KAJIAN PERBANDINGAN TEKNIK KLASIFIKASI ALGORITMA C4.5, *NAIVE BAYES* DAN *CART* UNTUK PREDIKSI KELULUSAN MAHASISWA DENGAN METODE CRISP-DM

(STUDI KASUS : STMIK ROSMA KARAWANG)

TESIS

Disusun sebagai salah satu syarat
untuk memperoleh gelar Magister Komputer
dari Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI

Oleh:

PRIATI

NPM: 2014210040

PROGRAM PASCASARJANA
MAGISTER SISTEM INFORMASI
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER
LIKMI
BANDUNG

2016

KAJIAN PERBANDINGAN TEKNIK KLASIFIKASI ALGORITMA C4.5,
***NAIVE BAYES* DAN *CART* UNTUK PREDIKSI KELULUSAN**
MAHASISWA DENGAN METODE CRISP-DM
(STUDI KASUS : STMIK ROSMA KARAWANG)

Oleh :

PRIATI

NPM: 2014210040

Bandung, Desember 2015
Menyetujui,

Dr. Djajasukma Tjahjadi, S.E., M.T.
Pembimbing

PROGRAM PASCASARJANA
MAGISTER SISTEM INFORMASI
SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER LIKMI
BANDUNG

2016

ABSTRAK

KAJIAN PERBANDINGAN TEKNIK KLASIFIKASI ALGORITMA C4.5, *NAÏVE BAYES* DAN *CART* UNTUK PREDIKSI KELULUSAN MAHASISWA DENGAN METODE CRISP-DM

(STUDI KASUS : STMIK ROSMA KARAWANG)

PRIATI

NPM : 2014210040

Program Studi Magister Sistem Informasi Bisnis

Program Pascasarjana Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI, Bandung, 2015

Pembimbing : Dr. Djajasukma Tjahjadi, S.E., M.M

Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Rosma merupakan salah satu perguruan tinggi swasta di Kab. Karawang. Sejak berdirinya hingga sekarang STMIK Rosma telah menghasilkan jumlah kelulusan mahasiswa yang banyak. Jumlah lulusan yang banyak ini akan sangat disayangkan apabila tidak diteliti lebih lanjut untuk mendapatkan informasi dan pengetahuan baru.

Beberapa penelitian tentang kelulusan mahasiswa telah banyak dilakukan. Dalam penelitian ini dilakukan komparasi algoritma C4.5, *naïve bayes*, dan *CART* yang diaplikasikan terhadap data mahasiswa STMIK Rosma tahun 2005 sampai 2008 untuk jenjang Diploma III dan Strata 1.

Dari hasil pengujian dengan mengukur kinerja dari ketiga algoritma tersebut menggunakan metode pengujian *Confusion Matrix* dan kurva *ROC*, diketahui bahwa algoritma C4.5 dan algoritma *CART* memiliki nilai *accuracy* yang sama tinggi yaitu 95,6012% serta paling rendah adalah *accuracy* algoritma *naïve bayes* sebesar 89,5894%. Nilai AUC untuk algoritma *naïve bayes* menunjukkan nilai tertinggi yaitu 0,97 disusul algoritma C4.5 dengan nilai AUC 0,923 dan algoritma CART dengan nilai AUC 0,922. Melihat nilai AUC dari ketiga algoritma tersebut, maka semua algoritma termasuk kelompok klasifikasi yang sangat baik, karena nilai AUC-nya antara 0,90-1,00.

Kata kunci: kelulusan, algoritma C4.5, *Naïve Bayes, CART, Confusion Matrix,* Kurva *ROC*

ABSTRACT

COMPARATIVE STUDY OF CLASSIFICATION ALGORITHM C4.5, NAÏVE BAYES AND CART FOR PREDICTION OF STUDENTS GRADUATION

WITH CRISP-DM METHODE

(CASE STUDY: STMIK Rosma KARAWANG)

PRIATI

NPM: 2014210040

Master of Business Information Systems

Postgraduate Program of College of Informatic and Computer Management LIKMI, Bandung, 2015

Supervisor: Dr. Djajasukma Tjahjadi, S.E., M.M

College of Informatics and Computer Management (STMIK) Rosma is one of the private universities in Karawang. Since its establishment until now STMIK Rosma has produced a number of graduate students. The number of graduates that many of these would be very unfortunate if it is not investigated further to gain new information and knowledge.

Several studies of graduation has been done. In this study a comparison algorithm C4.5, naïve bayes, and CART that applied to the student data STMIK Rosma 2005 to 2008 for Diploma III and bachelor degree.

From the test results to measure the performance of the three algorithms using testing methods Confusion Matrix and the ROC curve, it is known that the algorithm C4.5 and CART algorithms have the same accuracy values as high as 95.6012% and the lowest was naïve bayes algorithm at 89, 5894%. AUC values for naïve bayes algorithm shows the highest value is 0.97 followed by the algorithm C4.5 with AUC value of 0.923 and the CART algorithm are 0.922. Seeing the AUC value of the third algorithm, all algorithms including the classification of a very good group, because of its AUC values between 0.90 to 1.00.

Keywords: graduation, C4.5 algorithm, Naïve Bayes, CART, Confusion Matrix, ROC Curve

KATA PENGANTAR

Puji dan syukur kepada Allah SWT, Tuhan semesta alam, berkat rahmat dan karunia-Nya, tesis yang berjudul "KAJIAN PERBANDINGAN TEKNIK KLASIFIKASI ALGORITMA C4.5, *NAÏVE BAYES, CART* UNTUK PREDIKSI KELULUSAN MAHASISWA DENGAN METODE CRISP-DM (STUDI KASUS : STMIK ROSMA KARAWANG)", dapat terselesaikan.

Penelitian ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister Sistem Informasi Bisnis dari Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI, Bandung.

Peneliti menyadari bahwa dalam penyusunan tesis ini masih terdapat banyak kekurangan yang perlu ditambahkan. Oleh karena itu peneliti mengharapkan kritik dan saran yang membangun sehingga penelitian ini akan lebih berkembang.

Peneliti mengucapkan banyak terima kasih kepada semua pihak yang telah memberikan dukungan moril maupun materiil, terutama kepada:

1. Yth. Dr. Djajasukma Tjahjadi, S.E., M.T. sebagai Pembimbing Tesis, Wakil Ketua Bidang Akademik dan Kemahasiswaan Sekolah Tinggi Manajemen Informatika LIKMI, Bandung, yang telah meluangkan waktu, memberikan bimbingan, memberikan inspirasi dan motivasi, dalam penyelesaian tesis ini;
2. Yth. Bpk. Darmasyah, M.Kom Ketua STMIK Rosma Karawang, yang telah memberikan ijin dalam penelitian dan penulisan tesis ini.
3. Yth. Dr. Budi Permana, S.E., Ak., M.Sc. Ketua Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI, Bandung;
4. Yth. Dr. H. Ana Hadiana Ketua Program Studi Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI, Bandung;
5. Yth. Bpk Yudiana, M.Kom Wakik Ketua I STMIK Rosma Karawang, yang telah membantu dalam pengumpulan data mahasiswa.

6. Seluruh staf STMIK Rosma Karawang yang telah memberikan berbagai dukungan dalam penyelesaian tesis ini;
7. Kedua orang tua, Bapak dan Ibu, yang tanpa lelah memberikan doa yang tulus serta restu yang tak terkirakan;
8. Suami dan kedua anakku tercinta yang turut ikut merasakan perjuangan dalam penyusunan dan penyelesaian tesis ini;
9. Rekan-rekan se-almamater yang senantiasa men-*support* dan banyak memberikan masukan dalam penyusunan tesis ini;
10. Rekan – rekan Universitas Singaperbangsa Karawang dan Universitas Buana Perjuangan Karawang yang telah meluangkan waktu untuk berdiskusi dengan peneliti tentang penyusunan tesis ini;

Serta semua pihak yang tidak dapat disebutkan satu persatu mulai dari penyusunan proposal sampai dengan penyajian tesis ini.

Bandung, November 2015

Penulis,

Priati

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR RUMUS

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Teknologi informasi sudah berkembang pesat di berbagai aspek kehidupan manusia, baik di aspek industri, ekonomi, sosial dan berbagai aspek lainnya termasuk pada aspek pendidikan, khususnya pendidikan tinggi. Hal ini menyebabkan banyak data mengenai mahasiswa yang dihasilkan.

Pertumbuhan yang pesat dari akumulasi data atau informasi itu telah menciptakan kondisi dimana suatu institusi memiliki bergunung-gunung data tetapi miskin informasi yang bermanfaat. Tidak jarang “gunung” data itu dibiarkan begitu saja seakan-akan menjadi “kuburan data”. Pertanyaannya sekarang, apakah gunung data tersebut akan dibiarkan, tidak berguna lalu dibuang, ataukah dapat ditambang untuk menemukan “emas”, yaitu informasi yang lebih bermanfaat? Jawabannya ya, data mining hadir untuk menjawab tantangan tersebut.

Data mahasiswa yang semakin banyak jumlahnya dari waktu ke waktu sangat disayangkan apabila tidak dianalisa. Apabila data yang banyak tersebut dimanfaatkan dengan menganalisa pola apa yang ada, pihak kampus akan mendapatkan informasi untuk pengambilan keputusan.

Pada STMIK Rosma Karawang terdapat banyak data mahasiswa yang ada pada Pangkalan Data Pendidikan Tinggi (PDDIKTI) merupakan kumpulan data penyelenggaraan Pendidikan Tinggi seluruh Perguruan Tinggi yang terintegrasi secara nasional. PDDIKTI menjadi salah satu *instrument* pelaksanaan penjaminan mutu. Dalam pasal 56 ayat 2 UU No. 12 Tahun 2012 tentang Pendidikan Tinggi menyebutkan bahwa Pangkalan Data Pendidikan Tinggi sebagaimana dimaksud pada ayat (1) berfungsi sebagai sumber informasi bagi Lembaga Akreditasi, Pemerintah dan Masyarakat. Bagi Lembaga Akreditasi, Pangkalan Data Perguruan Tinggi berfungsi sebagai sumber informasi untuk melakukan akreditasi Program Studi dan Perguruan Tinggi, sedangkan bagi Pemerintah berfungsi sebagai sumber informasi untuk melakukan pengaturan,

perencanaan, pengawasan, pemantauan dan evaluasi serta pembinaan dan koordinasi Program Studi dan Perguruan Tinggi. Dan bagi masyarakat berguna untuk mengetahui kinerja Program Studi dan Perguruan Tinggi.

"Berdasarkan berlimpahnya data mahasiswa dan data jumlah kelulusan mahasiswa, informasi yang tersembunyi dapat diketahui dengan cara melakukan pengolahan terhadap *data tersebut sehingga berguna bagi pihak universitas"* (Oscar Ong, 2013).

Permasalahan yang sering terjadi adalah masih banyaknya jumlah mahasiswa yang lulus dengan lama studi melampaui waktu yang telah ditetapkan dengan memperoleh Indeks Prestasi Kumulatif (IPK) yang relatif rendah, sehingga dapat mempengaruhi mutu lulusan perguruan tinggi. Seiring dengan terus bertambahnya jumlah mahasiswa di STMIK Rosma Karawang maka jumlah data kemahasiswaan terus meningkat pula sehingga menyebabkan terjadinya penumpukan data yang belum diolah dengan optimal untuk menggali informasi dan pengetahuan baru yang dapat digunakan sebagai bahan pertimbangan pemimpin STMIK dalam proses pengambilan kebijakan dan keputusan. Terutama untuk memprediksi tingkat kelulusan mahasiswa.

Sejak berdirinya dalam rentang waktu tahun 2000 sampai sekarang tahun 2015 STMIK Rosma Karawang telah menghasilkan gudang data mahasiswa yang ada pada pangkalan data perguruan tinggi dengan jumlah data mahasiswa sebanyak 15246 *record*.

Klasifikasi data mahasiswa pada bidang pendidikan merupakan tugas penting dalam memprediksi kelulusan bahkan dapat membantu pihak perguruan tinggi dalam mengambil keputusan atau kebijakan. Dengan demikian sangat penting dalam memprediksi kelulusan secara dini untuk meningkatkan jumlah kelulusan mahasiswa dalam tahun-tahun berikutnya.

Banyak penelitian sudah dilakukan dalam prediksi kelulusan namun belum ada yang membandingkan hasil dari ketiga algoritma (C4.5, *naïve bayes* dan *CART*). Ketiga algoritma tersebut digunakan dalam memprediksi kelulusan mahasiswa dengan tujuan agar algoritma terpilih merupakan algoritma yang paling akurat sehingga dapat

melakukan prediksi kelulusan mahasiswa secara dini. Ketiga algoritma tersebut termasuk dalam sepuluh klasifikasi *data mining* yang paling popular (Wu & Kumar, 2009).

Saat ini penelitian terhadap *data mining* dan sistem pendidikan semakin banyak dilakukan. Penelitian mengenai *data mining* di dunia pendidikan telah lama ada (sejak tahun 1990an) dan baru dikelompokkan menjadi sebuah bidang penelitian *Educational Data mining* pada tahun 2005.

Mulai tahun 2008, organisasi ini mengadakan konferensi tahunan EDM yang membahas penelitian-penelitian *data mining* di dunia pendidikan di seluruh dunia (Rahmayuni, 2014).

Beberapa penelitian yang menggunakan teknik data mining pada data set akademik dan kemahasiswaan telah banyak dilakukan, antara lain adalah penelitian yang dilakukan oleh Ridwan, dkk. 2013., Septian, 2009., Sulistyo, 2014., Hartanto, 2014., Al-Radaideh, dkk, 2006., Jananto, 2010., Sunjana, 2010. Perbedaan penelitian ini dengan penelitian sebelumnya adalah pada data set yang digunakan adalah Pangkalan Data Pendidikan Tinggi (PDDIKTI) dan data dari bagian akademik STMIK Rosma Karawang, serta algoritma yang digunakan.

Dengan permasalahan diatas maka penulis tertarik untuk melakukan penelitian yang berjudul "KAJIAN PERBANDINGAN TEKNIK KLASIFIKASI ALGORITMA C4.5, NAIVE BAYES, DAN CART UNTUK PREDIKSI KELULUSAN MAHASISWA (STUDI KASUS : STMIK ROSMA KARAWANG)".

1.2 Rumusan Masalah

Pokok permasalahan dirumuskan melalui pertanyaan berikut :

a. Apakah Algoritma C4.5, Naive Bayes dan CART dapat digunakan untuk menentukan prediksi kelulusan mahasiswa STMIK Rosma Karawang?
b. Diantara algoritma C4.5, Naive Bayes dan CART manakah yang terbaik dalam menentukan prediksi kelulusan mahasiswa STMIK Rosma Karawang?
c. Apakah algoritma yang terpilih dapat menampilkan data prediksi hasil data mining kelulusan mahasiswa?

1.3 Tujuan Penelitian

Sesuai dengan rumusan masalah yang telah dikemukakan maka tujuan penelitian adalah sebagai berikut :

a. Membandingkan tingkat akurasi yang dihasilkan oleh teknik atau model data mining yaitu algoritma C4.5, *Naive Bayes* dan *CART* dalam memperkirakan kelulusan mahasiswa STMIK Rosma Karawang.
b. Menjabarkan algoritma C4.5, Naive Bayes dan CART ke dalam rule.
c. Menerapkan algoritma C4.5, *Naive Bayes* dan *CART* dalam melakukan prediksi terhadap kelulusan mahasiswa STMIK Rosma Karawang.

1.4 Ruang Lingkup

Untuk lebih fokus maka dalam penelitian ini hanya dibatasi pada masalah yang terkait dengan algoritma C4.5, *Naive Bayes* dan *CART* menggunakan klasifikasi data mining dengan cara menganalisis sejumlah atribut atau variabel yang menjadi parameter dalam prediksi kelulusan STMIK Rosma Karawang.

Adapun atribut atau variabel yang digunakan adalah Jenjang (JNJNG), Program Studi (PRODI), Tempat Lahir (TPTLHR), Jenis Kelamin (JNSKLMN), Jumlah SKS yang Sudah ditempuh (JMLSKS), Jumlah SKS Total, Indeks Prestasi Komulatif (IPK), Sekolaj, Kelas, Kerja, IP Semester 1 (IPS1), IP Semester 2 (IPS2), IP Semester 3 (IPS3), IP Semester 4 (IPS4), Tanggal Kelulusan (KELULUSAN).

Penelitian ini menggunakan perangkat lunak Weka *(Weikato Environment Knowledge and Analysis)* versi 3.6.1 yang merupakan aplikasi *data mining* berbasis *open source* (General Public Lisence) dan ber-*engine* Java dengan *Graphical User Interface* (GUI) menggunakan Java.

1.5 Manfaat Penelitian

Penelitian ini bermanfaat bagi pihak STMIK Rosma Karawang sebagai bahan pertimbangan dalam proses pengambilan keputusan dan kebijakan terutama dalam memprediksi tingkat dan kualitas kelulusan mahasiswa.

1.6 Sistematika Penulisan

Untuk mempermudah dalam penyusunan penelitian ini maka dibuat sistematika penulisan sebagai berikut:

BAB I : PENDAHULUAN

Bab ini berisi mengenai Latar Belakang, Rumusan Masalah, Tujuan Penelitian, Ruang Lingkup, Manfaat Penelitian dan Sistematika Penulisan.

BAB II : LANDASAN TEORI

Bab ini berisi mengenai teori-teori yang digunakan dalam menganalisis hasil penelitian.

BAB III : OBYEK DAN METODE PENELITIAN

Penelitian ini menggunakan CRISP-DM *(Cross Industry Standard Process for Data Mining)* dengan tahapan *Business/Research Understanding Phase*, *Data Understanding Phase* (Fase Pemahaman Data), *Data Preparation Phase* (Fase Pengolahan Data) *Modeling Phase* (Fase Pemodelan), *Evaluation Phase* (Fase Evaluasi) dan *Deployment Phase* (Fase Penyebaran)..

BAB IV : HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi mengenai pembahasan dan pengujian variabel yang telah ditentukan dengan penerapan teknik klasifikasi (ALGORITMA C4.5, *NAIVE BAYES* dan *CART*) menggunakan alat bantu WEKA 3.6.1 .

BAB V : KESIMPULAN DAN SARAN

Bab ini berisi tentang Kesimpulan dan Saran dari hasil penelitian.

BAB II

LANDASAN TEORI

2.1 Pengertian Data Mining

"Data Mining didefinisikan sebagai sebuah proses untuk menentukan hubungan pola dan tren baru yang bermakna dengan menyaring, menggunakan teknik pengenalan pola seperti teknik statistik dan matematika" (Larose, 2005).

"Data Mining merupakan sebuah analisa dari observasi data dalam jumlah besar untuk menentukan hubungan yang tidak diketahui sebelumnya dan dua metode baru untuk meringkas data agar mudah dipahami serta kegunaannya untuk memilih data" (Jefri, 2013).

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, *dan machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, dkk., 2005).

Jadi, data mining menurut peneliti adalah cara memperoleh beragam informasi dari banyaknya *data* histori yang dilakukan dengan teknik pengolahan tertentu sehingga mendapatkan pola data yang sesuai dengan domain aplikasi yang diinginkan.

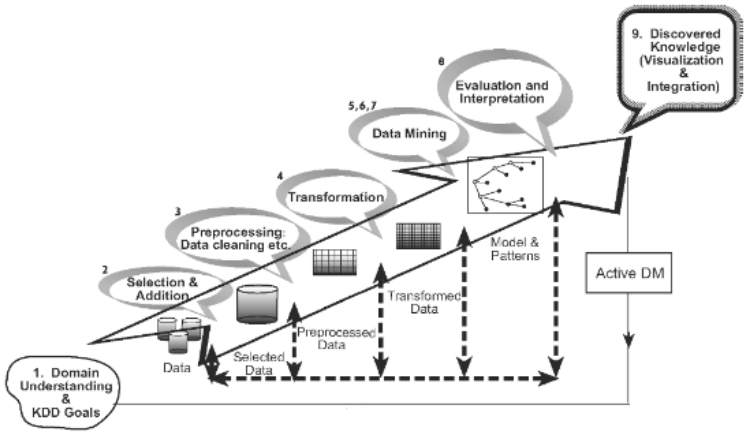
Perkembangan teknologi informasi yang begitu maju saat ini, menyebabkan tingkat akurasi suatu data sangat dibutuhkan dalam kehidupan sehari-hari. Setiap informasi yang ada menjadi suatu hal penting untuk menentukan setiap keputusan dalam situasi tertentu. Hal ini menyebabkan penyediaan informasi menjadi sarana untuk dianalisa dan diringkas menjadi suatu pengetahuan dari data yang bermanfaat ketika pengambilan suatu keputusan dilakukan (Huda, 2010).

Pada penelitian (Ridwan, dkk., 2013) menjelaskan bahwa faktor yang paling berpengaruh dalam penentuan klasifikasi kinerja akademik mahasiswa adalah Indeks Prestasi Komulatif (IPK), Indeks Prestasi Semester (IPS) semester 1, IPS semester 4 dan jenis kelamin. Pada penelitian ini peneliti menggunakan algoritma C4.5 dalam menentukan prediksi kelulusan berdasarkan atribut atau variabel atau variabel jenis

kelamin, asal sekolah SMA dan IPS semester satu sampai dengan semester enam. (Ridwan, dkk., 2013)

Data mining merupakan inti dari proses *Knowledge Discovery in Database* (KDD) (Mai, 2005). KDD adalah proses terorganisir untuk mengidentifikasi pola yang valid, baru, berguna, dan dapat dimengerti dari sebuah data set yang besar dan kompleks. Langkah-langkah dalam KDD (Maimon, 2005):

a. Pembentukan pemahaman domain aplikasi. Pada tahap ini menentukan tujuan dari *end-user* dan bagian terkait dimana KDD dilakukan. Mengembangkan pemahaman tentang domain aplikasi ini adalah awal langkah persiapan. Mempersiapkan adegan untuk memahami apa yang harus dilakukan dengan banyak keputusan (tentang transformasi, algoritma, representasi, dll). Orang-orang yang bertanggung jawab atas proyek KDD perlu memahami dan menentukan tujuan dari *end-user* dan lingkungan di mana proses KDD akan berlangsung (termasuk pengetahuan awal yang relevan). Sebagai hasil proses, mungkin ada revisi dan perbaikan dari langkah ini. Memiliki, memahami tujuan KDD, *preprocessing* data dimulai, sebagaimana didefinisikan dalam tiga langkah berikutnya (perhatikan bahwa beberapa metode disini mirip dengan algoritma *Data Mining*, tetapi digunakan dalam konteks *preprocessing*).



