Penerapan *Data Mining* menggunakan Metode *Decision Tree C4.5* untuk Prediksi Tingkat Kelulusan Mahasiswa (Studi Kasus : STMIK WIT)

Niken Adityas Lestari

Teknik Informatika, STMIK WIT

Abstract

One aspect of accreditation assessment is students and graduates. Graduation rates and the number of students will influence the accreditation process carried out by the government. The university database stores a variety of data, from those data when properly explored, it can be seen that new patterns or knowledge that can be developed to be applied to universities include the potential for students to graduate on time, late graduation, and drop out. This study uses Knowledge Discovery in Database (KDD) methods to analyze data in the application of data mining, extract what knowledge is considered according to size and boundary specifications, use databases together with the required preprocessing, sampling and transformation from the database. Data mining techniques can predict graduation on time, late and drop out, that is using the decision tree or decision tree with C4.5 algorithm based on GPA 4 attributes, employment status, study program, school origin, gender, parental salary and regional origin. The C4.5 algorithm can construct decision trees that are able to overcome continuous type attributes, overcome missing values and can prune complex trees. The data used in the prediction process are student data from 2006 to 2010, and the data used to test the prediction pattern is 2012 student data.

Keywords : *Data mining, Knowledge discovery in database (KDD), decision tree, C4.5 algorithm.*

Abstrak

Salah satu aspek penilaian akreditasi adalah mahasiswa dan lulusan. Tingkat kelulusan dan jumlah mahasiswa akan berpengaruh dalam proses akreditasi yang dilakukan oleh pemerintah. Database perguruan tinggi menyimpan berbagai data, dari data tersebut apabila digali dengan tepat maka dapat diketahui pola atau pengetahuan baru yang dapat dikembangkan untuk diterapkan pada perguruan tinggi diantaranya mengenai potensi mahasiswa lulus tepat waktu, lulus terlambat, dan drop out. Penelitian ini menggunakan metode *Knowledge Discovery in Database* (KDD) untuk menganalisis data dalam penerapan data mining, mengekstrak pengetahuan apa yang dianggap sesuai dengan spesifikasi ukuran dan batas, menggunakan database bersama dengan *preprocessing* yang diperlukan, pengambilan sampel dan transformasi dari database. Teknik data mining dapat melakukan prediksi kelulusan tepat waktu, terlambat dan *drop out*, yaitu menggunakan *decision tree* atau pohon keputusan dengan algoritma C4.5 berdasarkan atribut IPK4, status pekerjaan, program studi, asal sekolah, jenis kelamin, gaji orangtua dan asal daerah. Algoritma C4.5 dapat mengkonstruksi pohon keputusan yang mampu mengatasi atribut bertipe kontinu, mengatasi nilai yang hilang dan dapat melakukan pemangkasan pohon yang kompleks. Data yang digunakan dalam proses prediksi adalah data mahasiswa tahun 2006 sampai dengan 2010, dan data yang digunakan untuk menguji pola prediksi adalah data mahasiswa tahun 2012.

Kata Kunci : Data mining, *Knowledge discovery in database* (KDD), *decision tree*, algoritma C4.5.

1. Pendahuluan

Saat ini dikenal istilah *big data*, yang menggambarkan volume data yang sangat besar, terstruktur maupun tidak terstruktur, yang membanjiri dunia bisnis. *Big data* dapat dianalisis, sehingga organisasi atau perusahaan dapat mengambil keputusan strategis dengan lebih baik. Dalam *big data*, tentu saja akan kesulitan dalam membaca dan mengetahui pola-pola dan relasi-relasi data jika dilakukan secara manual. Untuk itu dibutuhkan suatu teknik yang relatif cepat dan mudah untuk menemukan pengetahuan, pola dan relasi antar data secara otomatis yang disebut dengan Data Mining atau penambangan data.

"Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse atau penyimpanan lainnya". (Meilani dan Susanti, 2014: 1-2)

Metode *decision tree* mengubah fakta yang sangat besar menjadi pohon keputusan yang memprediksikan aturan. Salah satu algoritma yang digunakan untuk membentuk pohon keputusan (*decision tree)* adalah C4.5.

Algoritma data mining C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan dan bersifat prediktif. Dasar algoritma C4.5 adalah pembentukan pohon keputusan (*decision tree*).

Setiap perguruan tinggi berusaha untuk terus memperbaiki manajemennya untuk meningkatkan mutu pendidikan dan meningkatkan akreditasi. Salah satu aspek penilaian akreditasi adalah mahasiswa dan lulusan. *Database* perguruan tinggi menyimpan data akademik, data administrasi, data mahasiswa dan lulusan. Dari data tersebut apabila digali dengan tepat maka dapat diketahui pola atau pengetahuan baru yang dapat dikembangkan untuk diterapkan pada perguruan tinggi diantaranya mengenai potensi mahasiswa lulus tepat waktu, lulus terlambat, dan *drop out*.

Penelitian ini menggunakan data mining untuk prediksi kelulusan mahasiswa dengan menggunakan tiga variabel target yaitu lulus tepat waktu, lulus terlambat dan *drop out*. Metode yang digunakan adalah *decision tree* C4.5. Dengan adanya penelitian ini, pihak perguruan tinggi dapat

nikenadityas54@gmail.com

mengetahui hasil prediksi kelulusan mahasiswa dan dapat memberikan tindakan preventif untuk mengurangi masalah keterlambatan kelulusan dan *drop out*.

2. Kerangka Teori

2.1. Data Mining

Istilah data mining memiliki beberapa padanan, seperti *knowledge discovery* ataupun penemuan pengetahuan tepat digunakan karena tujuan utama dari data mining memang untuk mendapatkan pengetahuan yang masih tersembunyi di dalam bongkahan data. (Susanto dan Suryadi, 2010:2).

Data mining bukanlah suatu bidang yang baru. Dalam aplikasinya, *data mining* sebenarnya merupakan bagian dari proses *Knowledge Discovery in Database* (KDD), bukan sebagai teknologi yang utuh dan berdiri sendiri. Menurut (Al Fatta, 2007:13) menjelaskan bahwa Data mining adalah proses yang menggunakan teknik statistik, perhitungan, kecerdasan buatan dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar.

