SIA mengalami gangguan. Kedua, fungsi Pengelompokan. Fungsi ini mengakomodasi pengelompokan data baik pengelompokan pembagian ke dalam kategori maupun pengelompokan pembulatan.



Gambar 4.8 Tab Pra Proses

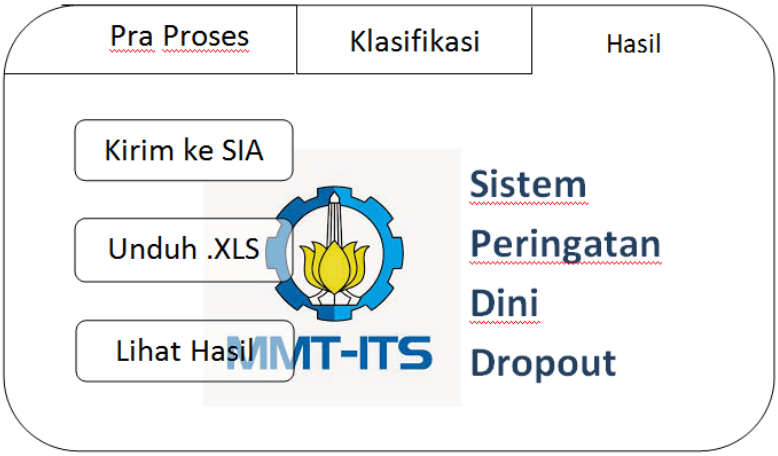
Gambar 4.9 menjelaskan visualisasi tab Klasifikasi yang di dalamnya terdapat 2 fungsi juga. Pertama, fungsi Buat Model. Fungsi ini akan melakukan uji coba persis seperti yang dilakukan dalam penelitian ini. Lalu pengguna memilih model yang terbaik yang cocok. Kedua, fungsi

Klasifikasi Dropout. Fungsi ini melakukan klasifikasi berdasarkan model yang telah dipilih.



Gambar 4.9 Tab Klasifikasi

Gambar 4.10 merupakan visualisasi tab Hasil yang terdiri dari 3 fungsi. Pertama, fungsi Kirim ke SIA. Fungsi ini akan mengirimkan hasi klasifikasi ke Sistem Informasi Akademik yaitu ke table Peringatan. Kedua, fungsi Unduh .XLS. Fungsi ini memungkinkan pengguna untuk mengunduh informasi hasil klasifikasi dalam format Ms Excel yang mungkin digunakan untuk bahan rapat dan koordinasi lainnya di luar sistem informasi ITS. Ketiga, fungsi Lihat Hasil. Fungsi ini memungkinkan pengguna untuk melihat hasil klasifikasi di layar monitor.



Gambar 4.10 Tab Hasil Klasifikasi

(Halaman Sengaja Dikosongkan)

BAB 5

PENUTUP

Dalam bab ini diuraikan beberapa hal yang merupakan kesimpulan dari penelitian Perancangan Sistem Peringatan Dini *Dropout* dengan Menggunakan Metode Klasifikasi Naive Bayes yang dibuat setelah dilakukan pengujian dan analisa. Selain itu juga diberikan saran-saran yang mendukung dalam upaya untuk lebih membawa kepada penyempurnaan penelitian-penelitian ke depannya.

5.1. Kesimpulan

- Klasifikasi Naive Bayes dapat digunakan untuk membangun Sistem Peringatan Dini *Dropout* berdasarkan penelitian dengan tingkat prediksi yang cukup tinggi meskipun di awal uji coba dimulai dengan tingkat prediksi yang kurang memuaskan. Tingkat akurasi dapat mencapai 92% dengan tingkat sensitifitas mencapai 94,1%. Dan yang paling mengejutkan, meski jumlah data *dropout* hanya sebesar 15%, model klasifikasi Naive Bayes mampu mencapai tingkat spesifisitas 80%. Capaian ini didapatkan dari uji coba Tahap II pada Percobaan 4 dengan atribut hasil pembulatan dan atribut hasil pembagian 5 kategori digunakan bersama-sama sebagai atribut prediktor.
- Atribut IPS semester 1 dan IPS semester 2 beserta atribut turunannya secara signifikan dapat meningkatkan akurasi, sensitifitas dan spesifisitas model klasifikasi Naive Bayes. Peningkatan ini dapat terjadi pada penambahan masing-masing atribut hasil pembulatan maupun pembagian 5 kategori, apalagi bila digabungkan. Hal ini wajar mengingat di dalam nilai tersebut terdapat pula unsur kehadiran dan aktifitas mahasiswa yang bersangkutan di dalam proses perkuliahan.
- Model klasifikasi Naive Bayes dapat digunakan untuk prediksi dropout dengan menggunakan data mahasiswa sebelum masuk MMT