Gambar 2.1. Langkah-Langkah KDD (Maimon, 2005)

b. Memilih dan menciptakan satu *dataset* untuk mendukung proses penemuan *knowledge* akan dilakukan. Penentuan data yang akan digunakan untuk proses KDD dilakukan pada tahap ini. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data untuk KDD ke dalam sebuah *dataset*,

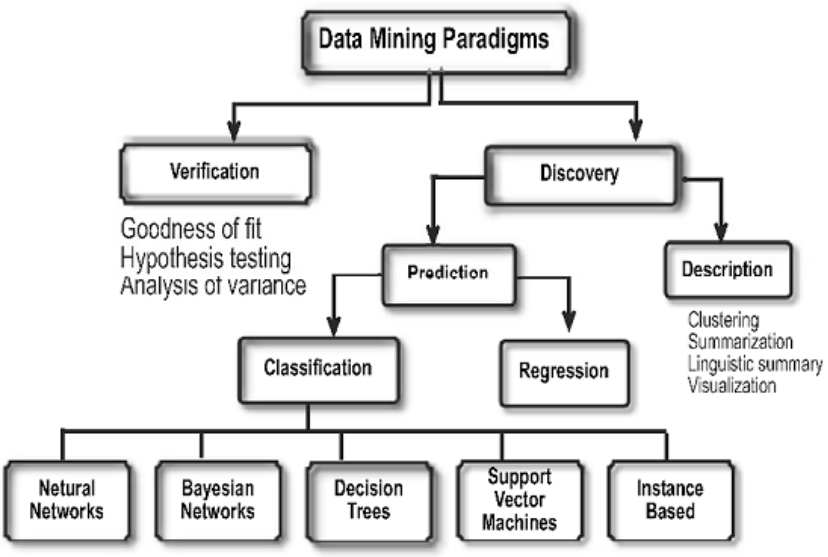
termasuk atribut atau variabel yang diperlukan dalam proses KDD. Terdapat interaktif dan iteratif dari KDD tersebut. Dimulai dengan data yang tersedia baik mengatur dan kemudian mengembang dan mengamati efeknya dalam KDD.

c. *Preprocessing* dan *cleansing*. Dalam tahap ini kehandalan data ditingkatkan. Termasuk *data clearing*, seperti menangani data yang tidak lengkap, menghilangkan gangguan atau *outlier*. Termasuk mengunakan metode statistik yang kompleks, atau melakukan penambangan spesifik data dengan algoritma dalam KDD.
d. Transformasi data. Pada tahap ini, generasi data yang lebih baik untuk *data mining* dipersiapkan dan dikembangkan, membuat data menjadi lebih baik menggunakan metode reduksi dimensi dan transformasi atribut atau variabel. Sebagai contoh, dalam pemeriksaan medis, hasil bagi atribut atau variabel mungkin sering menjadi faktor yang paling penting, dan tidak satu persatu. Di pemasaran, kita mungkin perlu mempertimbangkan efek di luar kendali kita serta upaya dan isu temporal (seperti mempelajari pengaruh akumulasi iklan). Namun, bahkan jika kita tidak menggunakan transformasi yang tepat di awal, kita dapat memperoleh efek mengejutkan bahwa petunjuk kepada kita tentang transformasi diperlukan (Pada iterasi berikutnya). Dengan demikian proses KDD mencerminkan kepada dirinya sendiri dan menyebabkan pemahaman tentang transformasi yang dibutuhkan (seperti pengetahuan ringkas dari sebuah ahli dalam bidang tertentu mengenai indikator terkemuka kunci).
e. Memilih tugas *data mining* yang cocok. Pada tahap ini ditentukan tipe *data mining* yang akan digunakan, apakah klasifikasi, regresi, atau *clustering*, tergantung pada tujuan KDD dan tahap sebelumnya.
f. Memilih algoritma *data mining*. Pemilihan algoritma yang paling tepat untuk menemukan pola dilakukan pada tahap ini. Ada dua tujuan utama dalam *Data Mining* : prediksi dan deskripsi. Prediksi sering disebut sebagai *Supervised Data Mining*, sementara deskriptif *data mining* meliputi aspek-aspek *unsupervised* dan visualisasi *data mining*. Sebagian besar data teknik pertambangan didasarkan pada pembelajaran induktif, dimana model yang dibangun secara eksplisit maupun implisit

oleh generalisasi dari jumlah yang memadai pelatihan data training. Asumsi yang mendasari pendekatan induktif adalah bahwa model terlatih ini berlaku untuk kasus masa depan. Strategi ini juga memperhitungkan tingkat *metalearning* untuk *set* tertentu dari data yang tersedia.

g. Penggunaan algoritma *data mining*. Pada tahap ini dilakukan implementasi dari algoritma *data mining* yang telah ditentukan pada tahap sebelumnya. Misalnya dengan men-*setting* parameter kontrol algoritma, seperti jumlah minimum kasus dalam daun tunggal dari pohon keputusan.
h. Evaluasi, pada tahap ini dilakukan evaluasi dan penerjemahan dari pola yang diperoleh, sehubungan dengan tujuan yang ditetapkan pada langkah pertama. Langkah ini berfokus pada komprehensibilitas dan kegunaan dari model induksi. Pada langkah ini pengetahuan ditemukan juga terdokumentasi untuk penggunaan lebih lanjut. Langkah terakhir adalah penggunaan dan umpan balik secara keseluruhan pada pola dan hasil penemuan diperoleh dengan *data mining.*

Penggunaan pengetahuan yang ditemukan yakni memasukkan pengetahuan ke dalam sistem lain untuk ditindaklanjuti. Pengetahuan menjadi aktif dalam arti bahwa kita dapat membuat perubahan ke sistem dan mengukur dampak Sebenarnya keberhasilan langkah ini menentukan efektivitas proses KDD secara keseluruhan. Ada banyak tantangan dalam langkah ini, kehilangan "kondisi laboratorium". Misalnya, pengetahuan itu ditemukan dari sebuah *snapshot* statis tertentu (biasanya sampel) dari data, tapi sekarang data menjadi dinamis. Gambar 2.2 adalah klasifikasi data mining (Maimon, 2005) :



Gambar 2.2. Klasifikasi Data Mining (Maimon, 2005)

Ada banyak metode *data mining* digunakan untuk tujuan yang berbeda dan tujuan pengklasifikasian tersebut untuk membantu dalam memahami berbagai metode, keterkaitan dan pengelompokan. Hal ini berguna untuk membedakan antara dua jenis *Data Mining*: verifikasi berorientasi (sistem memverifikasi hipotesis pengguna) dan penemuan berorientasi (sistem menemukan aturan baru dan pola mandiri). Pada gambar 2.2 menunjukan langkah-langkah dalam proses *data mining*, proses dalam tahapan *data mining* terdiri dari tiga langkah utama (Maimon, 2005), yaitu:

a. *Data Preparation*

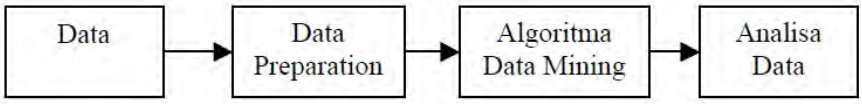
   Pada langkah ini, data dipilih, dibersihkan, dan dilakukan *preprocessed* mengikuti pedoman dan *knowledge* dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh.

b. Penggunaan Algoritma *Data Mining*

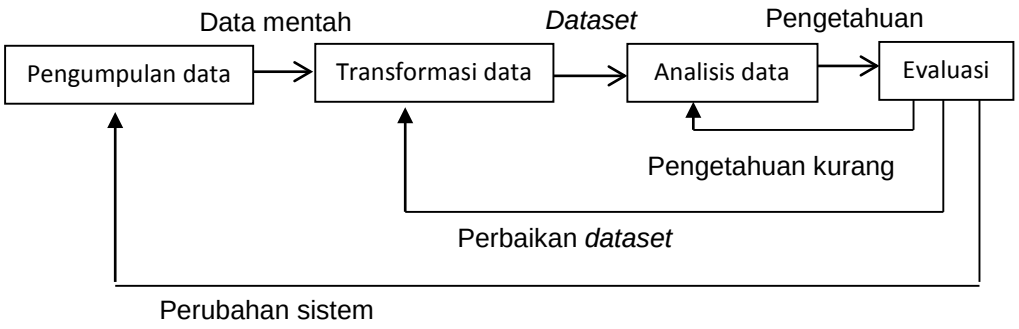
   Pada langkah ini digunakan algoritma *data mining* untuk memudahkan dalam melakukan identifikasi data dan mengintegrasikan keseluruhan data yang sudah ditemukan.

c. Tahap Analisa

   Keluaran dari data mining dievaluasi untuk melihat apakah *knowledge* domain ditemukan dalam bentuk *rule* yang telah diekstrak dari jaringan.



Gambar 2.3. Langkah-langkah dalam Proses *Data Mining* (Maimon, 2005)



Gambar 2.4. Aliran Informasi dalam *Data Mining* (Nilakant, 2004)

Gambar 2.4 ditunjukkan diagram yang menggambarkan aliran informasi dalam proses *data mining*. Proses *data mining* dalam gambar tersebut ditunjukkan sebagai proses yang iterative. Hasil evaluasi pengetahuan yang dihasilkan *data mining* dapat menimbulkan kebutuhan pengetahuan yang lebih lengkap, perbaikan kumpulan data *(dataset)* atau perubahan pada sistem.

2.2 Pengertian Kelulusan

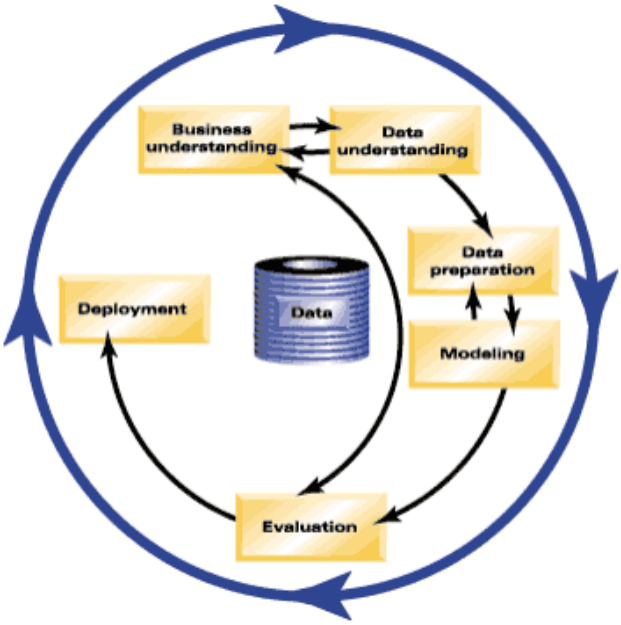
Mahasiswa yang memenuhi persyaratan kelulusan ditetapkan dalam yudisium kelulusan Fakultas atau Program Studi dan ditetapkan dengan keputusan Rektor. Tanggal kelulusan ditetapkan berdasarkan tanggal yudisium dan merupakan tanggal penetapan IPK akhir (Buku Panduan Akademik STMIK Rosma, 2010).

2.3 Metode *Data Mining*

Cross-Industry Standard Process for Data Mining (CRISP-DM) dikembangkan tahun 1996 oleh analisis dari beberapa industri seperti Daimler Chrysler, SPSS dan NCR. CRISP-DM menyediakan standar proses *data mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian.

Dalam CRISP-DM sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase seperti pada Gambar 2.5. Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Fase berikutnya dalam urutan bergantung kepada keluaran dari

fase sebelumnya. Hubungan penting antar fase digambarkan dengan panah. Sebagai contoh, jika proses berada pada fase *modeling.* Berdasar pada perilaku dan karakteristik model, proses mungkin kembali kepada fase *data preparation* untuk perbaikan lebih lanjut terhadap data atau berpindah maju kepada fase *evaluation.*



Gambar 2.5 Proses CRISP-DM (Larose, 2005)

Enam fase CRISP-DM *(Cross Industry Standard Process for Data Mining)* (Larose, 2005).

a. Fase Pemahaman Bisnis *(Business Understanding Phase)*

1) Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
2) Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan *data mining*.
3) Menyiapkan strategi awal untuk mencapai tujuan.

b. Fase Pemahaman Data *(Data Understanding Phase)*

1) Mengumpulkan data.

2) Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
3) Mengevaluasi kualitas data.
4) Jika diinginkan, pilih sebagian kecil kelompok data yang mungkin mengandung pola dari permasalahan

c. Fase Pengolahan Data *(Data Preparation Phase)*

1) Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.
2) Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
3) Lakukan perubahan pada beberapa variabel jika dibutuhkan.
4) Siapkan data awal sehingga siap untuk perangkat pemodelan.

d. Fase Pemodelan *(Modeling Phase)*

1) Pilih dan aplikasikan teknik pemodelan yang sesuai.
2) Kalibrasi aturan model untuk mengoptimalkan hasil.
3) Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama.
4) Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.

e. Fase Evaluasi *(Evaluation Phase)*

1) Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan.
2) Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
3) Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
4) Mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining*.

f. Fase Penyebaran *(Deployment Phase)*

1) Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikannya proyek.
2) Contoh sederhana penyebaran: Pembuatan laporan.

   Contoh kompleks Penyebaran: Penerapan proses data mining secara parallel pada departemen lain (Larose, 2005).

2.4 Alat Bantu *Data Mining* WEKA

Menurut *situs* resmi WEKA, www.cs.waikato.ac.nz/ml/weka, (*Waikato Environment for Knowledge Analysis*), *"WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes."*

The Waikato Environment for Knowledge Analysis (Weka) adalah rangkaian lengkap perpustakaan kelas *Java* yang mengimplementasikan banyak *state-of-the-art* pembelajaran mesin dan algoritma *data mining*. Weka tersedia secara bebas di *World Wide Web* dan menyertai teks baru pada dokumen *data mining* dan sepenuhnya menjelaskan semua algoritma yang dikandungnya. Aplikasi yang ditulis menggunakan *library class* pada Weka yang dapat dijalankan pada komputer manapun dengan kemampuan *browsing Web,* ini memungkinkan pengguna untuk menerapkan teknik pembelajaran mesin untuk data mereka sendiri terlepas dari *platform* komputer. (Witten, dkk, 2011).

Gambar 2.6. Tampilan Awal GUI WEKA (www.cs.waikato.ac.nz/ml/weka)

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

Software WEKA tidak hanya digunakan untuk akademik saja namun cukup banyak dipakai oleh perusahaan untuk meramalkan bisnis dari suatu perusahaan. Ian H. Witten merupakan latar belakang dibalik kesuksesan WEKA. Beliau merupakan profesor di *Universitas of Waikato, New Zealend*, yang menekuni *Digital Library, Text Mining, Machine Learning* dan *Information Retrieval*. Pada Weka ada beberapa metode pemilihan *variable* dari suatu *dataset,* diantaranya *BestFirst, ExhautiveSearch, FCBFSearch, GeneticSearch, GreedyStepwise, RaceSearch, RandomSearch, Rankerdan, RankerSearch*. Metode atau Teknik yang digunakan Weka adalah *Predictive* dan *Descriptive* karena Weka mendukung teknik-teknik *data preprocessing, clustering, classification, regression, visualization,* dan *feature Reduction*.

Semua teknik Weka adalah didasarkan pada asumsi bahwa data tersedia sebagai *flat file* tungggal atau hubungan, dimana setiap titik data digambarkan oleh sejumlah tetap atribut atau variabel (biasanya, atribut atau variabel *numeric* atau nominal, tetapi beberapa jenis atribut atau variabel lain juga didukung).

2.4.1 *Format Input* WEKA

WEKA mendukung beberapa *format file* untuk inputnya, yaitu:

a. *Comma Separated Values* (CSV): Merupakan *file* teks dengan pemisah tanda koma (,) yang cukup umum digunakan. File ini dapat dibuat dengan menggunakan *Microsoft Excel* atau membuat sendiri dengan menggunakan *notepad.*
b. *Format* C45: Merupakan *format file* yang dapat diakses dengan menggunakan aplikasi WEKA.
c. *Attribute-Relation File Format* (ARFF): Merupakan tipe *file teks* yang berisi berbagai *instance* data yang berhubungan dengan suatu set atribut atau variabel data yang dideskripsikan serta di dalam *file* tersebut.
d. *SQL Server/*MySql *Server*: Dapat mengakses *database* dengan menggunakan *SQL Server/MySql Server*.

2.4.2 Algoritma pada WEKA

J48 merupakan implementasi C4.5 pada WEKA. J48 menangani himpunan data dalam format ARFF, tidak mengandung kode untuk mengkonstruksi pohon keputusan. Kelas ini mereferensi kelas-kelas lain, kebanyakan pada paket Weka. *Classifiers* J48, yang mengerjakan semua proses konstruksi pohon.Adapun kelebihan C4.5 antara lain:

a. C4.5 mampu menangani atribut atau variabel dengan tipe diskrit atau kontinu.
b. C4.5 mampu menangani atribut atau variabel yang kosong *(Missing Value).*

C4.5 telah berkembang menjadi C5. C4.5 merupakan pengembangan dari ID3 *(Iterative Dichotomiser 3).* Naïve *bayes* merupakan implementasi algoritma *naïve bayes* pada WEKA dan S*impleCart* merupakan implementasi algoritma *CART.*

2.4.3 *Test Options* WEKA

WEKA memiliki empat jenis *test option*, yaitu:

a. *Use training set*: *Classifier* ini dilakukan dengan menggunakan data *training* itu sendiri.
b. *Supplied test set*: *Classifier* ini dievaluasi dengan memprediksi seberapa baik satu set *class* yang diambil dari sebuah *file*. *Test option* ini dilakukan dengan menggunakan data lain. Cara yang dilakukan adalah dengan meng*klik* set tombol yang menampilkan dialog yang memungkinkan untuk memilih *file* untuk menguji.

c. Cross-*validation*: *Classifier* ini dievaluasi oleh *cross-validation*, menggunakan jumlah *fold* yang dimasukkan ke dalam kolom teks *folds*. Pada *cross-validation* akan ada pilihan berapa *fold* yang akan digunakan. Nilai *default*-nya adalah 10.

d. *Percentage split*: *Classifier* ini dievaluasi dengan memprediksi seberapa baik persentase dari data yang digunakan selama pengujian. Jumlah data tersebut dibagi berdasarkan nilai yang dimasukkan ke dalam *field*.

2.5 Evaluasi dan Alat Ukur

Hasil evaluasi algoritma dapat ditampilkan dengan menggunakan *Confusion Matrix* (Tan, 2005). *Confusion Matrix* adalah salah satu alat ukur berbentuk *matrix* 2x2 yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi dataset terhadap kelas lulus dan tidak lulus pada algoritma yang dipakai tiap kelas yang diprediksi memiliki empat kemungkinan keluaran yang berbeda, yaitu *true positive* (TP) dan *true negatives* (TN) yang menunjukkan ketepatan klasifikasi. Jika prediksi keluaran bernilai positif sedangkan nilai aslinya adalah negatif maka disebut dengan *false positive* (FP) dan jika prediksi keluaran bernilai negatif sedangkan nilai aslinya adalah positif maka disebut dengan *false negative* (FN). Tabel 2.1 menyajikan bentuk *confusion matrix* seperti yang telah dijelaskan sebelumnya.

Tabel 2.1 *Confusion Matrix* untuk Klasifikasi Kelas (Tan S, 2005)

		Predicated Class	
		Yes	*No*
Actual Class	*Yes*	*True Positive (TP)*	*False Negative (FN)*
	No	*False Positive (FP)*	*True Negative (TN)*

Perhitungan akurasi dengan tabel *confusion matrix* adalah sebagai berikut:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+TN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.1)}$$

Pengukuran data dilakukan dengan *confusion matrix* (Tan S, 2005) dan ROC *Curve* (AUC) (Gorunescu, 2011) untuk mengevaluasi hasil dari algoritma *Decision Tree* C4.5. *Confusion matrix* merupakan sebuah table yang terdiri dari banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi. Tabel ini diperlukan untuk mengukur kinerja suatu model klasifikasi (Ariawan, 2009).

Tabel 2.2 *Confusion Matrix* (Ariawan, 2009)

		Predicated Class	
Actual Class		*Class = 1*	*Class = 0*
	Class = 1	F 11	F 10
	Class = 0	F 01	F 00

Perhitungan akurasi dengan tabel *confusion matrix* adalah sebagai berikut:

$$\text{Akurasi} = \frac{F\ 11 + F\ 00}{F\ 11 + F\ 10 + F\ 01 + F\ 00} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

ROC *(Receivier Operating Characteristic) Curve* adalah grafik antara sensitifitas *(true positive rate)* pada sumbu Y dengan 1-spesifisitas pada sumbu X *(false positive rate)*, seakan-akan menggambarkan tawar menawar antara sensitivitas dan spesifisitas, yang tujuannya adalah untuk menentukan *cut off point* pada uji *diagnostic* yang bersifat kontinyu. Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = Klasifikasi sangat baik

b. 0.80-0.90 = Klasifikasi baik

c. 0.70-0.80 = Klasifikasi cukup

d. 0.60-0.70 = Klasifikasi buruk

e. 0.50-0.60 = Klasifikasi salah

2.6 Teknik-Teknik *Data Mining*

Teknik–teknik ini terdiri atas algoritma spesifik yang dapat digunakan untuk setiap fungsi. Memahami bagaimana teknik-teknik ini bekerja dapat membantu dalam memilih teknik yang sesuai untuk memecahkan suatu problem. Beberapa teknik *data mining* antara lain:

2.6.1 Association Rule Mining

Association Rule Mining adalah teknik *mining* untuk menemukan aturan assosiatif antara suatu kombinasi item. Contoh aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah bisa diketahui berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tersebut pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu. Penting

tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, *support* yaitu persentase kombinasi item tersebut dalam database dan *confidence* yaitu kuatnya hubungan antar item dalam aturan assosiatif.

2.6.2 Classification

Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika-maka", berupa *decision tree*, formula matematis atau *neural network. Decision tree* adalah salah satu metode *classification* yang paling populer karena mudah untuk diinterpretasi oleh manusia.