Secara sederhana, data mining dapat diartikan sebagai proses mengekstrak atau "menggali" pengetahuan yang ada pada sekumpulan data. Menurut (Brijesh Kumar Bhardwaj, 2010;1) menjelaskan bahwa *Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying,extracting and evaluating variables related to the learning process of students as described.*

Banyak orang yang setuju bahwa data mining adalah sinonim dari *Knowledge-Discovery in Database* atau yang biasa disebut KDD. Dari sudut pandang yang lain, data mining dianggap sebagai satu langkah yang penting didalam proses KDD

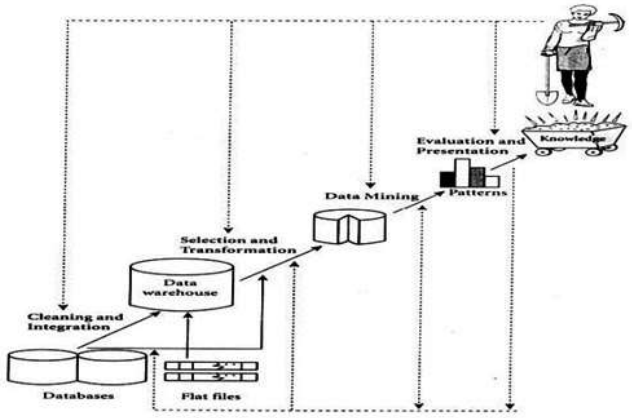
"Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya". (Meilani dan Susanti, 2014:1-2)

Menurut (Kadir, 2013:240) menjelaskan bahwa Data mining atau penambangan data adalah perangkat lunak yang digunakan untuk menemukan pola-pola tersembunyi, tren, maupun aturan-aturan yang terdapat dalam basis berukuran besar dan mengahsilkan aturan-aturan yang digunakan utnuk memperkirakan perilaku di masa mendatang.

"Kegunaan data mining dapat dibagi menjadi dua: deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari pola-pola yang dapat dipahami manusia dan menjelaskan karakteristik data. Sedangkan prediktif berarti data mining digunakan untuk mebentuk sebuah model pengetahuan yang akan digunakan untuk dilakukan prediksi". (Suyanto, 2017:3)

2.2 Metode *Knowledge Discovery in Database* (KDD)

Metode untuk menganalisis data dalam penerapan data mining ini adalah dengan menggunakan tahapan *Knowledge Discovery in Database* (KDD). Proses KDD adalah proses menggunakan metode *data mining* untuk mengekstrak pengetahuan apa yang dianggap sesuai dengan spesifikasi ukuran dan batas, menggunakan database bersama dengan *preprocessing* yang diperlukan, pengambilan sampel dan transformasi dari database. Tahapan *Knowledge Discovery in Database* (KDD) terdiri dari beberapa tahapan. Berikut adalah tahapan *Knowledge Discovery in Database* (KDD) yang ditunjukkan pada Gambar 2.1.



Gambar 2.1. Tahap penemuan *Knowledge* pada Data Mining (KDD) (Han & Kamber, 2006:6)

2.3 Decision Tree

Menurut (kusrini dan emha, 2009:13) menjelaskan bahwa Pohon keputusan atau *decision tree* merupakan metode klasifikasi dan prediksi yang sangat kuat dan tekenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat di ekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu.

2.4 Algoritma C4.5

Salah satu algoritma yang dapat digunakan untuk membuat pohon keputusan (*decision tree)* adalah algoritma C4.5.

C4.5 is a well-known algorithm used to generate a decision trees. It is an extension of the ID3 algorithm used to overcome its disadvantages. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classifier. (Kalpesh Adhatrao, 2013;42)

Menurut (kusrini dan emha, 2009:15) menjelaskan bahwa : Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

1. Pilihlah atribut sebagai akar
2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang

Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

2.5 *Waikato Environment Knowledge and Analisys* (WEKA)

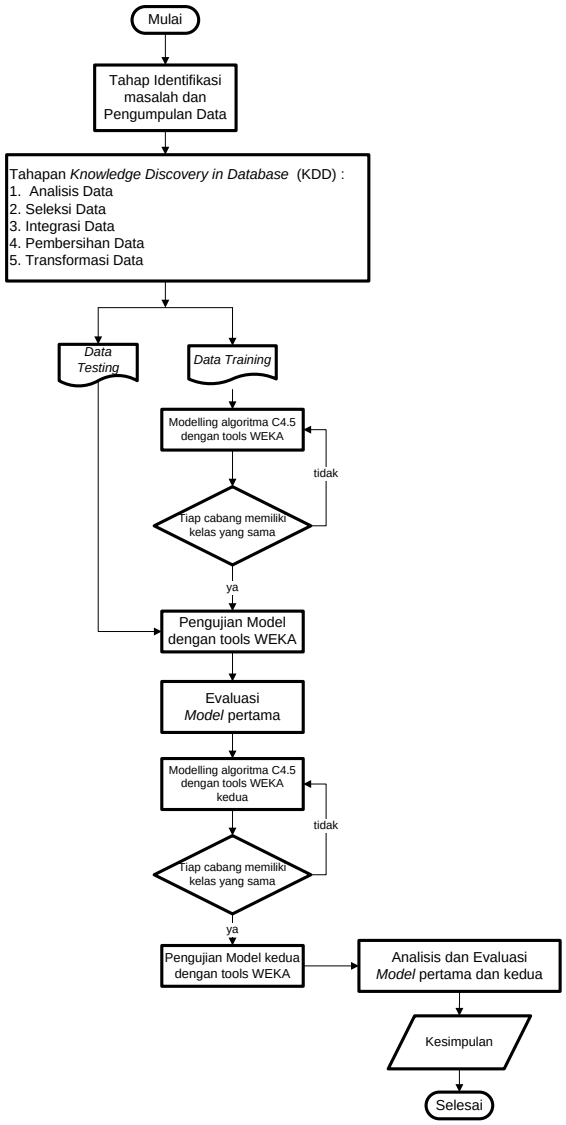
WEKA adalah sebuah perangkat lunak *data mining*/*machine learning* yang dibangun oleh *Department of Computer Science University of Waikato* di New Zealand.