ITS sampai data dari tes masuk (uji coba Tahap I) meskipun tingkat spesifisitas hanya sampai 20% (Percobaan 3). Artinya, pada awal masuk sebelum diterima menjadi mahasiswa, calon mahasiswa yang berpotensi dropout sebagian dapat terdeteksi meski tidak banyak yaitu sekitar 20 %.

- Variabel prediktor apapun yang dapat dipakai asal memiliki korelasi atribut lebih dari 0, mungkin untuk digunakan. Ukurannya berbeda dengan korelasi biasa. Korelasi atribut nilainya antara 0 dan 1 yang menunjukkan pengaruh dalam meningkatkan hasil prediksi.
- Pengelompokan data suatu atribut baik melalui pembulatan maupun pembagian ke dalam kategori dalam hal ini 5 kategori, dapat menghasilkan atribut baru. Hasil pengelompokan data tersebut, baik diletakkan dalam atribut baru ataupun menggantikan isi atribut lama, bisa jadi menambah atau malah mengurangi tingkat korelasi atribut terhadap variabel indikator.
- Sebuah atribut yang merupakan turunan dari atribut lain, dapat digabungkan penggunaannya. Dan bila mempunyai nilai korelasi atribut yang baik dengan kelas yang diprediksi, akan dapat meningkatkan hasil prediksi secara signifikan.
- Berdasarkan perbandingan dengan penelitian lain, Naive Bayes dapat bersaing dengan Decision Tree (algoritma C4.5) meskipun variabel pokok yang digunakan lebih sedikit.

5.2. Saran

- Berdasarkan uraian penelitian Djared Knowles pada tahun 2015 yang menekankan agar tidak bergantung pada satu metode terbaik sekalipun, untuk pengembangan Sistem Peringatan Dini Dropout yang lebih baik, perlu juga dilakukan penelitian sejenis dengan menggunakan metode lainnya. Tentunya penelitian tersebut tidak

hanya sekedar mengganti metode, melainkan dengan pengembangan metodologi yang lebih baik lagi.

- Pencarian terhadap atribut atau variabel prediktor yang memiliki pengaruh yang lebih baik lagi untuk kasus *dropout* masih dapat dilakukan mengingat tingkat korelasi atribut dari variabel-variabel prediktor yang digunakan masih dibawah 0,3 dari skala 0 sampai 1.
- Penelitian ini memanfaatkan perhitungan tingkat korelasi atribut yang telah disediakan oleh WEKA untuk mengoptimalkan metode klasifikasi Naive Bayes pada kasus *dropout* tanpa penekanan pada penanganan multikolinieritas. Untuk itu, perlu penelitian lebih lanjut untuk melihat efektifitas penanganan multikolinieritas terhadap tingkat prediksi. Terutama pada metode klasifikasi Naive Bayes yang mengasumsikan setiap atribut prediktornya adalah independen.

(Halaman Sengaja Dikosongkan)

DAFTAR PUSTAKA

Berry, M. J. A. and Linoff, G., (2000), Mastering Data Mining : The Art and Science of Customer Relationship Management, Wiley Computer Publishing, Canada.

Han, J. and Kamber, M., (2006). Data Mining: Concepts and Techniques, Elsevier.

Halim, A, (2015), Identifikasi Mahasiswa yang mempunyai Kecenderungan Lulus tidak Tepat Waktu pada Program Studi MMT ITS dengan Menggunakan Algoritma C4.5, Tesis Magister, Institut Teknologi Sepuluh Nopember, Surabaya.