2.6.2.1 Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membangun sebuah pohon keputusan (*decision tree*) dari data. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang juga merupakan algoritma untuk membangun sebuah pohon keputusan. Algoritma C4.5 secara rekursif mengunjungi tiap simpul keputusan, memilih percabangan optimal, sampai tidak ada cabang lagi yang mungkin dihasilkan (Larose, 2005).

Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, *CART*, dan C4.5 (Larose, 2005). Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Larose, 2005). Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut atau variabel dan *record*. Atribut atau variabel menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan pertandingan tenis, kriteria yang diperhatikan adalah cuaca, angin, dan temperatur. Salah satu atribut atau variabel merupakan atribut atau variabel yang menyatakan data solusi per *item* data yang disebut target atribut atau variabel. Atribut atau variabel memiliki nilai-nilai yang dinamakan *instance.* Misalkan atribut atau variabel cuaca memiliki *instance* berupa cerah, berawan dan hujan (Basuki & Syarif, 2003).

Pada penelitian Mujib Ridwan, dkk, (2013) menjelaskan bahwa faktor yang paling berpengaruh dalam penentuan klasifikasi kinerja akademik mahasiswa adalah Indeks Prestasi Komulatif (IPK), Indeks Prestasi Semester (IPS) semester 1, IPS semester 4 dan jenis kelamin. Pada penelitian ini peneliti menggunakan algoritma C4.5 dalam menentukan prediksi kelulusan berdasarkan attribut jenis kelamin, asal sekolah SMA dan IPS semester satu sampai dengan semester enam.

Berdasarkan hasil penelitian David Hartanto Kamagi (2014) bahwa Data mining dengan algoritma C4.5 dapat diimplementasikan untuk memprediksi tingkat kelulusan mahasiswa dengan empat kategori yaitu lulus cepat, lulus tepat, lulus terlambat dan drop out. Attribute yang paling berpengaruh dalam hasil prediksi adalah IPS semester enam, berhasil memprediksi kelulusan mahasiswa dengan presentase 87.5% dari enam puluh data training dan empat puluh data testing.

Ada beberapa tahapan dalam membangun sebuah pohon keputusan dengan Algoritma C4.5 yaitu (Kusrini, 2009) :

a. Menyiapkan *data training. Data training* biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
b. Menentukan akar dari pohon. Akar akan diambil dari atribut atau variabel yang terpilih, dengan cara menghitung nilai gain dari masing-masing atribut atau variabel, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut atau variabel, hitung dahulu nilai *entropy*. Untuk menghitung nilai *entropy* digunakan rumus:

$$Entropy(S) = \sum_{i=1}^{n} - \rho i * \log_2 \rho i \quad \text{.....(2.3)}$$

Keterangan :

S : himpunan kasus

A : atribut atau variabel

n : jumlah partisi S

pi : proporsi dari Si terhadap S

c. Kemudian hitung nilai *gain* yang menggunakan rumus:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad \text{.....(2.4)}$$

Keterangan :

S = himpunan kasus

A = fitur

n = jumlah partisi atribut atau variabel A

$|Si|$ = proporsi Si terhadap S

$|S|$ = jumlah kasus dalam S

d. Ulangi langkah ke-2 hingga semua *record* terpartisi.

e. Proses partisi pohon keputusan akan terhenti saat :

(1) Semua *record* dalam simpul N mendapat kelas yang sama.

(2) Tidak ada atribut atau variabel didalam *record* yang dipartisi lagi.

(3) Tidak ada *record* didalam cabang yang kosong.

Algoritma C4.5 dapat digunakan untuk menyusun sistem yang mempunyai kemampuan melihat pola kelulusan mahasiswa, untuk selanjutnya bisa menjadi strategi dalam proses perkuliahan (Huda, 2010).

Algoritma C4.5 merupakan algoritma klasifikasi pohon keputusan yang banyak digunakan karena memiliki kelebihan utama dari algoritma yang lainnya. Kelebihan algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterprestasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut atau variabel bertipe diskret dan dapat menangani atribut atau variabel bertipe diskret dan numerik (Han, 2001).

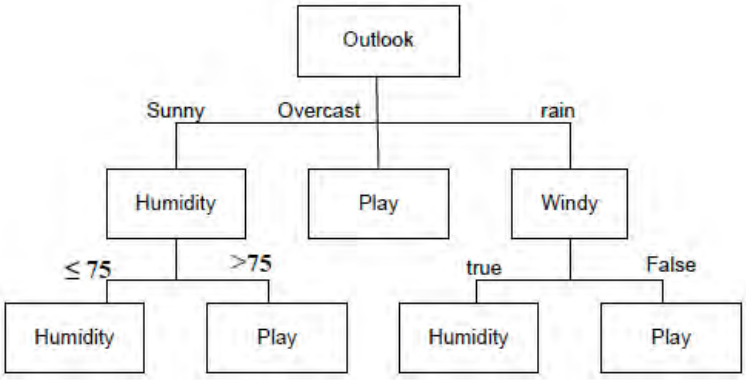
Dalam mengkonstruksi pohon, algoritma C4.5 membaca seluruh sampel *data training* dari *storage* dan memuatnya ke memori. Hal inilah yang menjadi salah satu kelemahan algoritma C4.5 dalam kategori “skalabilitas” adalah algoritma ini hanya dapat digunakan jika *data training* dapat disimpan secara keseluruhan dan pada waktu yang bersamaan di memori (Moertini, 2007).

Menurut hasil penelitian dari (Marselina Silvia Suhartinah, Ernastuti, 2010) bahwa dengan menggunakan algoritma C4.5 kesalahan yang dihasilkan dalam proses prediksi lebih sedikit karena C4.5 melakukan klasifikasi *record-record* ke dalam kelas tujuan yang

ada. Algoritma *Naïve Bayes* bila diimplementasikan menggunakan data yang digunakan dalam proses *training* akan menghasilkan nilai kesalahan yang lebih besar karena pada *Naïve Bayes* nilai suatu atribut atau variabel adalah independen terhadap nilai lainnya dalam satu atribut atau variabel yang sama. Namun memiliki akurasi yang lebih tinggi bila dimplementasikan ke data yang berbeda dari data *training* dan kedalam data yang jumlahnya lebih besar.

Algoritma C4.5 merupakan kelompok algoritma *Decision Tree*. Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* berupa data contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan *samples* merupakan *field-field* data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data (Sunjana, 2010).

Pada dasarnya konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (*rule)*. C4.5 adalah algoritma yang cocok untuk masalah klasifikasi dan data mining, C4.5 memetakan nilai atribut atau variabel menjadi *class* yang dapat diterapkan untuk klasifikasi baru (Wu & Kumar, 2009), seperti gambar 2.7.



Gambar 2.7 Pohon Keputusan (Bramer, 2007)

2.6.2.2 Naive Bayes

Naïve Bayes merupakan metode yang tidak memiliki aturan, *Naïve Bayes* merupakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. Klasifikasi *Naïve Bayes* adalah pengklasifikasian statistik

yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class.* Klasifikasi *Bayesian* didasarkan pada teorema *Bayes*, diambil dari nama seorang ahli matematika Inggris, Thomas Bayes (1702-1761), (Bramer,2007).

Naïve Bayes adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. *Naïve Bayes* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*. *Naïve Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar (Prasetyo, 2012). Prediksi Bayes didasarkan pada formula teorema Bayes dengan formula umum sebagai berikut :

$$P(H|X)= \frac{P(H|X) * P(H)}{P(X)} \quad \text{.....(2.5)}$$

Dimana :

X : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu class spesifik.

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) : Probabilitas hipotesis H (prior probability)

P(X|H) : Probabilitas X berdasar kondisi pada hipotesis H

P(X) : Probabilitas dari X

Naïve Bayes adalah penyederhanaan metode *Bayes*. Teorema *Bayes* disederhanakan menjadi:

$$P(H|X) = P(X|H) * P(X) \quad \text{.....(2.6)}$$

Bayes rule diterapkan untuk menghitung posterior dan probabilitas dari data sebelumnya. Dalam analisis Bayesian, klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi (*prior* dan *posterior)* untuk menghasilkan *probabilitas* menggunakan aturan *Bayes.*

2.6.2.3 CART

Ciri khas algoritma *CART* ini adalah noktah keputusan yang selalu bercabang 2 atau bercabang biner. Algoritma *CART* ini pertama kali digagas oleh Leo Breiman,

Jerome Friedman, Richard Olshen, dan Charles Stone (Larose, 2005). Algoritma ini juga masuk dalam *The Top 10 Algorithm in Data Mining* (Wu dan Kumar, 2009).

Langkah – langkah pada algoritma *CART* adalah sebagai berikut (Susanto, dkk., 2010):

a. Langkah pertama, susunlah calon cabang *(Candidate Split)*. Penyusunan ini dilakukan terhadap seluruh variabel *predictor* secara lengkap *(Exhaustive).* Daftar yang berisi calon cabang disebut daftar calon cabang mutakhir.
b. Langkah kedua adalah menilai kinerja keseluruhan calon cabang yang ada pada daftar calon cabang mutakhir dengan jalan menghitung nilai besaran kesesuaian, Φ (s|t).
c. Langkah ketiga adalah menentukan calon cabang manakah yang akan benar-benar dijadikan cabang dengan memilih calon cabang yang memiliki nilai kesesuaian Φ (s|t) terbesar. Setelah itu gambarkanlah percabangan. Jika tidak ada lagi noktah keputusan, pelaksanaan algoritma *CART* akan dihentikan.
d. Namun, jika masih terdapat noktah keputusan, pelaksanaan algoritma dilanjutkan dengan kembali ke langkah kedua, dengan terlebih dahulu membuang calon cabang yang telah berhasil menjadi cabang sehingga mendapatkan daftar calon cabang yang baru.
e. Pohon keputusan yang dihasilkan *CART* merupakan pohon biner dimana tiap simpul wajib memiliki dua cabang. *CART* secara rekursif membagi *records* pada data latihan ke dalam *subset-subset* yang memiliki nilai atribut atau variabel target (kelas) yang sama.

Algoritma *CART* mengembangkan pohon keputusan dengan memilih percabangan yang paling optimal bagi tiap simpul. Pemilihan dilakukan dengan menghitung segala kemungkinan pada tiap variabel.

Misalkan *Φ(s|t)* merupakan nilai “kebaikan” kandidat cabang s pada simpul t, maka nilai *Φ(s|t)* dapat dihitung sebagai persamaan berikut (Larose, 2005):

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)| \quad \text{.....(2.7)}$$

Dimana :

t_L=simpul anak kiri dari simpul t

t_R=simpul anak kanan dari simpul t

P_L=jumlah *record* pada t_L jumlah seluruh *record* pada data latihan

P_R=jumlah *record* pada t_R jumlah seluruh *record* pada data latihan

P j|t_L =jumlah *record* kelas j pada t_L jumlah *record* pada simpul t

P j|t_R =jumlah *record* kelas j pada t_R jumlah record *pada* simpul t

Maksimal ketika *record* yang berada pada cabang kiri atau kanan simpul memiliki kelas yang sama (seragam). Nilai maksimal yang dicapai sama dengan jumlah kelas pada data. Misalkan jika data terdiri atas dua kelas, maka nilai maksimal adalah 2.

$$\sum_{j=1}^{\#kelas} |P(j|t_L) - P(j|t_R)| \quad \text{..... (2.8)}$$

Semakin seragam *record* pada cabang kiri atau kanan, maka semakin tinggi nilai. Nilai maksimal $2P_LP_R$ sebesar 0.5 dicapai ketika cabang kiri dan kanan memiliki jumlah *record* yang sama. Kandidat percabangan yang dipilih adalah kandidat yang memiliki nilai *Φ(s|t)* paling besar.

2.6.3 Clustering

Berbeda dengan *association rule mining* dan *classification* dimana kelas data telah ditentukan sebelumnya, *clustering* melakukan pengelompokan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui itu. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/*cluster*. *Clustering* dapat dilakukan pada data yang memiliki beberapa atribut atau variabel yang dipetakan sebagai ruang multidimensi. Banyak algoritma *clustering* memerlukan fungsi jarak untuk mengukur kemiripan antar data, diperlukan juga metode untuk normalisasi bermacam atribut atau variabel yang dimiliki data. Beberapa kategori algoritma *clustering* yang banyak dikenal adalah metode partisi dimana pemakai harus menentukan jumlah *k* partisi yang diinginkan lalu setiap data diuji untuk dimasukkan pada salah satu partisi, metode lain yang telah lama dikenal adalah metode hierarki yang terbagi dua lagi: *bottom-up*

yang menggabungkan *cluster* kecil menjadi *cluster* lebih besar *dan top-down* yang memecah *cluster* besar menjadi *cluster* yang lebih kecil. Kelemahan metode ini adalah bila salah satu penggabungan atau pemecahan dilakukan pada tempat yang salah, tidak dapat didapatkan *cluster* yang optimal. Pendekatan yang banyak diambil adalah menggabungkan metode hierarki dengan metode *clustering* lainnya seperti yang dilakukan oleh *Chameleon* (Karypis, George. Han, Eui-Hong Sam).

Tabel 2.3 Daftar Jurnal *Data Mining* Dengan Teknik Klasifikasi

NO	PENELITI	ATRIBUT/VARIABEL	HASIL
1	Yusuf Sulistyo Nugroho (2014)	IPK, Jurusan SMA, Jenis Kelamin, Asal Sekolah, Jumlah SKS Persemester, Pernah menjadi Asisten	Variabel yang paling tinggi pengaruhnya terhadap predikat kelulusan adalah partisipasi mahasiswa menjadi Asisten
2	Indri Rahmayuni (2014)	No. Peserta, Nama Peserta, Tempat Lahir, Tanggal Lahir, Tahun Masuk SMU/SMK, Jurusan SMU/SMK, Nilai Ijazah/STTB, Nilai NEM/UAN, Pilihan Jurusan, Alamat Rumah, Nama Orang Tua, Pekerjaan Orang Tua, Asal Daerah, Pekerjaan Orang Tua, Keadaan Orang Tua, Jumlah Saudara, Jenis Kelamin, Agama, Kewarganegaraan, Pendidikan Orang Tua	Algoritma C4.5 memberikan akurasi yang lebih baik dari pada algoritma CART dalam klasifikasi data karakteristik mahasiswa
3	David Hartanto & Seng Hansun (2014)	IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, IP Semester 5, IP Semester 6, Jenis Kelamin, SMA, Jumlah SKS	Data Mining dengan Algoritma C4.5 dapat diimplementasikan untuk memprediksi tingkat kelulusan mahasiswa dengan 4 kategori yaitu Lulus Cepat, Lulus Tepat, Lulus Terlambat dan Drop Out. Atribut atau variabel yang paling berpengaruh adalah IP Semester 6.

NO	PENELITI	ATRIBUT/VARIABEL	HASIL
4	David Hartanto Kamagi, Seng Hansun (2014)	Nama, NIM, IP semester1 sampai dengan IP semester 6, jenis kelamin, SMA, jumlah SKS	Data mining dengan algoritma C4.5 dapat diimplementasikan untuk memprediksi tingkat kelulusan mahasiswa dengan empat kategori yaitu lulus cepat, lulus tepat, lulus terlambat dan drop out. Attribute yang paling berpengaruh dalam hasil prediksi adalah IPS semester enam. berhasil memprediksi kelulusan mahasiswa dengan presentase 87.5%.
5	Yuda Septian Nugroho (2013)	NIM, Nama, Jenjang, Program Studi, Provinsi, Seks, SKS, IPK, Tahun Lulus	Metode klasifikasi Naive Bayes dengan atribut atau variabel NIM, Nama, Jenjang, Program Studi, Propinsi, Seks, IPK, Tahun Lulus didapat sebuah hasil bahwa nilai akurasi terhadap klasifikasi kelulusan sebesar 82,08%
6	Liliana Swastina (2013)	Nama, Jenis Kelamin, Umur, Asal Sekolah, Jurusan Asal Sekolah, Nilai UAN, IPK Semester1, IPK Semester2	Algoritma *Decision Tree* C4.5 memprediksi lebih akurat dari pada algoritma *Naïve Bayes* dalam penentuan kesesuaian jurusan dan rekomendasi jurusan mahasiswa.
7	Mohammed M, Abu Tair, Alaa M. El-Haleed (2012)	Student_ID, Student_Name, Gender, Date_of_Birth, Place_of_Birth, Speciality, Enrollment_year, Graduation_year, City, Location, Address, Telephone_Number, Matriculation_GPA, Secondary_School_Type, Matriculation_Obtained_Place, Matriculation_Year, College_GPA, Grade	Data mining pendidikan untuk meningkatkan kinerja lulusan dan mengatasi masalah nilai rendah mahasiswa pascasarjana.

NO	PENELITI	ATRIBUT/VARIABEL	HASIL
8	Marselina Silvia Suhartinah, Ernastuti (2010)	NEM SMA, IP semester 1 dan IP semester 2, IPK DNU semester 1 dana 2, gaji orang tua dan pekerjaan orang tua.	Dengan menggunakan algoritma C4.5 kesalahan yang dihasilkan dalam proses prediksi lebih sedikit karena C4.5 melakukan klasifikasi record-record ke dalam kelas tujuan yang ada. Dengan algoritma naive bayes akan menghasilkan nilai kesalahan yang lebih besar karena pada naive bayes nilai suatu atribut atau variabel adalah independent terhadap nilai lainnya dalam satu atribut atau variabel yang sama. Namun memiliki akurasi akurasi yang lebih tinggi bila dimplementasikan ke data yang berbeda dari data training dan kedalam data yang jumlahnya lebih besar.
9	Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal	PSM, CTG, SEM, ASS, ATT, LW, ESM	Hasil percobaan menunjukkan bahwa *CART* adalah algoritma terbaik untuk klasifikasi data. Penelitian ini akan membantu untuk para siswa dan guru untuk meningkatkan kinerja siswa. Penelitian ini juga akan bekerja untuk mengidentifikasi para pelajar yang membutuhkan perhatian khusus dan juga akan bekerja untuk mengurangi gagal rasio dan mengambil tindakan yang tepat untuk pemeriksaan semester berikutnya.

BAB III

OBYEK DAN METODE PENELITIAN

3.1 Sejarah STMIK Rosma

Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Rosma merupakan bagian dari lembaga pendidikan formal yang didirikan oleh Yayasan Pendidikan Rosma Karawang, sebuah yayasan pendidikan yang dirintis oleh salah seorang pendidik, yaitu Alm. Waba, M.Pd. STMIK Rosma didirikan berdasarkan akta yayasan Rosma yang dibentuk pada tanggal 13 November 1996 dengan legalitas akta notaris Ida Rosida Suryana, SH No: 321. Pemberian status terdaftar kepada 4 (empat) program studi untuk jenjang pendidikan Strata Satu, Diploma Tiga dan Diploma Satu menurut Surat Keputusan Menteri Pendidikan dan Kebudayaan RI. No: 107/D/O/2000 dan SK DIRJEN DIKTI No: 3525/D/T/2003.

3.1.1 Visi dan Misi

Visi:

Menjadi Perguruan Tinggi swasta yang diterima sebagai panutan, pelopor dalam pengembangan dan penerapan ilmu serta teknologi informasi di Indonesia.

Misi:

a. Menyelenggarakan program studi yang menunjang pengembangan dan penerapan teknologi informasi dalam berbagai bidang ilmu.
b. Menyediakan sarana dan lingkungan yang kondusif bagi pelaksanaan kegiatan pembelajaran yang efektif dan efisien.
c. Menjaga keterkaitan dan relevansi seluruh kegiatan akademik dengan kebutuhan pembangunan social ekonomi dan industri Indonesia serta mengantisipasi semakin globalnya kehidupan masyarakat.
d. Melangsungkan kerjasama dengan berbagai pihak, baik dari dalam maupun dari luar, sehingga ilmu dan teknologi selalu mutakhir serta dapat diterapkan secara tepat guna.

e. Berusaha menghasilkan cendikiawan dan ilmuwan muda yang bermoral, terampil dan kreatif sehingga mampu mengambil dan menciptakan sendiri perannya dalam upaya pembangunan nasional.

3.1.2 Syarat Kelulusan

Berikut syarat kelulusan mahasiswa STMIK ROSMA Karawang untuk Strata satu (S1) (Buku Panduan Akademik, 2010):

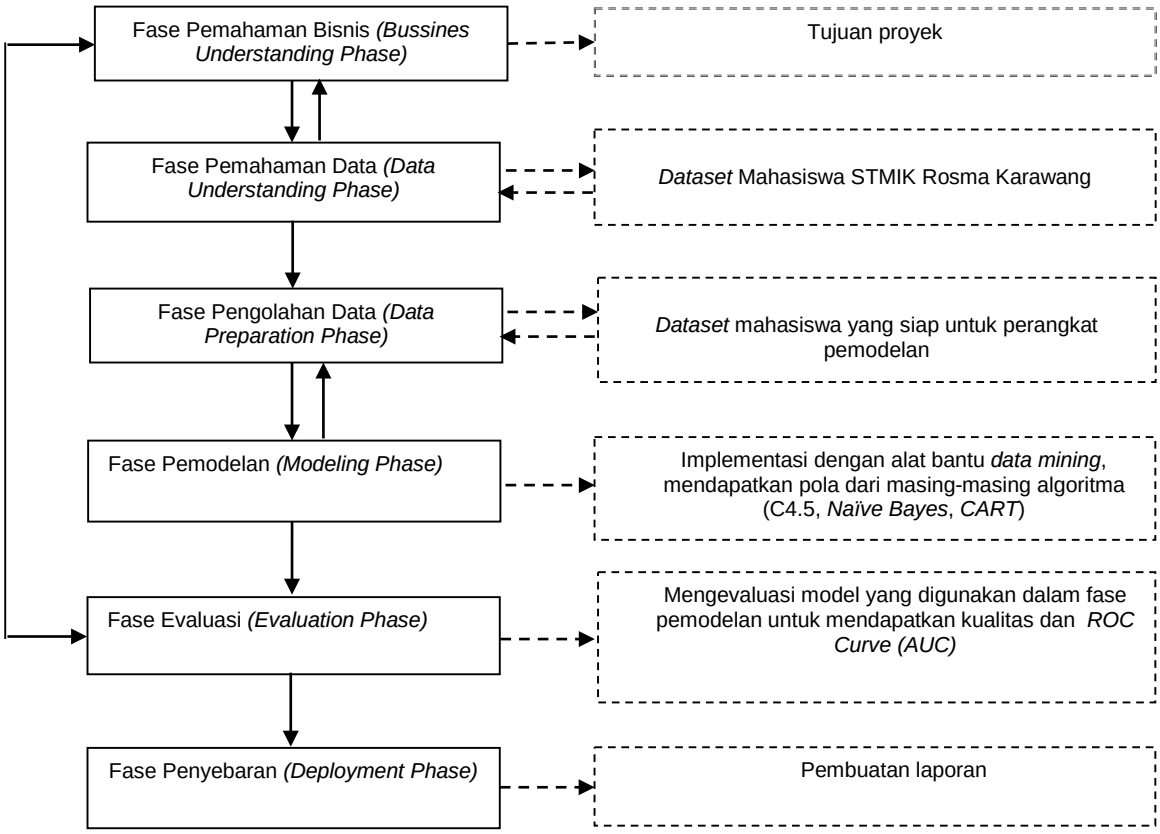
1. Telah menyelesaikan semua kewajiban dan/atau yang dibebankan dalam mengikuti suatu program studi sesuai dengan ketentuan yang berlaku.
2. Telah menyelesaikan kewajiban administrasi dan keuangan berkenaan dengan program studi yang diikuti sesuai ketentuan yang berlaku.
3. Telah dinyatakan lulus oleh STMIK Rosma dengan jumlah SKS minimal 144 (tidak termasuk matakuliah yang mempunyai Grade D atau E).