WEKA mudah digunakan dan diterapkan pada beberapa tingkatan yang berbeda. Tersedia implementasi algoritma-algoritma pembelajaran *stateof-the-art* yang dapat diterapkan pada dataset dari *command line*. WEKA mengandung *tools* untuk *pre-processing* data, klasifikasi, regresi, *clustering*, aturan asosiasi, dan visualisasi. *User* dapat melakukan *preprocess* pada data, memasukkannya dalam sebuah skema pembelajaran, dan menganalisa

classifier yang dihasilkan dan performansinya – semua itu tanpa menulis kode program sama sekali. Contoh penggunaan WEKA adalah dengan menerapkan sebuah metode pembelajaran ke dataset dan menganalisa hasilnya untuk memperoleh informasi tentang data, atau menerapkan beberapa metode dan membandingkan performansinya untuk dipilih.(Putu Gede Surya Cipta Nugraha, 2016;37)

3. Metodologi

Metode penelitian dapat dilihat pada Gambar 3.1 berikut ini :



Gambar 3.1. Metode Penelitian

Langkah – langkah penelitian dapat dijelaskan sebagai berikut :

1. Identifikasi Masalah

Pada Tahapan ini dilakukan analisa kebutuhan bisnis, analisa ini berfokus pada pemahaman tujuan kebutuhan berdasarkan penilaian bisnis. Kemudian pemahaman tersebut diubah menjadi sebuah rencana awal data mining yang dirancang untuk mencapai tujuan. Tujuan bisnis dari penelitian ini adalah mengenali pola mahasiswa untuk predikasi kelulusan.

2. Pengumpulan Data

Pengumpulan data penelitian ini dilakukan dengan cara sebagai berikut:

a. *Observasi*
   Merupakan metode pengumpulan data dengan cara mengadakan pengamatan langsung kepada objek penelitian.
b. Wawancara
   Merupakan teknik pengumpulan data dengan cara mengadakan tanya jawab atau wawancara langsung kepada bagian administrasi dan akademik STMIK WIT.
c. Studi Pustaka
   Studi pustaka, mengumpulkan data dengan mempelajari masalah yang berhubungan dengan objek yang diteliti serta bersumber dari buku- buku, literatur yang disusun oleh para ahli untu melengkapi data yang diperlukan dalam penelitian.

3. Analisa Data

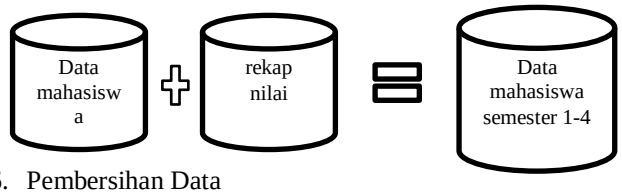
Pada tahapan ini dilakukan analisis untuk memahami data yang didapatkan dari hasil pengumpulan data. Data yang akan dianalisa ini adalah data mahasiswa STMIK WIT jenjang D3 dan S1 pada tahun 2006 sampai dengan tahun 2010. Kemudian dilakukan eksplorasi dan analisa struktur tabel yang ada pada database STMIK WIT. Output dari eksplorasi data ini adalah : tabel mahasiswa, tabel dosen, tabel mata kuliah, tabel FRS, rekap nilai dan tabel kelulusan.

4. Seleksi Data

Pada tahap ini tabel yang berhubungan dipilih untuk mempermudah proses pemilihan data. Daftar tabel yang telah di eksplorasi adalah : tabel dosen, tabel mahasiswa, tabel mata kuliah, tabel FRS dan tabel kelulusan. Dari data tersebut, dipilih tabel yang berhubungan dengan penelitian ini, yaitu data mahasiswa, nilai dan kelulusan.

5. Integrasi Data

Proses ini adalah menggabungkan beberapa atribut dari data mahasiswa, rekap nilai dan kelulusan menjadi satu tabel.



6. Pembersihan Data

Pada tahapan ini akan dilakukan pembersihan data dengan membuang data yang tidak konsisten atau *noise*, duplikasi data, memperbaiki kesalahan data dan bisa diperkaya dengan data eksternal yang relevan.

Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap dan bernilai null, sehingga jumlah dataset yang didapatkan dari proses pembersihan data ini adalah 977 *record*, seperti terlihat pada Tabel 3.1.

Tabel 3.1. *Dataset* jumlah mahasiswa STMIK WIT Angkatan 2006-2010

No	Program studi	Jenjang	Jumlah
1	Komputerisasi Akuntansi	D3	322
2	Manajemen Informatika	D3	296
3	Sistem Informasi	S1	142
4	Teknik Informatika	S1	217
	Total		977

7. Transformasi Data

Pada tahap ini, akan dilakukan proses pengubahan data ke dalam format yang sesuai untuk diproses dalam data mining.

a. Atribut Program Studi

Dikategorikan menjadi empat macam yaitu komputerisasi akuntansi, manajemen infomatika, Teknik informatika dan Sistem Informasi

b. Atribut jenis kelamin

Dikategorikan menjadi dua yaitu P untuk jenis kelami perempuan dan L untuk jenis kelamin laki-laki

c. Atribut asal daerah

Dikategorikan menjadi dua yaitu Cirebon dan luar Cirebon

d. Atribut asal sekolah

Dikategorikan berdasarkan asal sekolah mahasiswa yaitu SMA, SMK, MAN dan PAKET C

e. Atribut Indeks Prestasi Kumulatif (IPK)

Jenis data IPK diambil dari IPK semester 4. atribut dari IPK dikategorikan menjadi 3, seperti ditampilkan dalam Tabel 3.2.