Han, J and Kamber, M., (2006), Data Mining Concepts and Techniques, second edition. California : Morgan Kaufman.

Hidayat, M. M., (2013), Analisis Kemungkinan Drop Out Berdasarkan Perilaku Sosial Mahasiswa dalam Educational Data Mining Menggunakan Jaringan Syaraf Tiruan sebagai Classifier, Tesis Magister, Institut Teknologi Sepuluh Nopember, Surabaya.

Hilmiyah, F., (2017), Prediksi Kinerja Mahasiswa Menggunakan Support Vector Machine untuk Pengelola Program Studi di Perguruan Tinggi (Studi

Kasus: Program Studi Magister Statistika ITS), Tesis Magister, Institut Teknologi Sepuluh Nopember, Surabaya.

JCGM 200:2008 *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)* hal 35.

Khoirunnisak, M., (2010), Pemodelan Faktor-faktor yang Mempengaruhi Mahasiswa Berhenti Studi (Dropout) di Institut Teknologi Sepuluh Nopember Menggunakan Analisis *Bayesian Mixture Survival*, Tugas Akhir, Institut Teknologi Sepuluh Nopember, Surabaya.

Knowles, J. E., (2015), *Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin, Journal of Educational Data Mining*, Volume 7, No 3.

Natalius, S., (2010), Metoda *Naïve Bayes Classifier* dan Penggunaannya pada Klasifikasi Dokumen, Jurnal Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, Volume 17.

Phyu, Thair N., (2009), Survey of Classification Techniques in Data Mining, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

Taylor, J. R., (1999), *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, University Science Books, pp. 128–129. ISBN 0-935702-75-X.

Witten, I. H., Frank, E., Hall, M. A., (2011), Data Mining Practical Machine Learning Tools and Techniques (3rd ed), USA: Elsevier.

Xhemali, D., Hinde, C.J., Stone, R. G., (2009), Naive Bayes vs. Decision Tree vs. Neural Networks in the Classification of Training Web Pages, International Journal of Computer Science Issue, Vol. 4, No. 1, pp. 16-23.

(Halaman Sengaja Dikosongkan)

BIOGRAFI PENULIS

Nama : Maks Agustinus

Tempat/ Tahun Lahir : Surabaya/ 1974

Email : max_c08@yahoo.com

Penulis adalah mahasiswa yang berasal dari Kota Surabaya, JawaTimur. Penulis merupakan anak kedua dari lima bersaudara. Penulis menamatkan pendidikan dasar di SDN Pakis 374 Surabaya pada tahun 1986, menamatkan pendidikan Sekolah Menengah Pertama di SMPN 10 Surabaya pada tahun 1989 dan menamatkan pendidikan Sekolah Menengah Atas di SMAN 12 Surabaya pada tahun 1992. Kemudian Penulis melanjutkan studi ke jenjang sarjana di Jurusan Teknik Computer (Teknik Informatika) ITS dan lulus pada tahun 2000. Setelah itu, Penulis melanjutkan pendidikan ke jenjang Magister (S2) pada tahun 2016 di Departemen Manajemen Teknologi Fakultas Bisnis dan Manajemen Teknologi Institut Teknologi Sepuluh Nopember Surabaya (MMT-ITS) dengan mengambil bidang konsentrasi/ keahlian Manajemen Teknologi Informasi. Penulis memiliki ketertarikan pada bidang data mining dan perancangan sistem.

Setelah menamatkan pendidikan S1 pada tahun 2000, Penulis mulai bekerja di Universitas Kristen Petra Surabaya sebagai Programmer pada tahun 2001 dan masih aktif sampai biografi ini dibuat. Penulis sempat menjadi dosen luar biasa di Jurusan Manajemen Marketing Universitas Kristen Petra Surabaya pada mata kuliah Komputasi Bisnis pada tahun 2007. Kemudian dari tahun 2009-2014, Penulis dipercaya mengampu mata kuliah Sistem Informasi Manajemen di Jurusan Manajemen Bisnis Universitas Kristen Petra Surabaya. Selain itu, pada tahun 2014, Penulis sempat menjadi dosen pengganti pada mata kuliah Manajemen Strategi di jurusan yang sama.