Berikut syarat kelulusan mahasiswa STIMIK ROSMA Karawang untuk Diploma Tiga (D3) (Buku Panduan Akademik, 2010):

1. Telah menyelesaikan semua kewajiban dan/atau tugas yang dibebankan dalam mengikuti suatu program studi sesuai dengan ketentuan yang berlaku.
2. Telah menyelesaikan kewajiban administrasi dan keuangan berkenaan dengan program studi yang diikuti sesuai ketentuan yang berlaku.
3. Telah dinyatakan lulus oleh STMIK Rosma dengan jumlah SKS minimal 110 (tidak termasuk matakuliah yang mempunyai Grade D atau E).

3.2 Metode Penelitian *Data Mining*

Penelitian ini didesain dengan merujuk pada model CRISP-DM *(Cross Industry Standard Process for Data Mining)*. Berikut gambar tahapan yang dilakukan dalam penelitian.



Gambar 3.1 Tahapan Penelitian

3.2.1 Fase Pemahaman Bisnis *(Business Understanding Phase)*

Fase awal ini berfokus pada pemahaman tujuan dan kebutuhan proyek dari perspektif bisnis, kemudian mengubah pengetahuan ini ke dalam definisi masalah dan desain rencana awal *data mining* untuk mencapai tujuan proyek.

Tujuan proyek dalam penelitian ini adalah mengkaji dan membuat model hasil komparasi algoritma C4.5, *Naïve Bayes, CART*, serta menentukan algoritma mana yang paling akurat dan menghasilkan *rule* prediksi kelulusan mahasiswa STMIK Rosma Karawang sehingga dapat dijadikan acuan untuk meningkatkan jumlah kelulusan mahasiswa di tahun-tahun kelulusan berikutnya.

Strategi awal untuk mencapai tujuan adalah melakukan permintaan data mahasiswa kepada Bagian Akademik (BAAK) STMIK Rosma Karawang.

3.2.2 Fase Pemahaman Data *(Data Understanding Phase)*

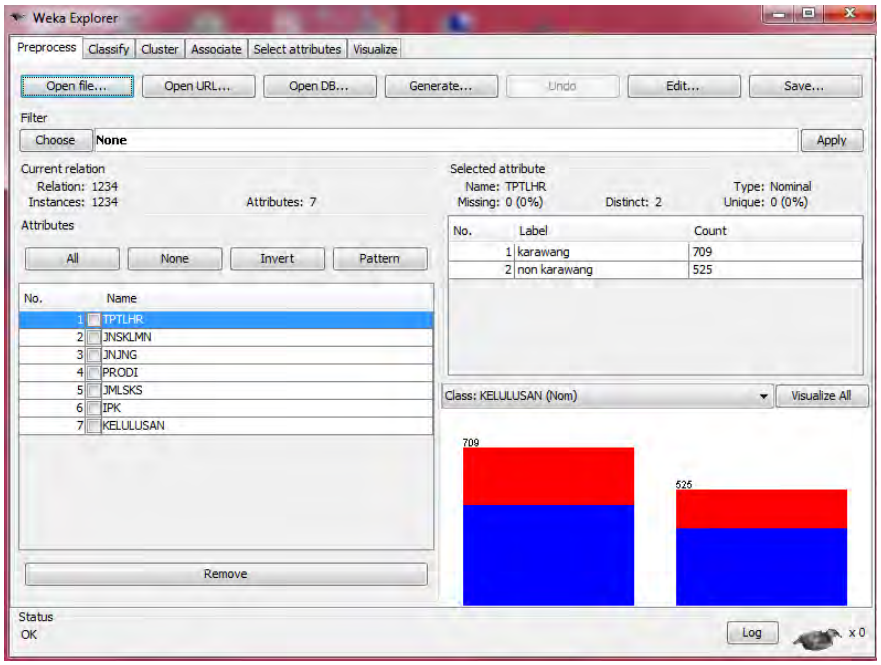
Fase pemahaman data dimulai dengan pengumpulan data awal dan dilanjutkan dengan aktifitas-aktifitas lain untuk mengenal data, mengidentifikasi permasalahan kualitas data, atau untuk mendeteksi *subset* data yang menarik untuk membentuk hipotesis bagi informasi yang tersembunyi.

Fase ini akan memilih dan menciptakan satu *dataset* mahasiswa STMIK Rosma Karawang yang didapat dari bagian Akademik (BAAK) untuk mendukung proses penemuan pengetahuan dengan mencari dari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan seluruh data untuk proses KDD *(Knowledge Discovery in Database).* Apabila dalam fase ini terjadi ketidaksesuaian maka analisis akan kembali kepada fase sebelumnya yaitu fase pemahaman bisnis.

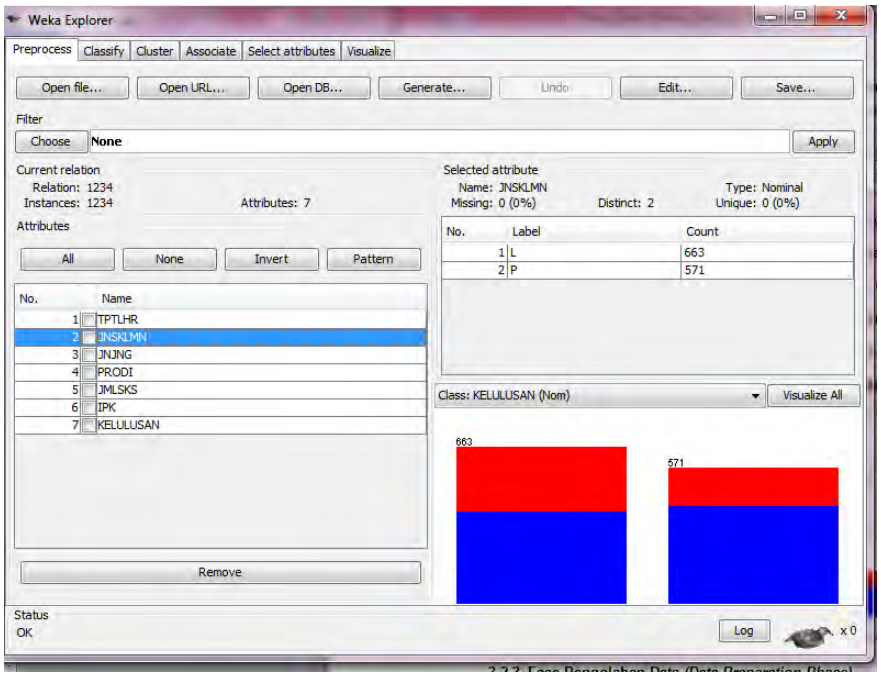
3.2.3 Fase Pengolahan Data *(Data Preparation Phase)*

Fase pengolahan data adalah proses setelah fase pemahaman data selesai dilakukan. Fase ini merupakan pekerjaan berat yang perlu dilakukan secara intensif.

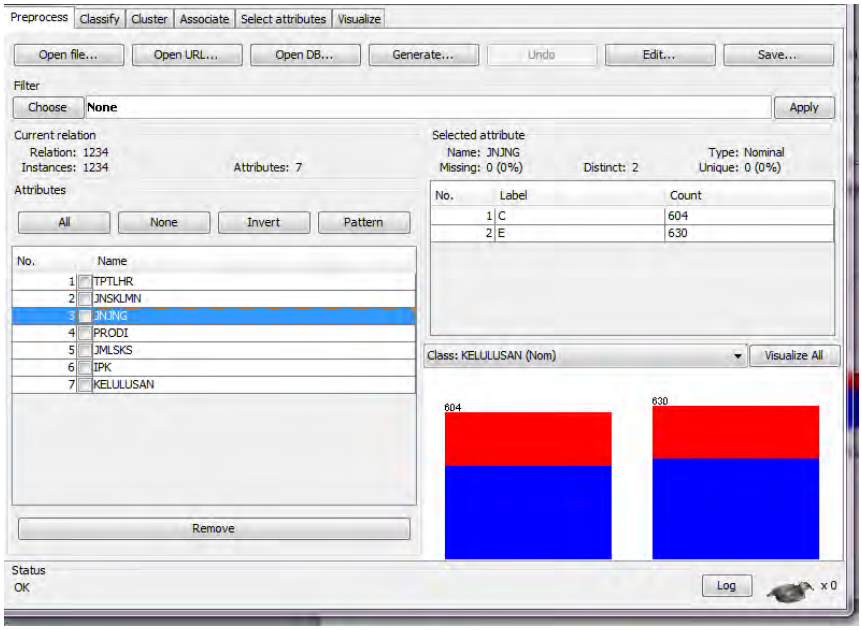
Fase ini meliputi seluruh aktifitas yang dilakukan untuk membangun dataset akhir (data yang akan digunakan sebagai masukan bagi aplikasi pemodelan) dari data mentah awal. Proses persiapan data biasanya dilakukan berulang kali untuk memastikan kualitas data telah dicapai. Aktifitas persiapan data antara lain pemilihan tabel, *record*, dan atribut, serta pembersihan dan transformasi data. Pada penelitian ini dilakukan pengolahan data dari masing-masing variabel :



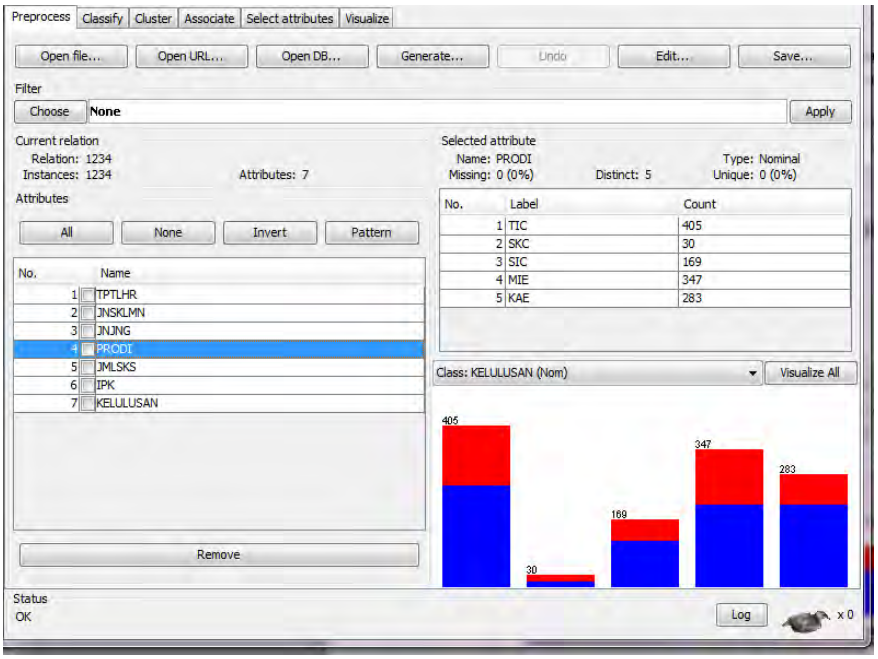
Gambar 3.2 Pengolahan Variabel Tempat Lahir (TPTLHR)



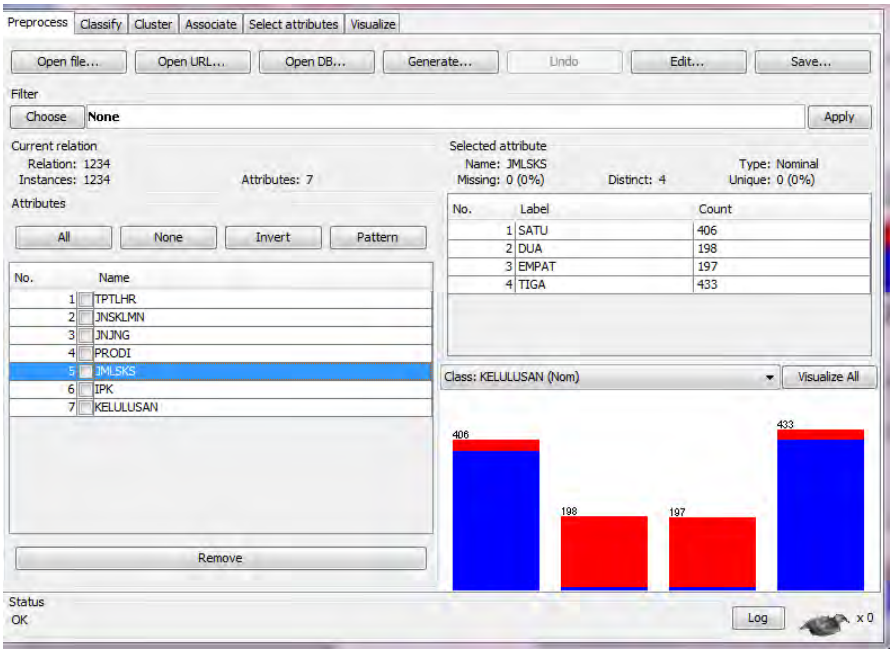
Gambar 3.3 Pengolahan Variabel Jenis Kelamin (JNSKLMN)



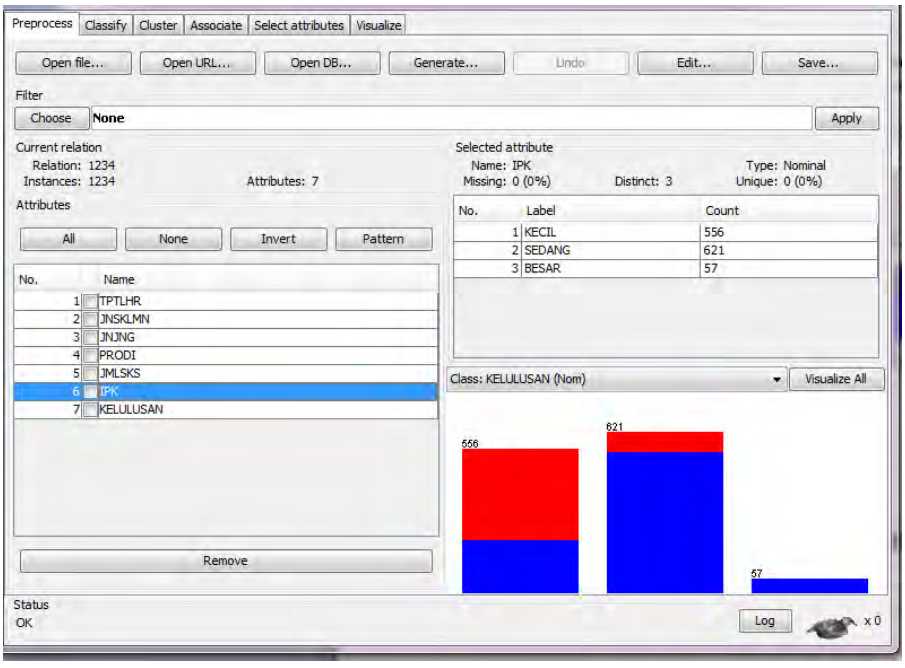
Gambar 3.4 Pengolahan Variabel Jenjang (JNJNG)



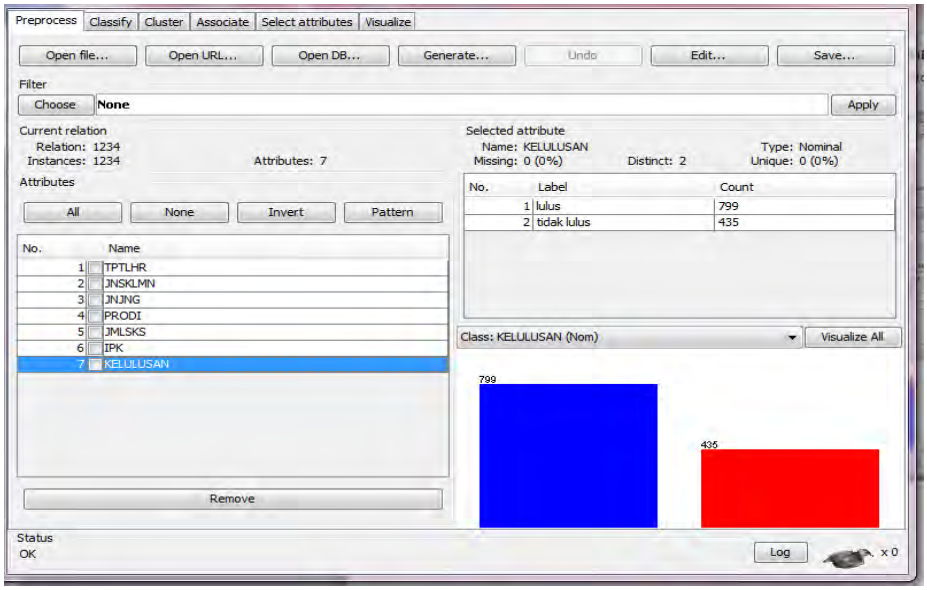
Gambar 3.5 Pengolahan Variabel Program Studi (PRODI)



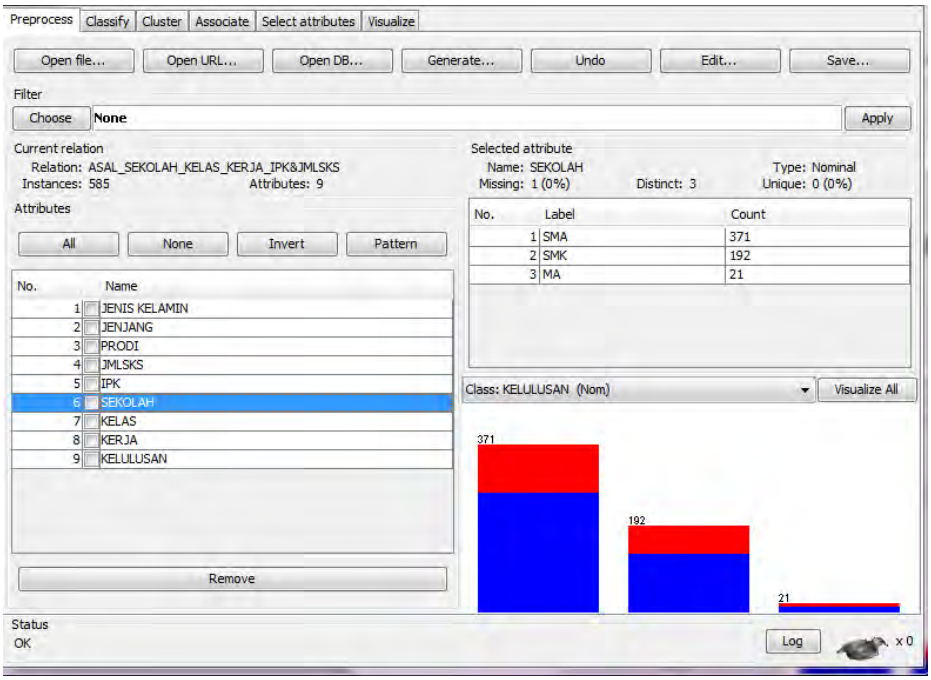
Gambar 3.6 Pengolahan Variabel Jumlah SKS (JMLSKS)



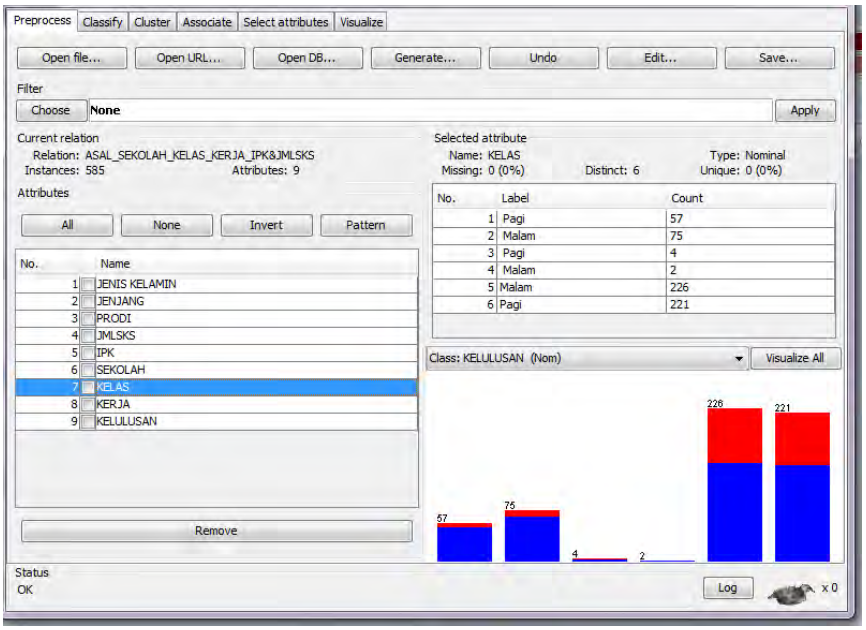
Gambar 3.7 Pengolahan Variabel Indeks Prestasi Komulatif (IPK)