Tabel 3.2 Kategori Indeks Prestasi Semester (IPS)

Kategori	Keterangan
Besar	IPS ≥ 3,00
Sedang	2,00 ≤ IPS ≤ 2,99
Kecil	IPS < 2,00

f. Variabel status pekerjaan mahasiswa

Jenis data pada variabel status pekerjaan mahasiswa dikategorikan menjadi dua, seperti yang ditampilkan dalam Tabel 3.3

Tabel 3.3 kategori status pekerjaan

Kategori	Keterangan
Bekerja	Bekerja
Mahasiswa	Tidak Bekerja

g. Atribut Gaji ortu

Atribut dari gaji orang tua dikategorikan menjadi tiga, terlihat pada Tabel 3.4

Tabel 3.4 Kategori Gaji Ortu

Kategori	Keterangan
Tinggi	>Rp. 3000.000
Sedang	Rp. 1.600.000 s.d Rp. 3000.000
Rendah	≤Rp. 1500.000

h. Atribut Kelas

Atribut ini dikategorikan menjadi dua yaitu pagi dan sore

Tabel 3.5 Cuplikan Dataset mahasiswa 2006-2010 yang belum ditransformasi

NIM	Nama Mahasiswa	PRODI	JK	Asal Daerah	Asal Sekolah	IPK 4	Gaji Ortu	Status Pekerjaan	KELAS	Status Kelulusan
0607048011	Alfani Junior	Komputerisasi Akuntansi	P	Cirebon	SMA	3,21	Rp 2.900.000	MAHASISWA	SORE	LULUS TEPAT WAKTU
0607048012	Bobby Wahyu A	Komputerisasi Akuntansi	L	Cirebon	MAN	1,11	Rp 2.900.000	BEKERJA	SORE	LULUS TERLAMBAT
0607048013	Delia Hindiarni	Komputerisasi Akuntansi	P	Luar Cirebon	SMA	3,00	Rp 1.900.000	BEKERJA	PAGI	LULUS TERLAMBAT
0607048014	Dwi Lukluk Dorlar	Komputerisasi Akuntansi	P	Luar Cirebon	SMA	3,21	Rp 3.000.000	MAHASISWA	SORE	LULUS TEPAT WAKTU
0607048015	Heni Yulianti	Komputerisasi Akuntansi	P	Luar Cirebon	SMA	2,42	Rp 2.900.000	BEKERJA	PAGI	LULUS TERLAMBAT
0607048016	Irman Antoni	Komputerisasi Akuntansi	L	Luar Cirebon	SMK	2,35	Rp 2.900.000	BEKERJA	PAGI	LULUS TERLAMBAT
0607048017	Iryan Setiawan	Komputerisasi Akuntansi	P	Luar Cirebon	PAKET C	2,63	Rp. 150.000	BEKERJA	PAGI	LULUS TERLAMBAT
0607048018	Lusi Juliarsih	Komputerisasi Akuntansi	P	Luar Cirebon	SMA	2,32	Rp 1.900.000	BEKERJA	PAGI	DROP OUT
0607048019	Lusiati	Komputerisasi Akuntansi	P	Cirebon	SMA	3,37	Rp 2.900.000	BEKERJA	PAGI	DROP OUT
0607048110	Andhin Nurcahi	Komputerisasi Akuntansi	P	Luar Cirebon	SMA	0,08	Rp 2.900.000	BEKERJA	PAGI	DROP OUT
0607048111	Lilis Nurhayati	Komputerisasi Akuntansi	P	Cirebon	SMA	3,53	Rp 1.300.000	BEKERJA	PAGI	DROP OUT
0607048112	Nopianalin	Komputerisasi Akuntansi	L	Luar Cirebon	SMA	3,26	Rp. 800.000	BEKERJA	PAGI	DROP OUT
0607048113	Puspa Indah	Komputerisasi Akuntansi	P	Luar Cirebon	SMK	3,00	Rp 2.900.000	MAHASISWA	PAGI	LULUS TEPAT WAKTU
0607048114	Ronald Sumarna	Komputerisasi Akuntansi	L	Luar Cirebon	SMA	2,00	Rp 1.900.000	MAHASISWA	SORE	LULUS TERLAMBAT
0607048115	Sri Mulyati	Komputerisasi Akuntansi	P	Luar Cirebon	SMK	3,63	Rp 1.900.000	BEKERJA	SORE	LULUS TERLAMBAT
0607048116	Ade Rusmin	Komputerisasi Akuntansi	L	Luar Cirebon	SMK	1,94	Rp 1.900.000	BEKERJA	SORE	DROP OUT
0607048117	Dwi Putri Marwar Ceni	Komputerisasi Akuntansi	P	Luar Cirebon	SMK	2,19	Rp 1.900.000	BEKERJA	SORE	LULUS TERLAMBAT
0607048118	Eni Rayanti	Komputerisasi Akuntansi	P	Luar Cirebon	SMA	1,47	Rp2.500.000	BEKERJA	PAGI	DROP OUT
0607048119	Hidayat	Komputerisasi Akuntansi	L	Luar Cirebon	SMK	1,37	Rp2.500.000	BEKERJA	PAGI	DROP OUT
0607048120	Novian	Komputerisasi Akuntansi	P	Cirebon	MAN	2,20	Rp 3.900.000	BEKERJA	PAGI	LULUS TERLAMBAT

Tabel 3.6 Seleksi Atribut

Atribut	Detail Penggunaan	
NIM	x	ID
Nama	x	No
Program studi	✓	Nilai Model
Jenis Kelamin	✓	Nilai Model
Asal Daerah	✓	Nilai Model
Asal Sekolah	✓	Nilai Model
IPK4	✓	Nilai Model
Status Pekerjaan	✓	Nilai Model
Gaji Ortu	✓	Nilai Model
Kelas	✓	Nilai Model
Status	✓	Label Target

Tabel 3.6 menjelaskan atribut-atribut yang akan digunakan dalam penelitian, indikator "yes" (√) menandakan bahwa atribut bersangkutan akan digunakan dalam penelitian, sedangkan indikator "No" (x) menandakan atribut tersebut akan dieliminasi.

Tabel 3.7 Cuplikan *Dataset* mahasiswa 2006-2010 yang telah ditransformasi

Program Studi	JK	Asal Daerah	ASAL SEKOLAH	IPK4	GAJI ORTU	STATUS PEKERJAAN	KELAS	STATUS
Komputerisasi Akuntansi	P	Cirebon	SMA	Besar	Sedang	Tidak Bekerja	SORE	Lulus Tepat Waktu
Komputerisasi Akuntansi	L	Cirebon	MAN	Kecil	Sedang	Bekerja	SORE	Lulus Terlambat
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Besar	Rendah	Bekerja	PAGI	Lulus Terlambat
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Besar	Sedang	Tidak Bekerja	SORE	Lulus Tepat Waktu
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Sedang	Tinggi	Bekerja	PAGI	Lulus Terlambat
Komputerisasi Akuntansi	L	Luar Cirebon	SMK	Sedang	Sedang	Bekerja	PAGI	Lulus Terlambat
Komputerisasi Akuntansi	P	Luar Cirebon	PAKET C	Sedang	Sedang	Bekerja	PAGI	Lulus Terlambat
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Sedang	Sedang	Bekerja	PAGI	Drop Out

8. *Modelling* dengan algoritma C4.5

Secara umum, fase algoritma C4.5 dalam membangun decision tree adalah sebagai berikut :

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



Gambar 3.2. *Flowchart* Algoritma C4.5

Pada penerapan *data mining* ini dibantu menggunakan aplikasi Weka 3.9. Weka merupakan sebuah perangkat lunak yang memiliki banyak algoritma *machine learning* untuk keperluan *data mining*.

9. Pengujian Model

Dalam tahap ini akan dilakukan pengujian model terhadap data *testing*, *rule* atau aturan pohon keputusan yang telah dihasilkan dari tahap sebelumnya akan diuji menggunakan data testing dengan bantuan tools Weka 3.9.

10. Evaluasi Hasil

Tahap ini digunakan untuk menguji kualitas dari data, apakah pola atau informasi yang ditemukan bersesuaian atau bertentangan dengan fakta sebelumnya. Evaluasi algoritma menggunakan *Confusion Matrix* berbentuk matrix 3x3.

11. Kesimpulan

Setelah dilakukan evaluasi kemudian maka dapat ditarik kesimpulan apakah informasi yang ditemukan sesuai dengan fakta sebelumnya atau tidak, sehingga akan diperoleh pengetahuan baru.

4. Hasil dan Pembahasan

4.1 *Modelling* Algoritma C4.5 Pertama

Jumlah data *training* yang akan digunakan untuk pengujian dalam pembentukan *decision tree* adalah sebanyak 977, sedangkan jumlah data testing adalah 130. Data *training* dan *testing* yang digunakan dalam *modelling* algoritma C4.5 ini merupakan data yang berbeda. Data *training* diambil dari data mahasiswa angkatan 2006 s.d 2010, sedangkan data *testing* diambil dari data mahasiswa tahun 2012.

Tabel 4.1 Cuplikan *Data training* yang telah melewati data *preprocessing*

Program Studi	J K	Asal Daerah	Asal Sekolah	IPK4	Gaji Ortu	Status Pekerjaan	Kelas	Status
Komputerisasi Akuntansi	P	Cirebon	SMA	Besar	Sedang	Tidak Bekerja	SORE	Lulus Tepat Waktu
Komputerisasi Akuntansi	L	Cirebon	MAN	Kecil	Sedang	Bekerja	SORE	Lulus Terlambat
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Besar	Rendah	Bekerja	PAGI	Lulus Terlambat
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Besar	Sedang	Tidak Bekerja	SORE	Lulus Tepat Waktu
Komputerisasi Akuntansi	P	Luar Cirebon	SMA	Sedang	Tinggi	Bekerja	PAGI	Lulus Terlambat

Tabel 4.1 merupakan hasil dari data *preprocessing* yang kemudian disebut dengan *data training*, data ini akan dipelajari untuk membentuk pola atau pohon keputusan. Algoritma yang akan digunakan adalah C4.5.

Algoritma C4.5 akan menghasilkan *decision tree* atau pohon keputusan. Cara menentukan atribut mana yang akan menjadi akar atau *root*, dilihat pada nilai *gain* tertinggi dari atribut-atribut yang ada. Perhitungan *node* secara manual untuk menentukan *root* dapat dilihat pada Tabel 4.2.

Tabel 4.2 Tabel Perhitungan Node

Node			Jml data	TW	T	DO	Entrophy	Gain
	Total		977	162	276	539	1,418426933	
	Sekolah Asal	SMA	418	96	135	187	1,533206573	0,029342056
		SMK	470	55	118	297	1,281241536	
		MAN	64	10	17	37	1,383501191	
		PAKET C	25	1	6	18	1,021119189	
	IPK	Besar	222	108	90	24	1,380750421	0,537290897
		Sedang	363	54	182	127	1,438411124	
		Kecil	392	0	4	388	0,082143051	
	Gaji Ortu	Tinggi	9	3	6	0	0,918295834	0,014262044
		Sedang	633	92	179	362	1,380761634	
		Rendah	335	67	91	177	1,461439762	
	Status Pekerjaan	Bekerja	639	57	177	405	1,241016342	0,062901185
		Mahasiswa	338	105	99	134	1,572009507	
	Program Studi	KA	322	86	97	139	1,555329537	0,063116799
		MI	296	32	66	198	1,217755316	
		TI	217	14	55	148	1,13354731	
		SI	142	30	58	54	1,531905284	
	Jenis Kelamin	L	592	67	145	380	1,263406195	0,040403887
		P	385	95	131	159	1,554265062	
	Asal daerah	Cirebon	749	122	216	411	1,418893431	0,000454012
		Luar Cirebon	228	40	60	128	1,414948964	
	Kelas	Pagi	545	85	149	311	1,391444441	0,001390926
		Sore	432	77	127	228	1,449321664	

Keterangan :