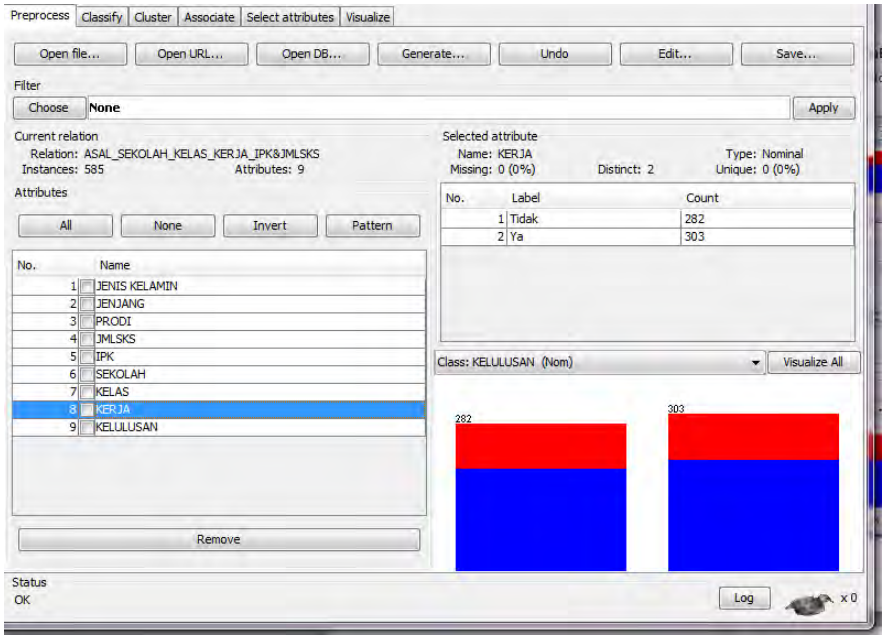
Gambar 3.8 Pengolahan Variabel Kelulusan



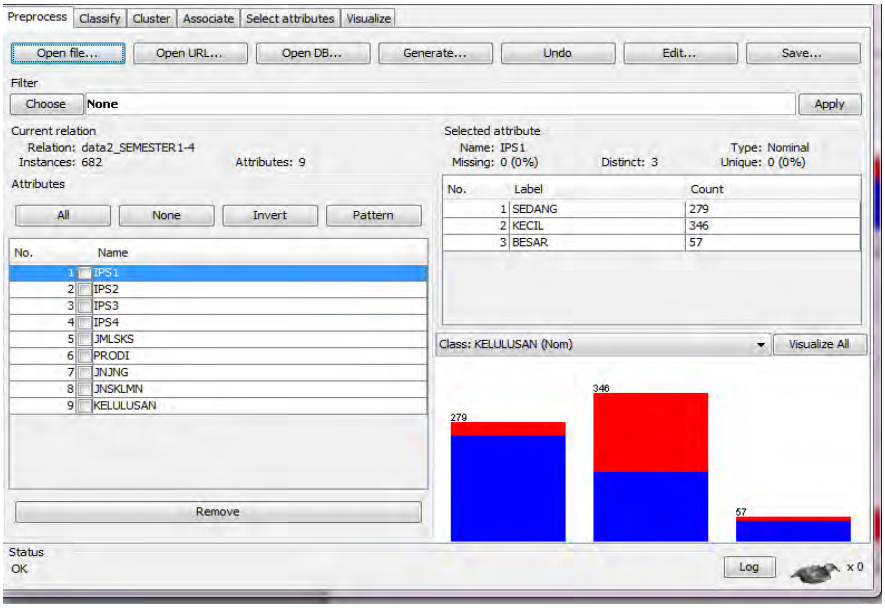
Gambar 3.9 Pengolahan Variabel Sekolah



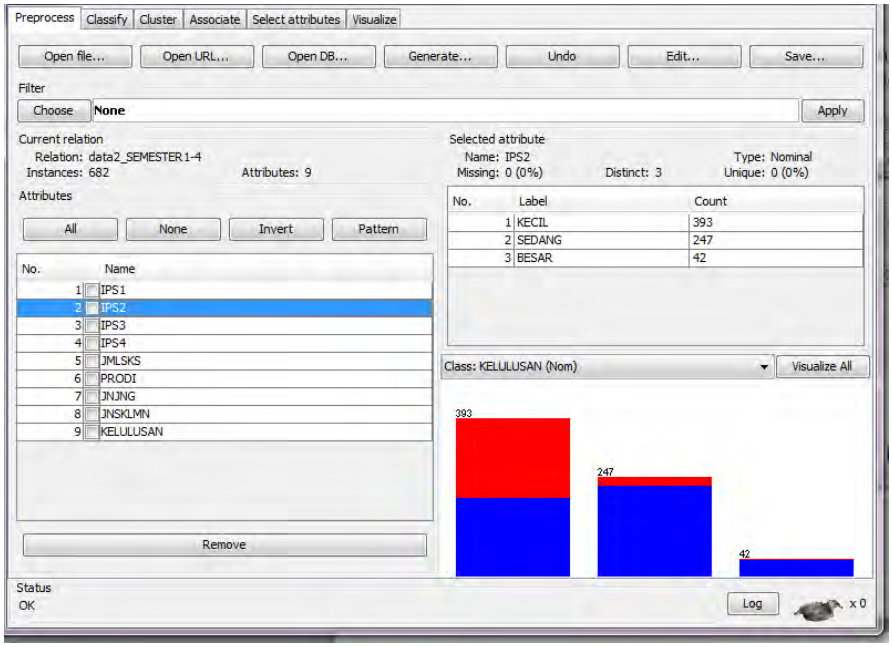
Gambar 3.10 Pengolahan Variabel Kelas



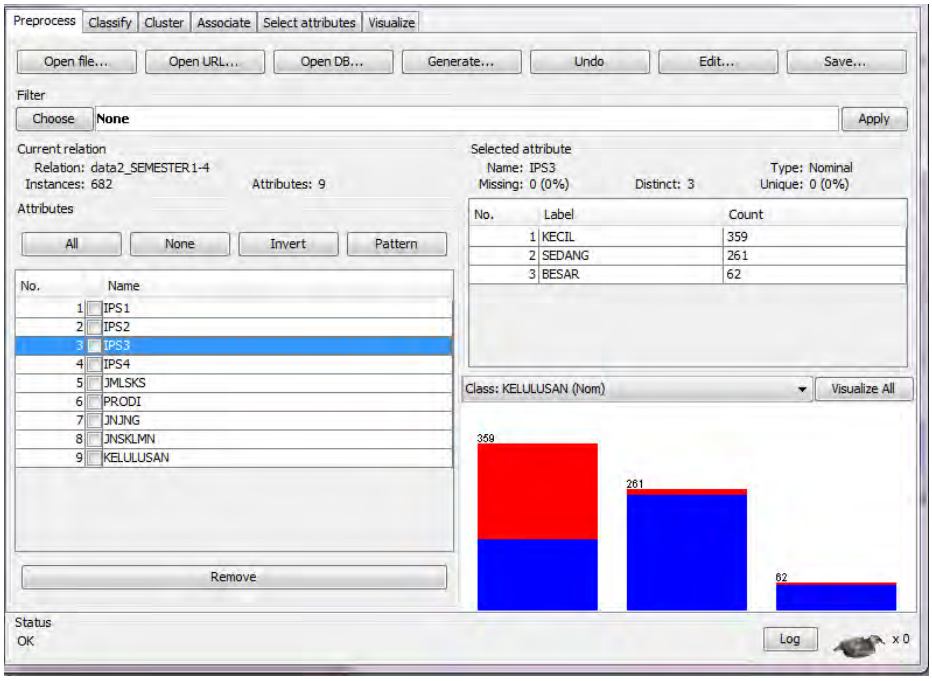
Gambar 3.11 Pengolahan Variabel Kerja



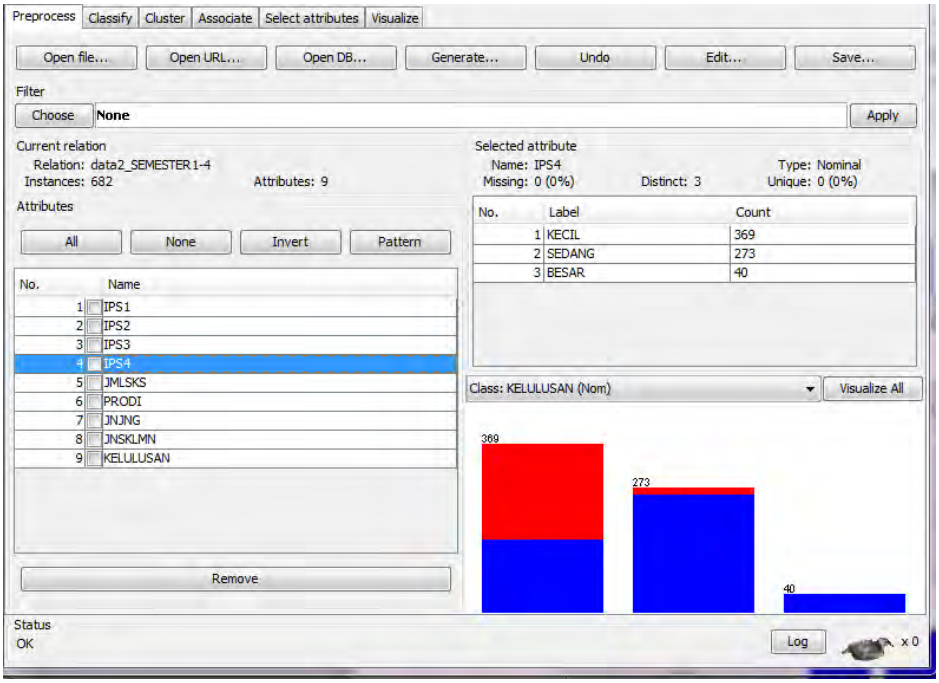
Gambar 3.12 Pengolahan Variabel Indeks Prestasi Semester 1 (IPS1)



Gambar 3.13 Pengolahan Variabel Indeks Prestasi Semester 2 (IPS2)



Gambar 3.14 Pengolahan Variabel Indeks Prestasi Semester 3 (IPS3)



Gambar 3.15 Pengolahan Variabel Indeks Prestasi Semester 4 (IPS4)

3.2.4 Fase Pemodelan *(Modeling Phase)*

Fase pemodelan adalah fase yang secara langsung melibatkan teknik *data mining* yaitu dengan melakuka pemilihan teknik *data mining* dan menentukan algoritma yang akan dilakukan.

Pada fase ini, berbagai jenis teknik pemodelan dipilih dan diaplikasikan serta parameter-parameternya dikalibrasi untuk mendapatkan hasil yang optimal. Biasanya terdapat beberapa teknik untuk jenis permasalahan *data mining* yang sama. Beberapa teknik juga memiliki kebutuhan akan bentuk data yang spesifik. Oleh karena itu, seringkali proses persiapan data dibutuhkan kembali.

Pemodelan yang akan digunakan adalah Algoritma C4.5, Algoritma *Naïve Bayes* dan Algoritma *CART.*

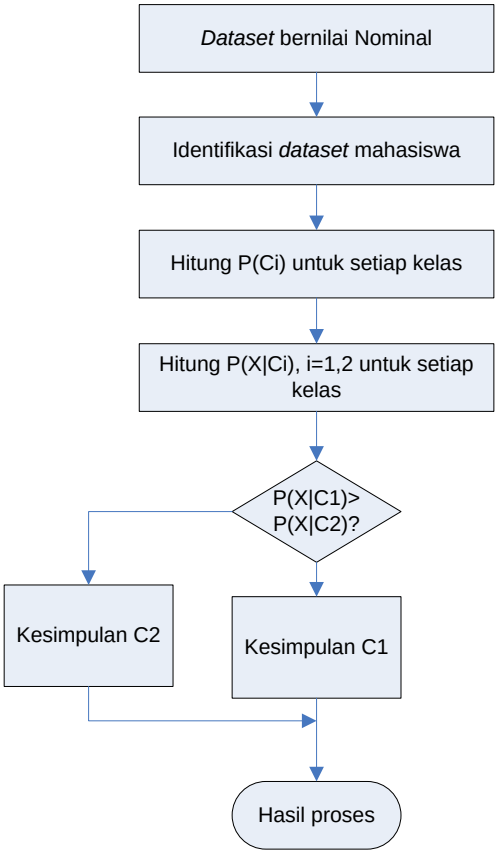
3.2.4.1 Algoritma C4.5

Secara umum langkah-langkah algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

a. Pilih atribut sebagai akar.
b. Buat cabang untuk tiap-tiap nilai.
c. Bagi kasus dalam cabang.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus nomor 2.6.

3.2.4.2 Algoritma *Naïve Bayes*



Gambar 3.16 Diagram Alir Algoritma *Naïve Bayes* (Nugroho, 2014)

Diagram alir *Naïve Bayes* diawali dengan data atribut nominal pada *dataset* mahasiswa, lalu hitung P(Ci) untuk setiap atribut. Dalam kasus *dataset* pada penelitian ini, atribut tahun kelulusan terdiri dari 2 kelas yaitu kelas lulus dan kelas tidak lulus.

Hitung P(X|Ci), i=1,2 untuk setiap kelas atau atribut. Setelah itu bandingkan. Jika P(X|C1)>P(X|C2) maka kesimpulan didapat bahwa kelas lulus. Jika P(X|C1)<P(X|C2) maka kesimpulan didapat bahwa kelas tidak lulus.

Setelah melakukan perhitungan sesuai dengan kaidah *Naïve Bayes*, didapatkan hasil perbandingan kelas C1>C2. Maka sesuai dengan diagram alir *Naïve Bayes* didapatkan kesimpulan “lulus”.

3.2.4.3 Algoritma *CART*

Langkah-langkah penerapan algoritma CART adalah sebagai berikut:

a. Menentukan *Predictor Variable* dan *Target Variable*.
b. Menyusun calon *(Split Candidate)*
c. Melakukan evaluasi performansi terhadap setiap split.
d. Memeriksa apakah masih ada *"diverse node",* apabila tidak ada dapat berhenti pada langkah *c* dan apabila masih ada lakukan evaluasi performansi setiap *split* seperti pada langkah *c*.

3.2.5 Fase Evaluasi *(Evaluation Phase)*

Fase evaluasi adalah mengevaluasi satu atau lebih algoritma yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektifitas sebelum digunakan, menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik, mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining.*

Pada fase ini, algoritma yang tampak memiliki kualitas tinggi dari perspektif analisis data telah dihasilkan. Sebelum dilanjutkan ke tahap penerapan, model yang dihasilkan dievaluasi dan di-*review* tiap langkah pembuatannya untuk memastikan model tesebut telah mencapai tujuan bisnis dengan tepat.

Tujuan utamanya adalah untuk menentukan apakah terdapat beberapa permasalahan bisnis yang tidak dicakup dengan baik. Pada akhir fase ini, keputusan mengenai pengaplikasian hasil *data mining* harus dapat dicapai.

3.2.6 Fase Penyebaran *(Deployment Phase)*

Fase penyebaran adalah fase terakhir yang dilakukan setelah fase-fase sebelumnya telah tertangani dengan baik. Pekerjaan yang dilakukan pada fase ini tergantung pada kebutuhan dan tujuan proyek *data mining*, yaitu dengan pembuatan laporan.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi mengenai pembahasan dan pengujian variabel yang telah ditentukan dengan menerapkan teknik klasifikasi algoritma C4.5, *Naïve Bayes* dan *CART* dengan alat bantu WEKA 3.6.1. Dalam hal ini penelitian dilakukan dengan lima skenario, yaitu:

4.1 Skenario Pertama

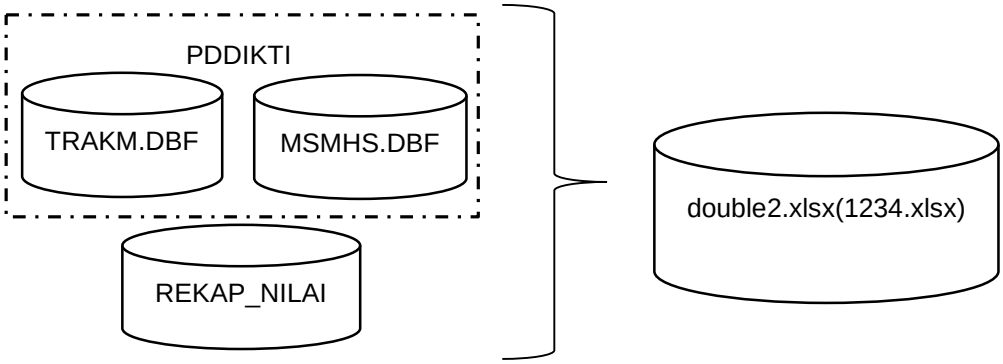
Skenario pertama merupakan percobaan pertama yang dilakukan oleh peneliti dengan fase sebagai berikut:

4.1.1 Fase Pemahaman Bisnis *(Business Understanding Phase)*

Tujuan proyek dalam penelitian ini adalah mengkaji dan membuat model hasil komparasi algoritma C4.5, *Naïve Bayes, CART*, serta menentukan algoritma mana yang paling akurat dan menghasilkan *rule* prediksi kelulusan mahasiswa STMIK Rosma Karawang sehingga dapat dijadikan acuan untuk meningkatkan jumlah kelulusan mahasiswa di tahun-tahun kelulusan berikutnya.

4.1.2 Fase Pemahaman Data *(Data Understanding Phase)*

Dataset mahasiswa yang didapat dari Bagian Akademik (BAAK) STMIK Rosma Karawang berupa dokumen *spreadsheet* dan *DBF*. Sumber data utama yang digunakan dalam penelitian ini adalah *Data* mahasiswa STMIK Rosma Karawang jenjang DIII dan S1 pada tahun 2000 sampai dengan tahun 2011 dengan format xlsx dan DBF. Data dalam tabel 4.1 adalah gabungan dari data yang ada di *folder* REKAP_NILAI, *Data* PDDIKTI yaitu master mahasiswa (MSMHS.DBF) dan transkrip nilai mahasiswa (TRAKM.DBF). Hasil penggabungan dari file di *folder* REKAP_NILAI, *file* MSMHS.DBF dan *file* TRAKM.DBF adalah *file* dengan nama 1234.xlsx seperti terlihat pada gambar 4.1.



Gambar 4.1 *Dataset* Mahasiswa 2000-2011

Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap atau double dan ditemukan banyak nilai kosong atau null yang disebut dengan *missing value.* Sehingga dataset yang didapatkan sejumlah 1234 record seperti terlihat pada tabel 4.1.

Tabel 4.1 *Dataset* Mahasiswa 2000-2011

No	Prodi	Jenjang	Angkatan	Jumlah
1	Manajemen Informatika (57401)	DIII (E)	2001-2011	347
2	Komputerisari Akuntansi (57402)	DIII (E)	2001-2011	283
3	Teknik Informatika (55201)	S1 (C)	2000-2010	405
4	Sistem Komputer (56201)	S1 (C)	2000-2008	30
5	Sistem Informasi (57201)	S1 (C)	2002-2010	169
Total				1234

4.1.3 Fase Pengolahan Data *(Data Preparation Phase)*

Fase ini adalah fase pengolahan yang menggunakan *dataset* dari fase sebelumnya. Variabel yang digunakan antara lain Tempat Lahir (TPTLHR), Jenis Kelamin (JNSKLMN), Jenjang (JNJNG), Program Studi (PRODI), Jumlah SKS (JMLSKS), IPK, Tanggal Kelulusan (KELULUSAN).

Data awal mahasiswa tersebut ditransformasi menjadi data kategori. Hal ini bertujuan agar dapat mempermudah dalam penggalian data dan mudah diproses oleh alat bantu *data mining*. Adapun pengkategorian data sebagai berikut:

(1) Variable TPTLHR (Tempat Lahir)

Jenis datanya dikategorikan Karawang dan Non-Karawang (tempat lahir diluar Karawang)

(2) Variable JNSKLMN (Jenis Kelamin)

Jenis datanya dikategorikan Laki-laki dengan inisial L dan Perempuan dengan inisial P.

(3) Variabel JNJNG (Jenjang)

Jenis datanya dikategorikan seperti pada tabel 4.2.

Tabel 4.2 Kode Jenjang Skenario Pertama

Kode	Jenjang
C	Strata Satu (S1)
E	Diploma Tiga (DIII)

(4) Variabel PRODI (Program Studi)

Kategori Prodi dapat dilihat pada tabel 4.3.

Tabel 4.3 Kode Program Studi Skenario Pertama

Kode	Prodi	Kategori
55201	Teknik Informatika (S1)	TIC
56201	Sistem Komputer (S1)	SKC
57201	Sistem Informasi (S1)	SIC
57401	Manajemen Informatika (D3)	MIE
57402	Komputerisasi Akuntansi (D3)	KAE

(5) Variabel JMLSKS (Jumlah SKS)

Jenis data JMLSKS merupakan data real yang dikategorikan menjadi 4 seperti terlihat pada tabel 4.4.

Tabel 4.4 Kategori Jumlah SKS Skenario Pertama

Kategori SATU	C≥144
Kategori DUA	C<144
Kategori TIGA	E≥110
Kategori EMPAT	E<110

(6) Variabel IPK (Indeks Prestasi Komulatif)

Jenis data IPK dikategorikan menjadi 3 seperti ditampilkan pada tabel 4.5.

Tabel 4.5 Kategori IPK

BESAR	IPK≥3,5
SEDANG	2,75<IPK<3,5
KECIL	IPK≤2,75

(7) Variabel KELULUSAN (Tanggal Kelulusan)

Variabel ini adalah data yang berjenis numerikal yang harus dilakukan proses inisiasi data terlebih dahulu kedalam bentuk nominal. Inisiasi tanggal kelulusan dilakukan dengan:

(a) Mahasiswa dari setiap angkatan yang sudah terdapat tanggal kelulusan dinyatakan “lulus”.

(b) Mahasiswa dari setiap angkatan yang belum terdapat tanggal kelulusan dinyatakan “tidak lulus”.