TW : Lulus tepat waktu

T : Lulus terlambat

DO : *Drop Out*

Perhitungan untuk mendapatkan nilai *entropy* dan *gain* seperti terlihat pada Tabel 4.2, yaitu dengan menggunakan rumus berikut :

$$Entropy(S) = \sum_{i}^{c} -p_i \log_2 p_i \quad \text{..... (1)}$$

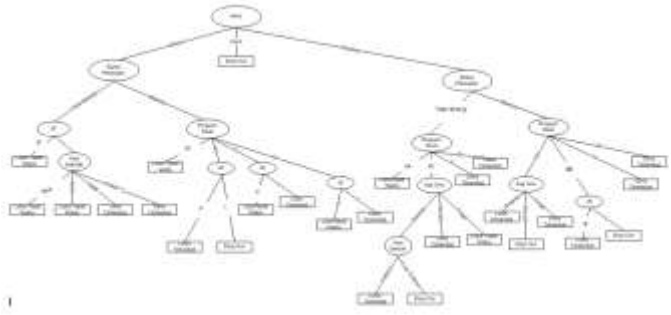
Menghitung nilai *Gain* pada atribut asal sekolah dihitung dengan menggunakan formula *gain* sebagai berikut:

$$Gain\ (S, A) = entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si) \quad \text{..... (2)}$$

Dari Tabel 4.2, atribut dengan nilai *gain* tertinggi dipilih sebagai *node* pertama pada pohon keputusan. Pada *node* selanjutnya akan diisi oleh atribut-atribut yang bernilai *gain* lebih rendah dan begitu seterusnya. *Node* yang tidak memiliki percabangan akan menunjukkan *output* dari setiap cabangnya yang dikenal dengan nama *leaf* atau daun.

Pada Tabel 4.2 terlihat bahwa atribut IPK4 memiliki nilai *gain* tertinggi diantara atribut yang lainnya, maka atribut IPK4 akan menjadi *root* pohon keputusan. Kemudian dilanjutkan dengan atribut status pekerjaan, asal sekolah

dan gaji orangtua. Diakhiri oleh status yang menyatakan keterangan lulus tepat waktu, lulus terlambat, *drop out*. Setalah semua *gain* diperoleh maka akan terbentuk pohon keputusan seperti yang terlihat pada Gambar 4.1.



Gambar 4.1
Decision Tree pada aplikasi Weka 3.9

Rule atau aturan yang dihasilkan dari 977 sampel *training* menghasilkan 8 *rule* lulus tepat waktu, 12 *rule* lulus terlambat dan 6 *rule drop out.* Aturan ini digunakan pada proses *testing* untuk memprediksi kelulusan, berikut keterangannya :

IPK4 = Besar
| STATUS PEKERJAAN = Tidak Bekerja
| | JK = P: Lulus Tepat Waktu (74.0/21.0)
| | JK = L
| | | ASAL SEKOLAH = SMA: Lulus Tepat Waktu (25.0/13.0)
| | | ASAL SEKOLAH = MAN: Lulus Tepat Waktu (1.0)
| | | ASAL SEKOLAH = SMK: Lulus Terlambat (15.0/5.0)
| | | ASAL SEKOLAH = PAKET C: Lulus Terlambat (0.0)
| STATUS PEKERJAAN = Bekerja
| | Program Studi = Komputerisasi Akuntansi
| | | JK = P: Lulus Terlambat (24.0/8.0)
| | | JK = L: Drop Out (9.0/4.0)
| | Program Studi = Manajemen Informatika: Lulus Tepat Waktu (32.0/17.0)
| | Program Studi = Teknik Informatika
| | | JK = P: Lulus Tepat Waktu (2.0)
| | | JK = L: Lulus Terlambat (13.0/6.0)
| | Program Studi = SISTEM INFORMASI
| | | JK = P: Lulus Terlambat (18.0/8.0)
| | | JK = L: Lulus Tepat Waktu (9.0/3.0)
IPK4 = Kecil: Drop Out (392.0/4.0)
IPK4 = Sedang
| STATUS PEKERJAAN = Tidak Bekerja
| | Program Studi = Komputerisasi Akuntansi: Lulus Tepat Waktu (44.0/19.0)
| | Program Studi = Manajemen Informatika
| | | GAJI ORTU = Sedang
| | | | Asal Daerah = Cirebon: Lulus Terlambat (11.0/2.0)
| | | | Asal Daerah = Luar Cirebon: Drop Out (4.0)
| | | GAJI ORTU = Rendah : Lulus Terlambat (0.0)
| | | GAJI ORTU = Tinggi : Lulus Tepat Waktu (6.0/1.0)
| | Program Studi = Teknik Informatika: Lulus Terlambat (35.0/11.0)
| | Program Studi = SISTEM INFORMASI: Lulus Terlambat (20.0/6.0)
| STATUS PEKERJAAN = Bekerja
| | Program Studi = Komputerisasi Akuntansi
| | | GAJI ORTU = Sedang : Lulus Terlambat (71.0/22.0)
| | | GAJI ORTU = Rendah : Drop Out (17.0)
| | | GAJI ORTU = Tinggi : Lulus Terlambat (1.0)
| | Program Studi = Manajemen Informatika
| | | JK = P: Lulus Terlambat (26.0/10.0)
| | | JK = L: Drop Out (58.0/31.0)
| | Program Studi = Teknik Informatika: Drop Out (37.0/16.0)
| | Program Studi = SISTEM INFORMASI: Lulus Terlambat (33.0/16.0)

4.2 Pengujian Model Pertama

Pengujian model menggunakan data *testing*, pengujian ini dilakukan untuk mengukur sejauh mana *classifier* melakukan klasifikasi dengan benar. Pengujian model menggunakan alat bantu yaitu tools Weka 3.9. Proses pengujian pada dasarnya membandingkan hasil prediksi dengan data sesungguhnya.

Dari pengujian tersebut dapat diketahui tingkat kebenaran prediksi, terlihat pada Tabel 4.3.

Tabel 4.3 Tingkat Kebenaran Prediksi

Jumlah *Rule*	Jumlah Data *Testing*	Prediksi Benar	Prediksi Salah
14	130	109	21

4.3 Evaluasi Model Pertama

Pengukuran data dilakukan dengan *confusion matrix* 3x3. Evaluasi pengukuruan ini membandingkan nilai akurasi dan *error rate* algoritma *decision tree* C4.5.