Tabel 4.6 *Dataset* Mahasiswa tahun 2000-2011 yang Belum Diinisiasi

NIMHSTRAKM	TPLHRMSMHS	KDJEKMSMHS	KDJENMSMHS	KDPSTMSMHS	SKSTTTRAKM	NLIPKTRAKM	TGLLSMSMHS
200301251028	KARAWANG	L	C	55201	154	2,77	11-Jul-09
200401151004	BANDUNG	L	C	57201	159	2,87	14-Jul-09
200401151028	CIAMIS	L	C	57201	155	2,92	11-Jul-09
200501151001	KARAWANG	L	C	57201	148	2,03	26-Agust-10
200501151002	BLITAR	L	C	57201	159	2,96	17-Jul-09
200501151003	BENGKULU	L	C	57201	152	2,58	13-Jul-09
200501151005	KARAWANG	P	C	57201	159	3,1	14-Jul-09
200501151006	SUMEDANG	L	C	57201	145	2,47	20-Agust-11
200501151008	KARAWANG	L	C	57201	148	2,32	20-Agust-11
200501151010	BONDOWOSO	L	C	57201	159	2,74	17-Jul-09
200501151012	LALANG LUAS	L	C	57201	159	3,06	14-Jul-09
200501151014	KARAWANG	P	C	57201	159	3,21	11-Jul-09
200501151015	PEMALANG	L	C	57201	159	3,07	11-Jul-09
200501151016	JAKARTA	P	C	57201	152	2,3	
200501251001	KARAWANG	L	C	55201	152	3,16	11-Jul-09
200501251002	BANDUNG	L	C	55201	162	2,96	11-Jul-09
200501251004	KARAWANG	L	C	55201	162	3,31	11-Jul-09
200501251007	KARAWANG	L	C	55201	162	3,02	15-Jul-09
200501251008	KARAWANG	L	C	55201	159	2,17	11-Jul-09
200501251009	TALANG PETAI	L	C	55201	162	2,93	11-Jul-09
200501251010	MAJALENGKA	L	C	55201	162	3,03	13-Jul-09
200501251014	BEKASI	L	C	55201	162	3,01	13-Jul-09
200501251016	KARAWANG	P	C	55201	162	2,89	13-Jul-09
200501251018	TALANG PETAI	L	C	55201	162	3,34	13-Jul-09

Tabel 4.7 *Dataset* Mahasiswa 2000-2011 yang Siap untuk Perangkat Pemodelan

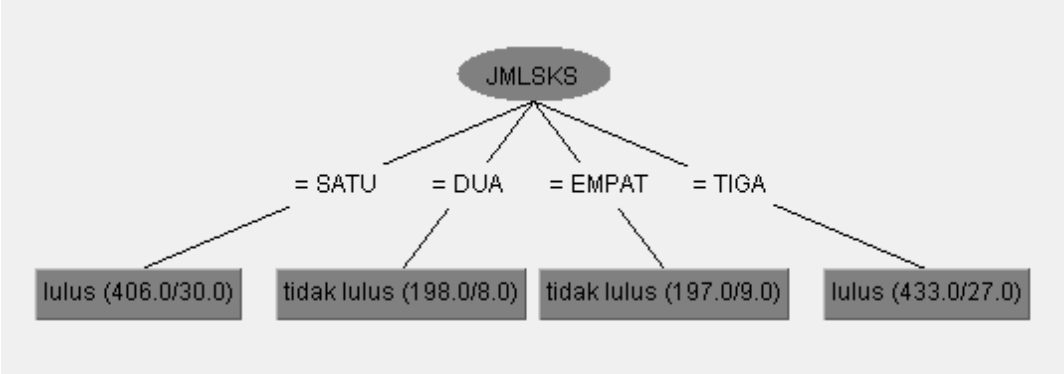
TPTLHR	JNSKLMN	JNJNG	PRODI	JMLSKS	IPK	KELULUSAN
karawang	L	C	TIC	SATU	KECIL	lulus
karawang	P	C	TIC	DUA	SEDANG	tidak lulus
karawang	L	C	TIC	SATU	KECIL	tidak lulus
non karawang	P	C	TIC	SATU	SEDANG	lulus
karawang	P	C	TIC	SATU	SEDANG	lulus
karawang	L	C	TIC	DUA	KECIL	tidak lulus
non karawang	P	C	TIC	SATU	SEDANG	lulus
non karawang	L	C	TIC	SATU	KECIL	lulus
karawang	P	C	TIC	SATU	SEDANG	lulus
karawang	L	C	TIC	SATU	KECIL	lulus
karawang	L	C	TIC	SATU	KECIL	lulus
karawang	P	C	TIC	SATU	SEDANG	tidak lulus
non karawang	P	C	TIC	SATU	SEDANG	tidak lulus
karawang	L	C	TIC	SATU	SEDANG	tidak lulus
karawang	L	C	TIC	SATU	KECIL	tidak lulus
non karawang	L	C	TIC	SATU	KECIL	lulus
non karawang	L	C	TIC	SATU	SEDANG	tidak lulus
non karawang	P	C	TIC	SATU	KECIL	lulus
non karawang	P	C	TIC	SATU	SEDANG	lulus

4.1.4 Fase Pemodelan *(Modeling Phase)*

Fase pemodelan akan secara langsung terlibat dalam teknik-teknik *data mining* yaitu dengan melakukan pemilihan teknik dan menentukan algoritma yang akan dilakukan. Dalam hal ini akan dilakukan tiga pemodelan yaitu:

4.1.4.1 Pemodelan Algoritma C4.5

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2000-2011 seperti pada tabel 4.7 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1. Gambar 4.2 adalah pohon keputusan yang terbentuk.



Gambar 4.2 Pohon Keputusan Akhir dari *Dataset* Mahasiswa tahun 2000-2011

Pada pohon keputusan yang terbentuk, hanya Jumlah SKS menjadi simpul akar.

Tabel 4.8 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma C4.5 pada skenario pertama.

Tabel 4.8 *Confusion Matrix* Algoritma C4.5 Skenario Pertama

```
=== Confusion Matrix ===

   a   b   <-- classified as
 782  17 |   a = lulus
  57 378 |   b = tidak lulus
```

Akurasi = $\frac{782+378}{782+17+57+378} x\ 100\% = 94{,}0032\%$.

Nilai kurva *ROC* yang dihasilkan adalah 0,913.

4.1.4.2 Pemodelan Algoritma *Naïve Bayes*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2000-2011 seperti pada tabel 4.7 dengan algoritma *Naïve Bayes* dan diolah dengan alat

bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.9 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *Naïve Bayes* pada skenario pertama.

Tabel 4.9 *Confusion Matrix* Algoritma *Naïve Bayes* Skenario Pertama

```
=== Confusion Matrix ===

   a   b   <-- classified as
 782  17 |   a = lulus
  57 378 |   b = tidak lulus
```

Akurasi = $\frac{782+378}{782+17+57+378} x\ 100\%$ = 94,0032%.

Nilai kurva *ROC* yang dihasilkan adalah 0,953.

4.1.4.3 Pemodelan Algoritma *CART*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2000-2011 seperti pada tabel 4.7 dengan algoritma *CART* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.10 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *CART* pada skenario pertama.

Tabel 4.10 *Confusion Matrix* Algoritma *CART* Skenario Pertama

```
=== Confusion Matrix ===

   a   b   <-- classified as
 782  17 |   a = lulus
  57 378 |   b = tidak lulus
```

Akurasi = $\frac{782+378}{782+17+57+378} x\ 100\%$ = 94,0032%.

Nilai kurva *ROC* yang dihasilkan adalah 0,907.

4.1.5 Fase Evaluasi *(Evaluation Phase)*

Fase evaluasi adalah fase untuk melakukan evaluasi ketiga algoritma yang dipakai dalam fase pemodelan yaitu Algoritma C4.5, *Naïve Bayes,* dan *CART.* Evaluasi dari keputusan akhir yang terbentuk mengindikasikan bahwa pola yang dihasilkan belum dapat memberikan informasi dan pengetahuan yang sesuai dengan tujuan proyek walaupun tingkat akurasi yang tinggi dan nilai kurva ROC yang dihasilkan dari tiga

algoritma yang digunakan dalam fase pemodelan adalah 0,90 – 1,00 adalah termasuk dalam kelompok klasifikasi sangat baik..

Menurut peneliti, Jumlah SKS total adalah informasi yang kurang menarik karena Jumlah SKS total merupakan salah satu syarat kelulusan sehingga tanpa perlu proses *mining* kelulusan sudah dapat diketahui. Dengan demikian peneliti melakukan skenario kedua.

4.2 Skenario Kedua

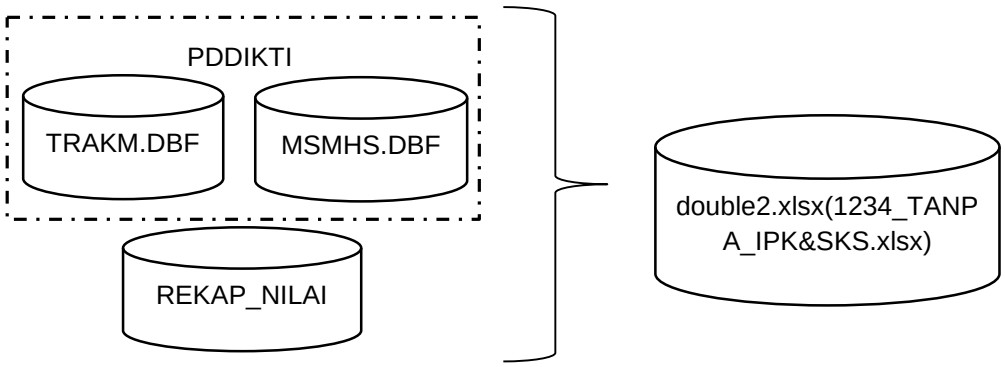
Skenario kedua merupakan percobaan kedua setelah pada skenario pertama tidak mendapatkan informasi yang menarik.

4.2.1 Fase Pemahaman Bisnis *(Business Understanding Phase)*

Fase pemahaman bisnis ini sudah dipaparkan pada bab III.

4.2.2 Fase Pemahaman Data *(Data Understanding Phase)*

Dataset mahasiswa yang didapat dari Bagian Akademik (BAAK) STMIK Rosma Karawang berupa dokumen *spreadsheet* dan *DBF*. Sumber data utama yang digunakan dalam penelitian ini adalah *Data* mahasiswa STMIK Rosma Karawang jenjang DIII dan S1 pada tahun 2000 sampai dengan tahun 2011 dengan format xlsx dan DBF. Hasil penggabungan dari file di *folder* REKAP_NILAI, *file* MSMHS.DBF dan *file* TRAKM.DBF adalah *file* dengan nama 1234_TANPA_IPK&SKS.xlsx seperti terlihat pada gambar 4.3.



Gambar 4.3 *Dataset* Mahasiswa 2000-2011 Tanpa IPK dan Jumlah SKS

4.2.3 Fase Pengolahan Data *(Data Preparation Phase)*

Fase pengolahan data skenario kedua menggunakan variabel Tempat Lahir (TPTLHR), Jenis Kelamin (JNSKLMN), Jenjang (JNJNG), Program Studi (PRODI), Tanggal Kelulusan (KELULUSAN) tanpa menyertakan variabel Jumlah SKS (JMLSKS) dan IPK. Tabel 4.11 adalah *dataset* yang siap untuk perangkat pemodelan.

Tabel 4.11 *Dataset* Mahasiswa 2000-2011 tanpa Variabel

Jumlah SKS dan IPK

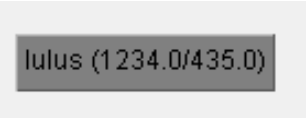
TPTLHR	JNSKLMN	JNJNG	PRODI	KELULUSAN
karawang	L	C	TIC	lulus
karawang	P	C	TIC	tidak lulus
karawang	L	C	TIC	tidak lulus
non karawang	P	C	TIC	lulus
karawang	P	C	TIC	lulus
karawang	L	C	TIC	tidak lulus
non karawang	P	C	TIC	lulus
non karawang	L	C	TIC	lulus
karawang	P	C	TIC	lulus
karawang	L	C	TIC	lulus
karawang	L	C	TIC	lulus
karawang	P	C	TIC	tidak lulus
non karawang	P	C	TIC	tidak lulus
karawang	L	C	TIC	tidak lulus
karawang	L	C	TIC	tidak lulus
non karawang	L	C	TIC	lulus
non karawang	L	C	TIC	tidak lulus
non karawang	P	C	TIC	lulus
non karawang	P	C	TIC	lulus

4.2.4 Fase Pemodelan *(Modeling Phase)*

Fase pemodelan pada skenario kedua sama dengan fase pemodelan pada skenario pertama yaitu dengan melakukan tiga pemodelan algoritma C4.5, *Naïve Bayes, CART.*

4.2.4.1 Algoritma C4.5

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2000-2011 seperti pada tabel 4.11 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1. Gambar 4.4 adalah pohon keputusan yang terbentuk.



Gambar 4.4 Pohon Keputusan Yang dihasilkan Dari Dataset Mahasiswa tahun 2000-2011 tanpa variabel Jumlah SKS dan IPK

Pada pohon keputusan yang terbentuk tidak terdapat pola yang sesuai dengan tujuan proyek. Tabel 4.12 adalah *Confusion Matrix* yang terbentuk dari pemodelan algoritma C4.5 pada skenario kedua.

Tabel 4.12 *Confusion Matrix* Algoritma C4.5 Skenario Kedua

```
=== Confusion Matrix ===

   a   b   <-- classified as
 799   0 |   a = lulus
 435   0 |   b = tidak lulus
```

Akurasi = $\frac{799+0}{799+0+0+435} x\ 100\%$ = 64,7488%

Nilai kurva *ROC* yang dihasilkan adalah 0,497.

4.2.4.2 Algoritma *Naïve Bayes*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2000-2011 seperti pada tabel 4.11 dengan algoritma *Naïve Bayes* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.13 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *Naïve Bayes* pada skenario kedua.

Tabel 4.13 *Confusion Matrix* Algoritma *Naïve Bayes* Skenario Kedua

```
=== Confusion Matrix ===

   a   b   <-- classified as
 785  14 |   a = lulus
 419  16 |   b = tidak lulus
```

Akurasi = $\frac{785+16}{785+16+14+419} x\ 100\%$ = 64,9109%

Nilai kurva *ROC* yang dihasilkan adalah 0,576.

4.2.4.3 Algoritma *CART*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2000-2011 seperti pada tabel 4.11 dengan algoritma *CART* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.14 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *CART* pada skenario kedua.

Tabel 4.14 *Confusion Matrix* Algoritma *CART* Skenario Kedua

```
=== Confusion Matrix ===

   a   b   <-- classified as
 799   0 |   a = lulus
 435   0 |   b = tidak lulus
```

Akurasi = $\frac{799+0}{799+0+0+435} x\ 100\%$ = 64,7488%

Nilai kurva *ROC* yang dihasilkan adalah 0,497.

4.2.5 Fase Evaluasi *(Evaluation Phase)*

Fase evaluasi adalah fase untuk melakukan evaluasi ketiga algoritma yang digunakan. Evaluasi dari keputusan akhir yang terbentuk mengindikasikan bahwa tidak ditemukan pola serta nilai akurasi yang rendah dari ketiga algoritma dan nilai kurva ROC yang dihasilkan dari tiga algoritma yang digunakan dalam fase pemodelan adalah 0,60 – 0,70 adalah termasuk dalam kelompok klasifikasi buruk. Dengan demikian peneliti melakukan skenario ketiga.

4.3 Skenario Ketiga

Skenario ketiga merupakan percobaan ketiga setelah pada skenario kedua tidak mendapatkan informasi yang sesuai dengan tujuan proyek dan menghasilkan klasifikasi yang buruk.

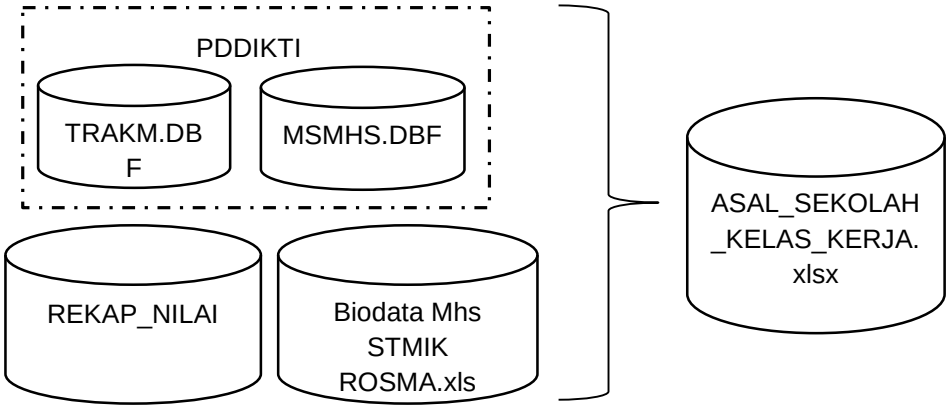
4.3.1 Fase Pemahaman Bisnis *(Business Understanding Phase)*

Fase pemahaman bisnis ini sudah dipaparkan pada bab III.

4.3.2 Fase Pemahaman Data *(Data Understanding Phase)*

Dataset mahasiswa yang didapat dari Bagian Akademik (BAAK) STMIK Rosma Karawang berupa dokumen *spreadsheet* dan *DBF*. Sumber data utama yang digunakan

dalam penelitian ini adalah *Data* mahasiswa STMIK Rosma Karawang jenjang DI, DIII dan S1 pada tahun 2005 sampai dengan tahun 2009 dengan format xlsx dan DBF. Data dalam tabel 4.15 adalah gabungan dari data yang ada di *folder* REKAP_NILAI, *Data* PDDIKTI yaitu master mahasiswa (MSMHS.DBF) dan transkrip nilai mahasiswa (TRAKM.DBF). Hasil penggabungan dari file di *folder* REKAP_NILAI, *file* MSMHS.DBF dan *file* TRAKM.DBF adalah *file* dengan nama ASAL_SEKOLAH_KELAS_KERJA.xlsx seperti terlihat pada gambar 4.5.



Gambar 4.5 *Dataset* Mahasiswa 2005-2009

Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap atau double dan ditemukan banyak nilai kosong atau null yang disebut dengan *missing value.* Sehingga dataset yang didapatkan sejumlah 585 *record* seperti terlihat pada tabel 4.15.

Tabel 4.15 Data Jumlah Mahasiswa STMIK Rosma 2005-2009

No	Prodi	Jenjang	Jumlah
1	Manajemen Informatika (57401)	DIII (E)	125
2	Komputerisari Akuntansi (57402)	DIII (E)	111
3	Teknik Informatika (55201)	S1 (C)	213
4	Sistem Komputer (56201)	S1 (C)	17
5	Sistem Informasi (57201)	S1 (C)	98
6	Manajemen Informatika (57601)	D1 (G)	8
7	Komputerisasi Akuntansi (57602)	D1 (G)	13
Total			585

4.3.3 Fase Pengolahan Data *(Data Preparation Phase)*

Fase ini adalah fase pengolahan yang menggunakan *dataset* dari fase sebelumnya. Variabel yang digunakan antara lain JENIS KELAMIN, JENJANG, PRODI, JMLSKS, IPK, SEKOLAH, KELAS, KERJA, Tanggal Kelulusan (KELULUSAN).

Data awal mahasiswa tersebut ditransformasi menjadi data kategori. Hal ini bertujuan agar dapat mempermudah dalam penggalian data dan mudah diproses oleh alat bantu *data mining*. Adapun pengkategorian data sebagai berikut:

i. Variabel JENIS KELAMIN

Jenis datanya dikategorikan Laki-laki dengan inisial L dan Perempuan dengan inisial P.

ii. Variabel JENJANG

Jenis datanya dikategorikan seperti pada tabel 4.16.

Tabel 4.16 Kode Jenjang Skenario Ketiga

Kode	Jenjang
C	Strata Satu (S1)
E	Diploma Tiga (DIII)
G	Diploma Satu (DI)

iii. Variabel PRODI (Program Studi)

Kategori Prodi dapat dilihat pada tabel 4.17.

Tabel 4.17 Kode Program Studi Skenario Ketiga

Kode	Prodi	Kategori
55201	Teknik Informatika (S1)	TIC
56201	Sistem Komputer (S1)	SKC
57201	Sistem Informasi (S1)	SIC
57401	Manajemen Informatika (D3)	MIE
57402	Komputerisasi Akuntansi (D3)	KAE
57601	Manajemen Informatika (D1)	MIG
57602	Komputerisasi Akuntansi (D1)	KAG

iv. Variabel JMLSKS (Jumlah SKS)

Jenis data JMLSKS merupakan data real yang dikategorikan menjadi 6 seperti terlihat pada tabel 4.18.

Tabel 4.18 Kategori Jumlah SKS Skenario Ketiga

Kategori	Keterangan
SATU	C≥144
DUA	C<144
TIGA	E≥110
EMPAT	E<110
LIMA	G≥40
ENAM	G≤40

v. Variabel IPK (Indeks Prestasi Komulatif)

Jenis data IPK dikategorikan menjadi 3 seperti ditampilkan pada tabel 4.5.

vi. Variabel SEKOLAH

Variabel ini dikategorikan menjadi 3, yaitu SMA. SMK dan MA.

vii. Variabel KELAS

Variabel ini dikategorikan menjadi 2, yaitu Kelas Pagi dan Kelas Malam.

viii. Variabel KERJA

Variabel ini dikategorikan menjadi 2 seperti terlihat pada tabel 4.19.

Tabel 4.19 Kategori KERJA

Kategori	Keterangan
Ya	Mahasiswa Bekerja
Tidak	Mahasiswa Tidak Bekerja

ix. Variabel KELULUSAN (Tanggal Kelulusan)

Variabel ini adalah data yang berjenis numerikal yang harus dilakukan proses inisiasi data terlebih dahulu kedalam bentuk nominal. Inisiasi tahun lulus dilakukan dengan:

(a) Mahasiswa dari setiap angkatan yang sudah terdapat tahun kelulusan dinyatakan “lulus”.

(b) Mahasiswa dari setiap angkatan yang belum terdapat tahun kelulusan dinyatakan “tidak lulus”.

Tabel 4.20 *Dataset* Mahasiswa 2005-2009 yang sudah diinisiasi

NIM	TEMPAT LAHIR	JENIS KELAMIN	JENJANG	PRODI	JUMLAH SKS	IPK	KELULUSAN	ASAL SEKOLAH	KELAS	KERJA
200501151001	KARAWANG	L	C	57201	159	2,96	lulus	SMAN 5 Karawang	Pagi	Tidak
200501151002	BLITAR	L	C	57201	152	2,91	lulus	SMK Telkom	Malam	Ya
200501151003	BENGKULU	L	C	57201	159	3,1	lulus	SMAN 2 Padang	Pagi	Tidak
200501151005	KARAWANG	P	C	57201	145	2,81	lulus	SMAN 1 Tempuran	Pagi	Tidak
200501151006	SUMEDANG	L	C	57201	148	2,32	lulus	SMKN 7 Bandung	Malam	Ya
200501151008	KARAWANG	L	C	57201	159	2,74	lulus	SMU Sunan Gunung Jati	Pagi	Tidak
200501151010	BONDOWOSO	L	C	57201	159	3,06	lulus	SMAN 5 Karawang	Pagi	Tidak
200501151012	LALANG LUAS	L	C	57201	159	3,21	lulus	SMK Telkom	Malam	Ya
200501151014	KARAWANG	P	C	57201	159	3,07	lulus	SPK DepKes RI Yogyakarta	Pagi	Tidak
200501151015	PEMALANG	L	C	57201	152	2,3	lulus	SMAN 1 Lubuk Pinang	Malam	Ya
200501151016	JAKARTA	P	C	57201	152	3,16	tidak lulus	SMAN 4 Karawang	Pagi	Tidak
200501251001	KARAWANG	L	C	55201	162	2,96	lulus	SMK Texmaco	Pagi	Tidak
200501251002	BANDUNG	L	C	55201	162	3,31	lulus	STM Negeri 4 Bandung	Malam	Ya
200501251004	KARAWANG	L	C	55201	162	3,02	lulus	SMUN 1 Telukjambe	Malam	Ya
200501251007	KARAWANG	L	C	55201	159	2,17	lulus	SMU Yos Sudarso Karawang	Pagi	Tidak
200501251008	KARAWANG	L	C	55201	162	2,93	lulus	SMK Bakti Purwokerto	Pagi	Tidak
200501251009	TALANG PETAI	L	C	55201	162	3,03	lulus	MA. YPPA Cipulus	Pagi	Tidak
200501251010	MAJALENGKA	L	C	55201	162	3,01	lulus	SMAN 5 Karawang	Pagi	Tidak
200501251014	BEKASI	L	C	55201	162	2,89	lulus	SMUN 1 Rengasdengklok	Pagi	Tidak
200501251016	KARAWANG	P	C	55201	162	3,34	lulus	SMA Bhineka Karawang	Malam	Ya
200501251018	TALANG PETAI	L	C	55201	162	2,6	lulus	SMKN 1 Cirebon	Malam	Ya
200501251019	JAKARTA	L	C	55201	162	2,86	lulus	SMK Taruna Karya	Pagi	Tidak
200501251020	KARAWANG	P	C	55201	162	2,86	lulus	MAN Rengasdengkok	Malam	Ya

Tabel 4.21 *Dataset* Mahasiswa 2005-2009 yang Siap Perangkat Pemodelan

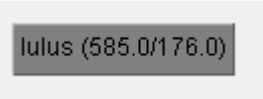
JENIS KELAMIN	JENJANG	PRODI	JMLSKS	IPK	SEKOLAH	KELAS	KERJA	KELULUSAN
L	C	SIC	SATU	SEDANG	SMA	Pagi	Tidak	lulus
L	C	SIC	SATU	SEDANG	SMK	Malam	Ya	lulus
L	C	SIC	SATU	SEDANG	SMA	Pagi	Tidak	lulus
P	C	SIC	SATU	SEDANG	SMK	Pagi	Tidak	lulus
L	C	SIC	SATU	KECIL	SMK	Malam	Ya	lulus
L	C	SIC	SATU	KECIL	SMA	Pagi	Tidak	lulus
L	C	SIC	SATU	SEDANG	SMA	Pagi	Tidak	lulus
L	C	SIC	SATU	SEDANG	SMK	Malam	Ya	lulus
P	C	SIC	SATU	SEDANG	SMK	Pagi	Tidak	lulus
L	C	SIC	SATU	KECIL	SMA	Malam	Ya	lulus
P	C	SIC	SATU	SEDANG	SMA	Pagi	Tidak	tidak lulus
L	C	TIC	SATU	SEDANG	SMK	Pagi	Tidak	lulus
L	C	TIC	SATU	SEDANG	SMK	Malam	Ya	lulus
L	C	TIC	SATU	SEDANG	SMA	Malam	Ya	lulus
L	C	TIC	SATU	KECIL	SMA	Pagi	Tidak	lulus
L	C	TIC	SATU	SEDANG	SMK	Pagi	Tidak	lulus
L	C	TIC	SATU	SEDANG	MA	Pagi	Tidak	lulus
L	C	TIC	SATU	SEDANG	SMA	Pagi	Tidak	lulus
L	C	TIC	SATU	SEDANG	SMA	Pagi	Tidak	lulus
P	C	TIC	SATU	SEDANG	SMA	Malam	Ya	lulus
L	C	TIC	SATU	KECIL	SMK	Malam	Ya	lulus
L	C	TIC	SATU	SEDANG	SMK	Pagi	Tidak	lulus
P	C	TIC	SATU	SEDANG	MA	Malam	Ya	lulus
L	C	TIC	SATU	KECIL	SMK	Pagi	Tidak	lulus

4.3.4 Fase Pemodelan *(Modeling Phase)*

Fase pemodelan pada skenario k etiga sama dengan fase pemodelan pada skenario pertama dan kedua dengan melakukan tiga pemodelan yaitu:

4.3.4.1 Algoritma C4.5

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2009 seperti pada tabel 4.21 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1. Gambar 4.6 adalah pohon keputusan yang terbentuk.



Gambar 4.6 Pohon Keputusan Yang dihasilkan Dari *Dataset* Mahasiswa tahun 2005-2009 pada skenario ketiga

Pada pohon keputusan yang terbentuk tidak terdapat pola yang sesuai dengan tujuan proyek. Tabel 4.22 adalah *Confusion Matrix* yang terbentuk dari pemodelan algoritma C4.5 pada skenario kedua.

Tabel 4.22 *Confusion Matrix* Algoritma C4.5 Skenario Ketiga

```
=== Confusion Matrix ===

   a   b   <-- classified as
 400   9 |   a = lulus
 167   9 |   b = tidak lulus
```

Akurasi = $\frac{400+9}{400+9+9+167} x\ 100\%$ = 69.9145%

Nilai kurva *ROC* yang dihasilkan adalah 0,525.

4.3.4.2 Algoritma *Naïve Bayes*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2005-2009 seperti pada tabel 4.21 dengan algoritma *Naïve Bayes* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.23 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *Naïve Bayes* pada skenario kedua.

Tabel 4.23 *Confusion Matrix* Algoritma *Naïve Bayes* Skenario Ketiga

```
=== Confusion Matrix ===

   a   b   <-- classified as
 369  40 |   a = lulus
 131  45 |   b = tidak lulus
```

Akurasi = $\frac{369+45}{369+45+40+131} x\ 100\%$ = 70,7692%.

Nilai kurva *ROC* yang dihasilkan adalah 0,659.

4.3.4.3 Algoritma *CART*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2005-2009 seperti pada tabel 4.21 dengan algoritma *CART* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.24 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *CART* pada skenario ketiga.

Tabel 4.24 *Confusion Matrix* Algoritma *CART* Skenario Ketiga

```
=== Confusion Matrix ===

   a   b   <-- classified as
 386  23 |   a = lulus
 154  22 |   b = tidak lulus
```

Akurasi = $\frac{386+22}{386+22+23+154} x\ 100\%$ = 69,7436%

Nilai kurva *ROC* yang dihasilkan adalah 0,594.

4.3.5 Fase Evaluasi *(Evaluation Phase)*

Evaluasi dari keputusan akhir yang terbentuk mengindikasikan bahwa tidak ditemukan pola serta nilai akurasi yang rendah dari ketiga algoritma dan nilai kurva *ROC* yang dihasilkan dari tiga algoritma yang digunakan dalam fase pemodelan adalah 0,50 – 0,60 adalah termasuk dalam kelompok klasifikasi salah. Dengan demikian peneliti melakukan skenario keempat.

4.4 Skenario Keempat

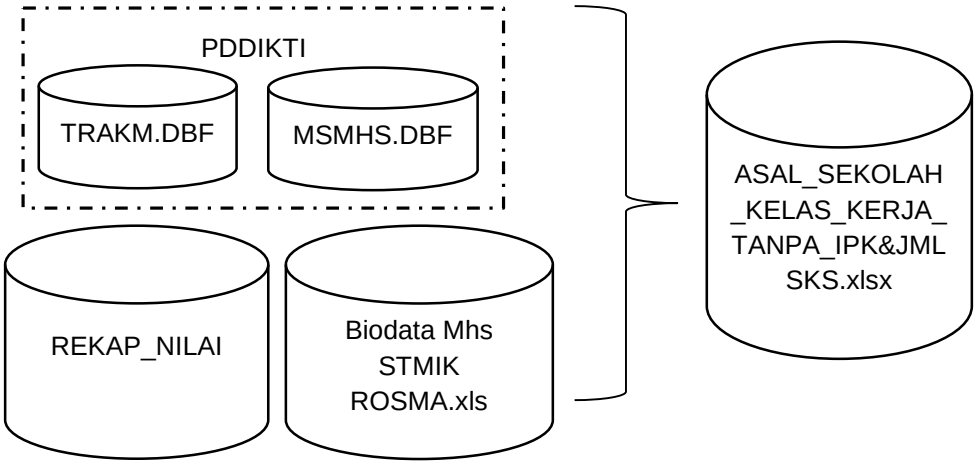
Skenario keempat merupakan percobaan keempat setelah pada skenario ketiga menghasilkan klasifikasi yang salah.

4.4.1 Fase Pemahaman Bisnis *(Business Understanding Phase)*

Fase pemahaman bisnis pada skenario keempat ini masih sama dengan fase pemahaman bisnis pada skenario-skenario sebelumnya.

4.4.2 Fase Pemahaman Data *(Data Understanding Phase)*

Sumber data utama yang digunakan dalam penelitian ini adalah *dataset* mahasiswa STMIK Rosma Karawang jenjang DI, DIII dan S1 pada tahun 2005 sampai dengan tahun 2009 dengan format xlsx dan DBF. Hasil penggabungan dari file di *folder* REKAP_NILAI, *file* Biodata Mhs STMIK Rosma, *file* MSMHS.DBF dan *file* TRAKM.DBF adalah *file* dengan nama ASAL_SEKOLAH_KELAS_KERJA_TANPA_IPK&JMLSKS.xlsx seperti terlihat pada gambar 4.7.



Gambar 4.7 Dataset Mahasiswa 2005-2009 Tanpa variabel

IPK dan Jumlah SKS

4.4.3 Fase Pengolahan Data *(Data Preparation Phase)*

Fase pengolahan data skenario keempat menggunakan variabel Jenis Kelamin, Jenjang, Program Studi (PRODI), Sekolah, Kelas, Kerja, Tanggal Kelulusan (KELULUSAN) tanpa menyertakan variabel IPK dan Jumlah SKS (JMLSKS). Tabel 4.25 adalah *dataset* yang siap untuk perangkat pemodelan.

Tabel 4.25 *Dataset* Mahasiswa 2005-2009 tanpa variabel

IPK dan Jumlah SKS yang Siap Perangkat Pemodelan

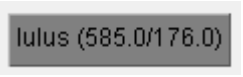
JENIS KELAMIN	JENJANG	PRODI	SEKOLAH	KELAS	KERJA	KELULUSAN
L	C	SIC	SMA	Pagi	Tidak	lulus
L	C	SIC	SMK	Malam	Ya	lulus
L	C	SIC	SMA	Pagi	Tidak	lulus
P	C	SIC	SMK	Pagi	Tidak	lulus
L	C	SIC	SMK	Malam	Ya	lulus
L	C	SIC	SMA	Pagi	Tidak	lulus
L	C	SIC	SMA	Pagi	Tidak	lulus
L	C	SIC	SMK	Malam	Ya	lulus
P	C	SIC	SMK	Pagi	Tidak	lulus
L	C	SIC	SMA	Malam	Ya	lulus
P	C	SIC	SMA	Pagi	Tidak	tidak lulus
L	C	TIC	SMK	Pagi	Tidak	lulus
L	C	TIC	SMK	Malam	Ya	lulus
L	C	TIC	SMA	Malam	Ya	lulus
L	C	TIC	SMA	Pagi	Tidak	lulus
L	C	TIC	SMK	Pagi	Tidak	lulus
L	C	TIC	MA	Pagi	Tidak	lulus
L	C	TIC	SMA	Pagi	Tidak	lulus
L	C	TIC	SMA	Pagi	Tidak	lulus
P	C	TIC	SMA	Malam	Ya	lulus
L	C	TIC	SMK	Malam	Ya	lulus
L	C	TIC	SMK	Pagi	Tidak	lulus
P	C	TIC	MA	Malam	Ya	lulus
L	C	TIC	SMK	Pagi	Tidak	lulus

4.4.4 Fase Pemodelan *(Modelig Phase)*

Fase pemodelan pada skenario k eempat sama dengan fase pemodelan pada skenario pertama, kedua dan ketiga, dengan melakukan tiga pemodelan yaitu:

4.4.4.1 Algoritma C4.5

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2009 seperti pada tabel 4.25 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1 dengan algoritma C4.5. Gambar 4.8 adalah pohon keputusan yang terbentuk.



Gambar 4.8 Pohon Keputusan Yang dihasilkan Dari *Dataset* Mahasiswa

tahun 2005-2009 pada skenario keempat

Pada pohon keputusan yang terbentuk tidak terdapat pola yang sesuai dengan tujuan proyek. Tabel 4.26 adalah *Confusion Matrix* yang terbentuk dari pemodelan algoritma C4.5 pada skenario keempat.

Tabel 4.26 *Confusion Matrix* Algoritma C4.5 Skenario Keempat

```
=== Confusion Matrix ===

   a   b   <-- classified as
 395  14 |   a = lulus
 170   6 |   b = tidak lulus
```

Akurasi = $\frac{395+6}{395+6+14+170} x\ 100\%$ = 68,547%.

Nilai kurva *ROC* yang dihasilkan adalah 0,537.

4.4.4.2 Algoritma Naïve Bayes

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2009 seperti pada tabel 4.25 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1 dengan algoritma *Naïve Bayes*. Hasil yang didapat belum memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.26 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *Naïve Bayes* pada skenario keempat.

Tabel 4.27 *Confusion Matrix* Algoritma *Naïve Bayes* Skenario Keempat

```
=== Confusion Matrix ===

   a   b   <-- classified as
 378  31 |   a = lulus
 148  28 |   b = tidak lulus
```

Akurasi = $\frac{378+28}{378+28+31+148} x\ 100\%$ = 69,4017%.

Nilai kurva *ROC* yang dihasilkan adalah 0,647.

4.4.4.3 Algoritma CART

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2005-2009 seperti pada tabel 4.25 dengan algoritma *CART* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat belum memberikan informasi atau

pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.28 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *CART* pada skenario keempat.

Tabel 4.28 *Confusion Matrix* Algoritma *CART* Skenario Keempat

```
=== Confusion Matrix ===

   a   b   <-- classified as
 379  30 |   a = lulus
 148  28 |   b = tidak lulus
```

Akurasi = $\frac{379+28}{379+28+30+148} x\ 100\%$ = 69.5726%

Nilai kurva *ROC* yang dihasilkan adalah 0,601.

4.4.5 Fase Evaluasi *(Evaluation Phase)*

Evaluasi dari keputusan akhir yang terbentuk mengindikasikan bahwa tidak ditemukan pola serta nilai akurasi yang rendah dari ketiga algoritma. Nilai kurva *ROC* yang dihasilkan dari tiga algoritma yang digunakan dalam fase pemodelan adalah 0,50 – 0,60 adalah termasuk dalam kelompok klasifikasi salah. Dengan demikian peneliti melakukan skenario kelima.

4.5 Skenario Kelima

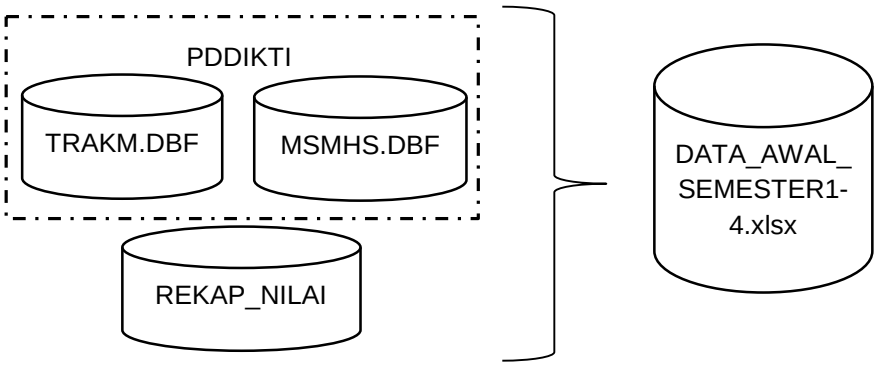
Skenario keempat merupakan percobaan keempat setelah pada skenario ketiga menghasilkan klasifikasi yang salah.

4.5.1 Fase Pemahaman Bisnis *(Business Understanding Phase)*

Fase pemahaman bisnis pada skenario kelima ini masih sama dengan fase pemahaman bisnis pada skenario-skenario sebelumnya yang telah dipaparkan pada bab III.

4.5.2 Fase Pemahaman Data *(Data Understanding Phase)*

Sumber data utama yang digunakan dalam penelitian ini adalah data mahasiswa STMIK Rosma Karawang jenjang DIII dan S1 pada tahun 2005 sampai dengan tahun 2008 dengan format xlsx dan DBF. Data dalam tabel 4.29 adalah gabungan dari data yang ada di folder REKAP_NILAI, data PDDIKTI yaitu master mahasiswa (MSMHS.DBF) dan transkrip nilai (TRAKM.DBF). Hasil penggabungan tersebut adalah DATA_AWAL_SEMESTER1-4.xlsx seperti pada gambar 4.9.



Gambar 4.9 *Dataset* Mahasiswa 2005-2008

Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap atau double dan ditemukan banyak nilai kosong atau null yang disebut dengan *missing value.* Sehingga dataset yang didapatkan sejumlah 682 *record* seperti terlihat pada tabel 4.29

Tabel 4.29 *Dataset* Jumlah Mahasiswa STMIK Rosma 2005-2008

No	Prodi	Jenjang	Angkatan	Jumlah
1	Manajemen Informatika (57401)	DIII (E)	2005-2008	197
2	Komputerisari Akuntansi (57402)	DIII (E)	2005-2008	142
3	Teknik Informatika (55201)	S1 (C)	2005-2008	227
4	Sistem Komputer (56201)	S1 (C)	2005-2008	25
5	Sistem Informasi (57201)	S1 (C)	2005-2008	91
Total				682

4.5.3 Fase Pengolahan Data *(Data Preparation Phase)*

Data mahasiswa terdiri dari beberapa variabel antara lain IP Semester 1 (IPS1), IP Semester 2 (IPS2), IP Semester 3 (IPS3), IP Semester 4 (IPS4), Jumlah SKS yang Sudah ditempuh (JMLSKS), PRODI, Jenjang (JNJNG), Jenis Kelamin (JNSKLMN), Tanggal Kelulusan (KELULUSAN). Pengkategorian datanya sebagai berikut:

(1) Variabel Indeks Prestasi Semester (IPS)

Jenis data IPS diambil dari semester 1, semester 2, semester 3 dan semester 4 serta dikategorikan menjadi 3 seperti ditampilkan pada tabel 4.30.

Tabel 4.30 Kategori IP Semester

BESAR	IPS≥3,50
SEDANG	2,75<IPS<3,50
KECIL	IPS≤2,75

(2) Variabel Jumlah SKS yang telah ditempuh (JMLSKS)

Jenis data JMLSKS merupakan data real yang dikategorikan berdasarkan rata-rata jumlah SKS yang telah ditempuh oleh mahasiswa selama empat semester dari semester satu sampai semester empat seperti terlihat pada tabel 4.31.

Tabel 4.31 Kategori Jumlah SKS yang Telah Ditempuh

Kategori	Keterangan
KECIL	SKS<67
BESAR	SKS≥67

(3) Variabel Program Studi (PRODI)

Kategori Prodi dapat dilihat pada tabel 4.32.

Tabel 4.32 Kode Program Studi Skenario Kelima

Kode	Prodi	Kategori
55201	Teknik Informatika (S1)	TIC
56201	Sistem Komputer (S1)	SKC
57201	Sistem Informasi (S1)	SIC
57401	Manajemen Informatika (D3)	MIE
57402	Komputerisasi Akuntansi (D3)	KAE

(4) Variabel Jenjang (JNJNG)

Jenis data Jenjang dikategorikan seperti pada tabel 4.33.

Tabel 4.33 Kode Jenjang

Kode	Jenjang
C	Strata Satu (S1)
E	Diploma Tiga (DIII)

(5) Variabel Jenis Kelamin (JNSKLMN)

Jenis data dari variabel Jenis Kelamin dikategorikan menjadi Laki-laki dengan inisial L dan Perempuan dengan inisial P.

Data yang berjenis numerikal seperti tahun lulus harus dilakukan proses inisiasi data terlebih dahulu kedalam bentuk nominal. Untuk melakukan inisiasi tahun lulus dapat dilakukan dengan:

(a) Mahasiswa dari setiap angkatan yang sudah terdapat tahun kelulusan dinyatakan "lulus".

Mahasiswa dari setiap angkatan yang belum terdapat tahun kelulusan dinyatakan "tidak lulus".

Tabel 4.34 adalah *dataset* mahasiswa tahun 2005-2008 yang belum diinisiasi dan tabel 4.35 adalah *dataset* mahasiswa tahun 2005-2008 yang siap untuk perangkat pemodelan hasil pengolahan data pada skenario kelima.