Dengan bantuan tools Weka, maka di dapatkan tabel *confusion matrix* untuk metode C4.5 seperti yang ditunjukkan oleh Gambar 4.2

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 28  7  0 |  a = Lulus Tepat Waktu
  8 42  5 |  b = Lulus Terlambat
  0  1 39 |  c = Drop Out
```

Gambar 4.2
Confusion Matrix Metode *Decision Tree* C4.5

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

$$Akurasi = \frac{AA+BB+CC}{AA+AB+AC+BA+BB+BC+CA+CB+CC} \times 100\% \quad \text{..... (3)}$$

$$Error\ Rate = \frac{AB+AC+BA+BC+CA+CB}{AA+AB+AC+BA+BB+BC+CA+CB+CC} \times 100\% \quad \text{..... (4)}$$

Tabel 4.4 Tabel Hasil Akurasi dan *Error Rate*

Dataset	Nilai Akurasi	Nilai *Error Rate*
Data *Testing*	84%	16%

4.4 *Modelling* Algoritma C4.5 Kedua

Pada uji *modelling* yang pertama, telah diketahui bahwa IPK4 merupakan *root* pohon keputusan karena memiliki nilai *gain* tertinggi diantara atribut-atribut lainnya. IPK 4 menjadi atribut utama yang digunakan dalam prediksi kelulusan mahasiswa.

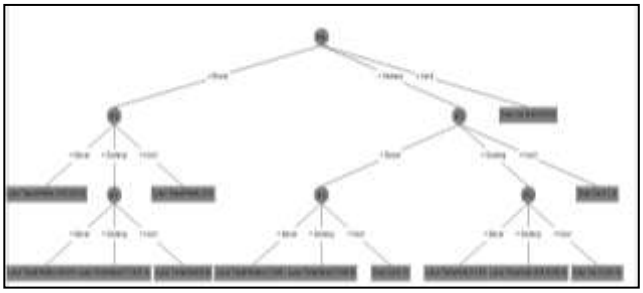
Pada *modeling* yang kedua, proses prediksi kelulusan akan dilakukan berdasarkan Indeks prestasi per semester. Atribut IPK4 akan di *breakdown* menjadi empat kategori

yaitu IP semester 1, 2, 3 dan 4. Atribut IP semester 1 sampai dengan 4 akan dijadikan atribut untuk memprediksi kelulusan, sehingga akan diketahui IP semester berapa saja yang dapat memprediksi kelulusan dan IP semester berapa saja yang tidak masuk dalam pola prediksi kelulusan.

Tabel 4.5 Cuplikan Data *Training* Untuk *Modelling* Yang Kedua

IPS1	IPS2	IPS3	IPS4	STATUS
Sedang	Besar	Besar	Besar	Lulus Tepat Waktu
Sedang	Sedang	Sedang	Sedang	Lulus Terlambat
Sedang	Sedang	Sedang	Sedang	Lulus Terlambat
Sedang	Besar	Besar	Besar	Lulus Tepat Waktu

Modelling algoritma C4.5 untuk data *training* diatas menggunakan alat bantu yaitu tools Weka 3.9. dengan menggunakan tools Weka maka dapat langsung terbentuk pohon keputusan.



Gambar 4.3
Decision Tree C4.5 Pada *Modelling* yang Kedua menggunakan Weka 3.9

Rule atau aturan yang dihasilkan dari 977 sampel *training* menghasilkan 8 *rule* lulus tepat waktu, 12 *rule* lulus terlambat dan 6 *rule drop out.* Aturan ini digunakan pada proses *testing* untuk memprediksi kelulusan, berikut keterangannya :

IP4 = Besar
| IP2 = Besar: Lulus Tepat Waktu (105.0/25.0)
| IP2 = Sedang
| | IP3 = Besar: Lulus Tepat Waktu (40.0/19.0)
| | IP3 = Sedang: Lulus Terlambat (77.0/31.0)
| | IP3 = Kecil: Lulus Terlambat (0.0)
| IP2 = Kecil: Lulus Tepat Waktu (0.0)
IP4 = Sedang
| IP2 = Besar
| | IP3 = Besar: Lulus Tepat Waktu (12.0/6.0)
| | IP3 = Sedang: Lulus Terlambat (10.0/5.0)
| | IP3 = Kecil: Drop Out (1.0)
| IP2 = Sedang
| | IP3 = Besar: Lulus Terlambat (13.0/5.0)
| | IP3 = Sedang: Lulus Terlambat (254.0/100.0)
| | IP3 = Kecil: Drop Out (13.0/1.0)
| IP2 = Kecil: Drop Out (12.0)
IP4 = Kecil: Drop Out (440.0/33.0)

4.5 Pengujian Model Kedua

Pengujian model kedua ini menggunakan data *testing* yang sama dengan model pertama, hanya atribut saja yang berbeda. Pengujian ini dilakukan untuk melihat nilai akurasi dan *error* dari aturan atau *rule* yang dihasilkan pada proses *modelling* sebelumnya.

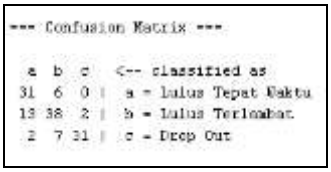
Dari pengujian tersebut dapat diketahui tingkat kebenaran prediksi, terlihat pada Tabel 4.6.

Tabel 4.6 Tingkat Kebenaran Prediksi

Jumlah *Rule*	Jumlah Data *Testing*	Prediksi Benar	Prediksi Salah
11	130	100	30

4.6 Evaluasi Model Kedua

Dengan bantuan tools Weka 3.9, maka di dapatkan tabel *confusion matrix* untuk metode C4.5 seperti yang ditunjukkan oleh Gambar 4.7.



Gambar 4.4
Confusion Matrix Metode *Decision Tree* C4.5

Tabel 4.7 Tabel Hasil Akurasi dan *Error Rate*

Dataset	Nilai Akurasi	Nilai *Error Rate*
Data *Testing*	77%	23%

4.7 Pola Pengetahuan

Pada model pertama, dari beberapa atribut yang diuji, ternyata IPK 4 adalah atribut utama dalam pola prediksi kelulusan. Selanjutnya pada model kedua, diketahui bahwa atribut utama pembentuk pohon keputusan adalah IP semester 4, sedangan IP semester 1 tidak masuk dalam pohon keputusan. Artinya pada model kedua IP yang sangat menentukan kelulusan adalah IP di semester 4.

Model pertama merupakan model yang akan digunakan untuk prediksi kelulusan karena setelah dilakukan pengujian pada model kedua dimana pengujian berdasarkan IP semester 1 sampai dengan 4 diketahui bahwa IP semester 4 yang menjadi atribut utama dalam pola prediksi kemudian diikuti oleh atribut IP semester 2 dan 3.