Tabel 4.34 *Dataset* Mahasiswa tahun 2005-2008 yang belum diinisiasi

NIM	IPS1	IPS2	IPS3	IPS4	JUMLAHSKS	PRODI	JENJANG	JENISKELAMIN	KELULUSAN
2005 01 15 1 001	3,05	2,35	2,67	2,27	83	SIC	C	L	26-Agust-10
2005 01 15 1 002	3,15	2,30	3,09	2,88	90	SIC	C	L	17-Jul-09
2005 01 15 1 003	2,59	2,70	2,83	3,00	87	SIC	C	L	13-Jul-09
2005 01 15 1 005	3,45	2,43	2,87	3,13	90	SIC	C	P	14-Jul-09
2005 01 15 1 006	2,90	1,65	2,52	2,18	83	SIC	C	L	20-Agust-11
2005 01 15 1 008	2,95	2,00	2,24	2,00	79	SIC	C	L	20-Agust-11
2005 01 15 1 010	2,90	2,09	3,13	2,96	90	SIC	C	L	17-Jul-09
2005 01 15 1 012	3,00	2,48	3,17	2,88	90	SIC	C	L	14-Jul-09
2005 01 15 1 014	2,80	2,48	3,04	3,58	90	SIC	C	P	11-Jul-09
2005 01 15 1 015	3,10	3,09	3,22	2,92	90	SIC	C	L	11-Jul-09
2005 01 15 1 016	2,70	1,78	2,35	2,33	90	SIC	C	P	11-Jul-09
2005 01 15 3 017	2,85	0,02	2,00	0,07	46	SIC	C	L	
2005 01 15 3 018	2,90	1,47	2,68	0,00	64	SIC	C	P	
2005 01 15 3 019	3,20	3,14	3,04	1,00	70	SIC	C	P	
2005 01 15 7 020	3,70	3,48	3,26	3,50	90	SIC	C	P	15-Jul-09
2005 01 25 1 001	2,64	3,05	3,08	3,20	88	TIC	C	L	11-Jul-09
2005 01 25 1 002	3,09	2,50	3,17	3,17	92	TIC	C	L	11-Jul-09
2005 01 25 1 003	2,23	2,32	2,67	0,00	68	TIC	C	L	
2005 01 25 1 004	3,23	3,32	3,58	3,25	92	TIC	C	L	11-Jul-09
2005 01 25 1 005	0,17	0,00	0,00	0,00	18	TIC	C	L	
2005 01 25 1 006	2,18	2,27	0,50	0,00	52	TIC	C	L	
2005 01 25 1 007	3,64	3,09	3,21	2,50	92	TIC	C	L	15-Jul-09
2005 01 25 1 008	2,73	2,36	2,29	0,81	89	TIC	C	L	11-Jul-09
2005 01 25 1 009	3,00	2,91	2,83	3,38	92	TIC	C	L	11-Jul-09

Tabel 4.35 *Dataset* Mahasiswa tahun 2005-2008 yang Siap Perangkat Pemodelan

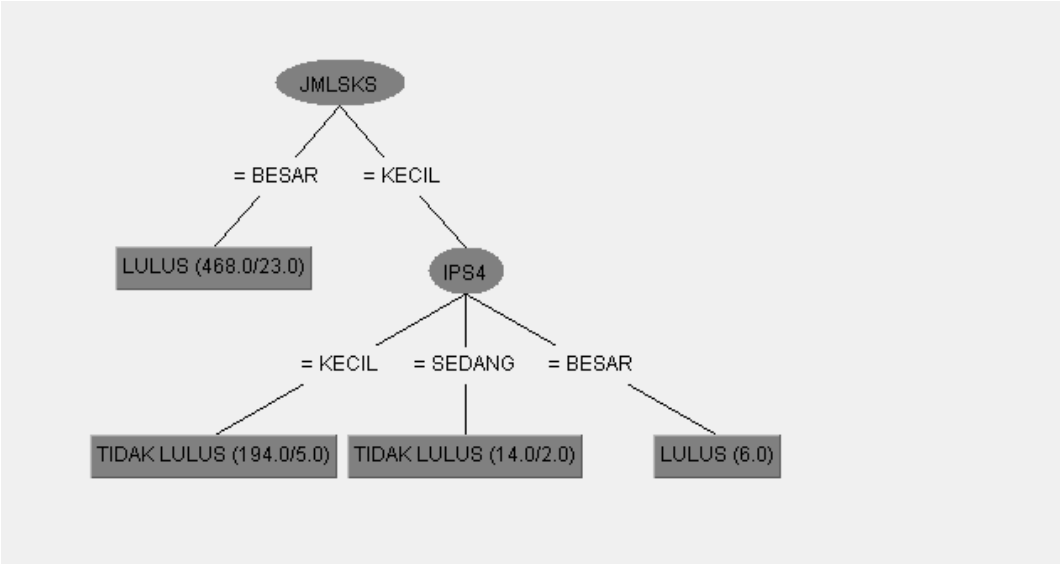
IPS1	IPS2	IPS3	IPS4	JMLSKS	PRODI	JNJNG	JNSKLMN	KELULUSAN
SEDANG	KECIL	KECIL	KECIL	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
KECIL	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	P	LULUS
SEDANG	KECIL	KECIL	KECIL	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	KECIL	KECIL	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
SEDANG	KECIL	SEDANG	BESAR	BESAR	SIC	C	P	LULUS
SEDANG	SEDANG	SEDANG	SEDANG	BESAR	SIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	BESAR	SIC	C	P	LULUS
SEDANG	KECIL	KECIL	KECIL	KECIL	SIC	C	P	TIDAK LULUS
BESAR	SEDANG	SEDANG	BESAR	BESAR	SIC	C	P	LULUS
KECIL	SEDANG	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
SEDANG	KECIL	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	BESAR	TIC	C	L	TIDAK LULUS
SEDANG	SEDANG	BESAR	SEDANG	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	KECIL	TIC	C	L	TIDAK LULUS
KECIL	KECIL	KECIL	KECIL	KECIL	TIC	C	L	TIDAK LULUS
BESAR	SEDANG	SEDANG	KECIL	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	BESAR	TIC	C	L	LULUS
SEDANG	SEDANG	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
SEDANG	SEDANG	SEDANG	SEDANG	BESAR	TIC	C	L	LULUS
KECIL	KECIL	KECIL	KECIL	KECIL	TIC	C	L	TIDAK LULUS

4.5.4 Fase Pemodelan *(Modeling Phase)*

Fase pemodelan pada skenario k elima sama dengan fase pemodelan pada skenario pertama, kedua, ketiga dan keempat dengan melakukan tiga pemodelan yaitu:

4.5.4.1 Algoritma C4.5

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2008 seperti pada tabel 4.31 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1 dengan algoritma C4.5. Gambar 4.10 adalah pohon keputusan yang terbentuk.



Gambar 4.10 Pohon Keputusan Yang dihasilkan Dari *Dataset* Mahasiswa tahun 2005-2008 pada skenario kelima

Dari sembilan variabel yang digunakan terlihat hanya dua variabel yang membentuk pohon, yaitu variabel JMLSKS (Jumlah SKS) dan IPS4 (IP Semester 4). Sedangkan variabel IPS1, IPS2, IPS3, Jenjang, Jenis Kelamin dan PRODI tidak terlihat dari pohon keputusan. Yang menjadi simpul akar adalah JMLSKS (Jumlah SKS) karena memiliki *gain* tertinggi. Jika JMLSKS BESAR maka "lulus" sedangkan jika JMLSKS KECIL maka lihat IPS4 (IP Semester 4). Jika IPS4 KECIL dan SEDANG maka "tidak lulus" sedangkan jika IPS4 BESAR maka "lulus".

Confusion Matrix untuk algoritma C4.5 dapat dilihat pada tabel 4.36. *Confusion Matrix* menunjukkan ketepatan klasifikasi atau kesesuaian dengan prediksi yang dilakukan dengan metode C4.5.

Tabel 4.36 *Confusion Matrix* Algoritma C4.5 Skenario Kelima

```
=== Confusion Matrix ===

   a   b   <-- classified as
 451   7 |   a = LULUS
  23 201 |   b = TIDAK LULUS
```

Akurasi = $\frac{451+201}{451+7+23+201}$ x 100% = 95.6012 %

Nilai kurva *ROC* yang dihasilkan adalah 0,923.

Keterangan tabel 4.36 adalah:

(a) Jumlah data *real* yang LULUS dan diprediksi LULUS adalah 451.

(b) Jumlah data *real* yang TIDAK LULUS dan diprediksi TIDAK LULUS adalah 201.

(c) Jumlah data *real* yang TIDAK LULUS dan diprediksi LULUS adalah 23

(d) Jumlah data *real* yang LULUS dan diprediksi TIDAK LULUS adalah 7.

4.5.4.2 Algoritma *Naïve Bayes*

Berikut ini adalah hasil penelitian dengan menggunakan *dataset* mahasiswa tahun 2005-2008 seperti pada tabel 4.31 yang selanjutnya akan diolah dengan alat bantu *data mining* WEKA 3.6.1 dengan algoritma *Naïve Bayes*. Hasil yang didapat sudah memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.37 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *Naïve Bayes* pada skenario kelima.

Tabel 4.37 *Confusion Matrix* Algoritma *Naïve Bayes* Skenario Kelima

```
=== Confusion Matrix ===

   a   b   <-- classified as
 407  51 |   a = LULUS
  20 204 |   b = TIDAK LULUS
```

Akurasi = $\frac{407+204}{407+204+51+20}$ x 100% = 89,5894 %

Nilai kurva *ROC* yang dihasilkan adalah 0,97.

4.5.4.3 Algoritma *CART*

Penelitian selanjutnya adalah dengan menggunakan *dataset* mahasiswa tahun 2005-2008 seperti pada tabel 4.31 dengan algoritma *CART* dan diolah dengan alat bantu *data mining* WEKA 3.6.1. Hasil yang didapat sudah memberikan informasi atau pengetahuan yang sesuai dengan tujuan proyek. Tabel 4.38 adalah *Confusion Matrix* yang terbentuk dari pemodelan dengan algoritma *CART* pada skenario kelima.

Tabel 4.38 *Confusion Matrix* Algoritma *CART* Skenario Kelima

```
=== Confusion Matrix ===

   a   b   <-- classified as
 451   7 |   a = LULUS
  23 201 |   b = TIDAK LULUS
```

Akurasi = $\frac{451+201}{451+201+7+23} x\ 100\%$ = 95.6012%

Nilai kurva *ROC* yang dihasilkan adalah 0,922.

4.5.5 Fase Evaluasi *(Evaluation Phase)*

Pebandingan hasil perhitungan nilai kurca *ROC* (*AUC)* untuk algoritma C4.5, *Naïve Bayes*, dan *CART* dapat dilihat pada Tabel 4.39.

Tabel 4.39 Komparasi Nilai AUC

Algoritma	Nilai AUC
C4.5	0.923
Naïve Bayes	0.97
CART	0.922

Analisis Hasil Komparasi

Model yang dihasilkan algoritma C4.5, *Naïve Bayes*, dan *CART* diuji menggunakan metode *Cross Validation*, terlihat algoritma C4.5 dan algoritma *CART* memiliki nilai *accuracy* yang sama dan paling tinggi, sedangkan yang terendah adalah *Naïve Bayes.*

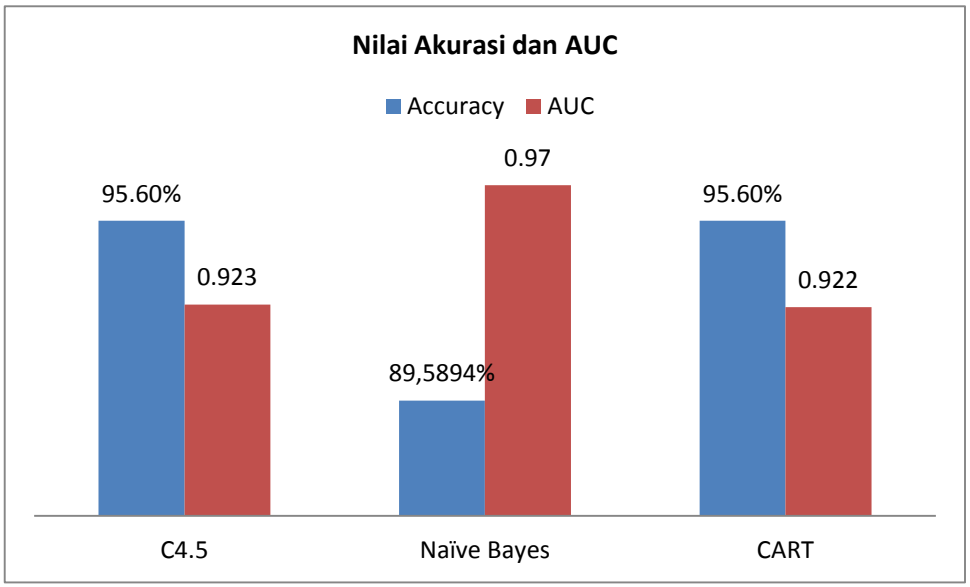
Dalam hal ini, dengan menggunakan algoritma C4.5 dan *CART* kesalahan dalam proses prediksi lebih sedikit karena kedua algoritma ini melakukan klasifikasi record-record kedalam kelas tujuan yang ada. Untuk implementasi algoritma *Naïve Bayes* dengan menggunakan data dalam proses training akan menghasilkan nilai kesalahan yang lebih besar karena pada algoritma ini nilai suatu variabel adalah independen terhadap nilai lainnya dalam satu variabel yang sama namun memiliki akurasi yang lebih tinggi bila diimplementasikan ke data yang berbeda dari data training dan kedalam data yang jumlahnya lebih besar.

Tabel 4.40 Komparasi Nilai *Accuracy* dan AUC

	C4.5	*Naïve Bayes*	*CART*
Accuracy	95,6012%	89,5894%	95,6012%
AUC	0,923	0,97	0,922

Tabel 4.40 membandingkan A*ccuracy* dan AUC dari tiap algoritma. Terlihat bahwa nilai *Accuracy* algoritma C4.5 dan algoritma *CART* memiliki nilai yang sama yaitu 95,6012%. Nilai AUC paling tinggi adalah *Naïve Bayes* yaitu 0,97. Hal ini dikarenakan nilai *false positive* pada algoritma ini lebih kecil daripada nilai *false positive* pada algoritma C4.5 dan *CART*.

Berdasarkan pengelompokkan Tabel 4.40 maka dapat disimpukan bahwa algoritma C4.5, *Naïve Bayes*, dan *CART* termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00. Grafik nilai akurasi dan nilai AUC masing-masing algoritma terlihat pada gambar 4.11.



Gambar 4.11 Nilai Akurasi dan AUC masing-masing Algoritma

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Hasil penelitian yang diperoleh sudah sesuai dengan tujuan penelitian yaitu sebagai berikut:

a. Dapat membandingkan tingkat akurasi yang dihasilkan masing-masing algoritma. Dengan alat bantu WEKA prediksi tingkat kelulusan mahasiswa STMIK Rosma Karawang, Algoritma C4.5 menghasilkan akurasi 95,6012%, Naïve Bayes 89,5894% dan CART 95,6012%.
b. Algoritma C4.5, *Naïve Bayes* dan *CART* dapat digunakan untuk memprediksi kelulusan mahasiswa. Untuk mengukur ketiga fungsi algoritma tersebut digunakan *confusion matrix* dan kurva ROC dengan hasil bahwa algoritma yang memiliki tingkat akurasi yang paling tinggi adalah algoritma C4.5 dan algoritma *CART.* Sedangkan algoritma yang menghasilkan kurva ROC paling tinggi adalah algoritma *Naïve Bayes.* Akurasi algoritma C4.5 dan algoritma *CART* memberikan hasil yang sama yaitu 95,6012%. Hal ini terjadi karena algoritma C4.5 membangun pohon dengan jumlah pohon tiap simpul sesuai dengan jumlah nilai simpul tersebut, seperti kasus data yang peneliti lakukan terhadap data nilai mahasiswa yang dikelompokkan kedalam dua kelompok (lulus dan tidak lulus) sehingga akan sama dengan algoritma *CART* dengan konsep cabang pohon biner.
c. Algoritma C4.5 dan algoritma *CART* memberikan akurasi yang lebih baik daripada *Naïve Bayes* dalam klasifikasi data mahasiswa STMIK Rosma Karawang.
d. Algoritma C4.5 dan algoritma *CART* memberikan hasil lebih baik karena data mahasiswa STMIK Rosma merupakan data kelompok yang cocok dengan sifat klasifikasi algoritma C4.5 dan algoritma *CART.*
e. Kelulusan mahasiswa dapat diprediksi lebih dini yaitu pada semester 4.
f. *Data mining* dengan algoritma C4.5 dan *CART* dapat diimplementasikan untuk memprediksi kelulusan mahasiswa STMIK Rosma dengan dua kategori yaitu lulus

dan tidak lulus. Variabel yang berpengaruh dalam hasil prediksi adalah Jumlah SKS yang telah ditempuh (JMLSKS) dan Indek Prestasi Semester 4 (IPS4).

g. Dapat menjabarkan masing-masing algoritma kedalam *rule*.
h. Dapat menerapkan masing-masing algoritma dalam melakukan prediksi terhadap kelulusan mahasiswa STMIK Rosma Karawang.

5.2 Saran

Untuk penelitian selanjutnya dapat menambah variabel lain, selain dari variabel yang dilakukan peneliti serta menggunakan perpaduan algoritma lainnya seperti *k-nearest neighbor* dan *neural network.*

DAFTAR PUSTAKA

Andi Wahju Rahardjo, Emanuel dan Hartono Arie. 2008. *Pengembangan Aplikasi Pengenalan Karaketer Alfanumerik Dengan Menggunakan Algoritma Neural Network Three-Layer Backpropagation.* Jurnal Informatika. Vol.4, No.1. 49 – 58.

Ariawan, Iwan. 2009. *Catatan materi kuliah dr Iwan Ariawan, MS.* [Online]. http://www.scribd.com/doc/15123416/Kurva-Receiver-Operating-Characteristic. Diakses tanggal 19 Oktober 2015, 20:15 WIB

Al-Radaideh, Q.A. 2006. *Mining Student Data Using Decission Tree. International Arab Conference on Informational Technology* (ACIT).

Basuki, Ahmad dan Syarif, Iwan. 2003. *Decision Tree*. Surabaya : Politeknik Electronika Negeri Surabaya ITS.

Bramer, M. 2007. *Principles of Data Mining.* London: Springer

Gorunescu, Florin. 2011. *Data Mining Concept Model Technique.*

Han, J, dan Kamber, M. 2001. *Data Mining Concepts and Techniques.* Morgan Kaufman Pub. USA

Huda, Nuqson Masykur. 2010. *Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa.* Semarang

Jananto A. 2010. *Penggunaan Algoritma SLIQ untuk Pengklasifikasian Kinerja Akademik Mahasiswa.* Jurnal Teknologi Informasi DINAMIK Vol XV, No.1 : 66-72

Jefri. 2013. *Implementasi Algoritma C4.5 Dalam Aplikasi Untuk Memprediksi Jumlah Mahasiswa Yang Mengulang Mata Kuliah Di STMIK AMIKOM Yogyakarta,* Yogyakarta

Kamagi, David Hartanto dan Hansun Seng. 2014. *Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa.* ULTIMATICS Vol. VI, No. 1

Karypis, George, dkk. 2007. *CHAMELEON : A Hierarchical Clustering Algorithm Using Dynamic Modeling*. Diambil dari http://www.users.cs.umn.edu/~kumar/papers/chameleon.ps. Diakses tanggal 19 Oktober 2015, 20:20 WIB.

Kursini, Luthfi. E. T. 2009. *Algoritma Data Mining.* PT Andi Offset.

Larose, Daniel. T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Willey & Sons. Inc.

M, Mohammed, dkk. 2012. *Mining Educational Data to Improve Students' Performance: A Case Study.* International Journal of Information and Communication Technology Research Volume 2 No. 2

Maimon, Rockah. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer Heidelberg. Berlin

Moertini, Veronica Sri. 2007. *Pengembangan Skalabilitas Algoritma Klasifikasi C4.5 Dengan Pendekatan Konsep Operator Relasi, studi kasus: pra-pengolahan dan klasifikasi citra batik.* Bandung

Nilakant, K. 2004. *Application of Data Mining in Constraint Based Intelligent Tutoring System.*www.cosc.canterbury.ac.nz/research/reports/HonsReps/2004/hons_0408.pdf. diakses tanggal 12 Oktober 2015

Nugroho, Yuda Septian. 2014. *Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro*. Fasilkom UDINUS Semarang

Nugroho, Yusuf Sulistyo. 2014. *Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta*. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Yogyakarta

Oscar Ong, Johan. 2013. *Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University.* Jurnal Ilmiah Teknik Industri. vol. 12. No. 1. pp. 10-13.

Prasetyo, Eko. 2012. *Data Mining : Konsep dan Aplikasi menggunakan MATLAB, 1st ed.* PT Andi Offset

Rahmayuni, Indri. 2014. *Klasifikasi Data Karakteristik Mahasiswa Menggunakan Algoritma C4.5 Dan Cart (Studi Kasus Educational Data Mining)*. Jurnal Teknologi Informasi & Pendidikan ISSN : 2086 – 4981. Vol. 7 No. 1

Ridwan, Mujib, dkk. 2013. *Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier.* Malang

STMIK ROSMA Karawang. 2010. *Buku Panduan Akademik Mahasiswa Tahun Ajaran 2010-2011*. Karawang. Jawa Barat

Susanto, Sani, dkk. 2010. *Pengantar Data Mining Menggali Pengetahuan Dari Bongkahan Data*. ANDI Yogyakarta.

Suhartinah, Marselina Silvia dan Ernastuti. 2010. *Graduation Prediction Of Gunadarma University Students Using Algorithm And Naive Bayes C4.5 Algorithm.* Undergraduate Program, Faculty of Industrial Engineering, Gunadarma University

Sunjana. 2010. *Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi Decision Tree.* Seminar Nasional Aplikasi Teknologi Informasi.

Swastina, Liliana. 2013. *Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa.* Jurnal GEMA AKTUALITA, Vol. 2 No. 1

Tan S, Kumar P, Steinbach M. 2005. *Introduction to Data Mining.* Addison Wesley. J. Taylor, Ed. Stanford

Turban, E, dkk. 2005. *Decicion Support Systems and Intelligent Systems*. PT Andi Offset

Witten, Ian H, dkk. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA

Wu & Kumar. 2009. *The Top Ten Algorithms in Data Mining.* USA: CRC Press

Yadav, Surjeet Kumar, dkk. 2012. *Data Mining Applications: A comparative Study for Predicting Student's performance.* International Journal Of Innovative Technology & Creative Engineering (ISSN:2045-711). Vol.1 No.12

http://www.cs.waikato.ac.nz/ml/weka/ diakses tanggal 19 Oktober 2015, 19:59 WIB