Dari uraian tersebut, telah diperoleh pengetahuan baru dari pola prediksi kelulusan ini, yaitu untuk memprediksi lulus tepat waktu, lulus terlambat dan *drop out* tidak dapat dilakukan di tahun pertama perkuliahan, melainkan dapat dilakukan setelah mahasiswa menempuh perkuliahan sampai dengan semester 4. Nilai IPK semester 4 sangat tepat digunakan sebagai atribut utama dalam prediksi kelulusan, kemudian diikuti atribut lainnya yaitu : status pekerjaan, asal sekolah, jenis kelamin, program studi, gaji orangtua dan asal daerah. Berikut ini adalah pola prediksi kelulusan tepat waktu, terlambat dan *drop out* yang dihasilkan dari pohon keputusan pada model pertama :

Prediksi Lulus Tepat Waktu (TW)

1. Jika IPK 4 besar, tidak bekerja dan jenis kelaminnya perempuan maka lulus tepat waktu
2. Jika IPK 4 besar, tidak bekerja, jenis kelamin laki-laki dan asal sekolah SMA maka lulus tepat waktu
3. Jika IPK 4 besar, tidak bekerja, jenis kelamin laki-laki dan asal sekolah MAN maka lulus tepat waktu
4. Jika IPK 4 besar dan program studinya manajemen informatika maka lulus tepat waktu
5. Jika IPK4 besar, program studinya Teknik Informatika dan jenis kelaminnya perempuan maka lulus tepat waktu

6. Jika IPK 4 besar, program studinya sistem informasi dan jenis kelaminnya laki-laki maka lulus tepat waktu
7. Jika IPK 4 sedang, tidak bekerja dan program studinya komputerisasi akuntansi maka lulus tepat waktu
8. Jika IPK 4 sedang, tidak bekerja, program studinya manajemen informatika, gaji ortu tinggi maka lulus tepat waktu

Prediksi Lulus Terlambat (T)

1. Jika IPK 4 besar, tidak bekerja, jenis kelamin laki-laki dan asal sekolah SMK maka lulus terlambat
2. Jika IPK 4 besar, bekerja, program studinya komputerisasi akuntansi dan jenis kelaminnya perempuan maka lulus terlambat
3. Jika IPK4 besar, program studinya Teknik Informatika dan jenis kelaminnya laki-laki maka lulus terlambat
4. Jika IPK 4 besar, program studinya sistem informasi dan jenis kelaminnya perempuan maka lulus terlambat
5. Jika IPK 4 sedang, tidak bekerja, program studinya manajemen informatika, gaji ortu sedang dan asal daerah Cirebon maka lulus terlambat
6. Jika IPK 4 sedang, tidak bekerja, program studinya manajemen informatika, gaji ortu rendah maka lulus Terlambat
7. Jika IPK 4 sedang, tidak bekerja dan program studinya teknik informatika maka lulus terlambat
8. Jika IPK 4 sedang, tidak bekerja dan program studinya sistem informati maka lulus terlambat
9. Jika IPK 4 sedang, bekerja, program studinya komputerisasi akuntansi dan gaji ortu sedang maka lulus telambat
10. Jika IPK 4 sedang, bekerja, program studinya komputerisasi akuntansi dan gaji ortu tinggi maka lulus telambat
11. Jika IPK 4 sedang, bekerja, program studinya manajemen informatika dan jenis kelaminnya perempuan maka lulus terlambat
12. Jika IPK 4 sedang, bekerja, program studinya sistem informasi maka lulus terlambat

Prediksi *Drop Out* (DO)

1. Jika IPK 4 besar, bekerja, program studinya komputerisasi akuntansi dan jenis kelamin laki-laki maka drop out
2. Jika IPK 4 kecil maka drop out
3. Jika IPK 4 sedang, tidak bekerja, program studinya manajemen informatika, gaji ortu sedang dan asal daerah Luar Cirebon maka lulus drop out
4. Jika IPK 4 sedang, bekerja, program studinya komputerisasi akuntansi dan gaji ortu rendah maka drop out
5. Jika IPK 4 sedang, bekerja, program studinya manajemen informatika dan jenis kelaminnya laki-laki maka drop out

Jika IPK 4 sedang, bekerja, program studinya teknik informatika maka drop out

.5. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, dapat diambil kesimpulan sebagai berikut :

1 Data mining dengan algoritma C4.5 dapat diimplementasikan untuk memprediksi kelulusan mahasiswa STMIK WIT dengan tiga kategori yaitu lulus tepat waktu, lulus terlambat dan *drop out*. Variabel yang berpengaruh dalam hasil prediksi adalah IPK4, status pekerjaan, asal sekolah, asal daerah, jenis kelamin, gaji ortu dan program studi, sedangkan variabel yang tidak bepengaruh pada hasil prediksi adalah kelas.

2 Pada penelitian ini, penggunaan algoritma C4.5 mampu melakukan prediksi dengan baik (84%) terhadap masa studi mahasiswa yang tepat waktu, terlambat dan *drop out*. Pembentukan pohon keputusan (*Decision Tree*) dapat digunakan oleh pengelola akademik di dalam memetakan mahasiswa yang berpotensi mengalami keterlambatan masa studi dan *drop out* di masa mendatang. Penerapan *Educational Data Mining* (EDM) memberikan kemajuan dan kontribusi pada dunia pendidikan dan pada bidang riset data mining.

3 Dengan alat bantu Weka, hasil uji coba menggunakan 130 data *testing*, pola yang dibentuk pada model pertama memiliki nilai akurasi kecocokan sebesar 84%, sedangkan model kedua 77%.

Pola atau aturan yang dihasilkan dari model pertama dengan menggunakan 977 data *training* menghasilkan 8 *rule* lulus tepat waktu, 12 *rule* lulus terlambat dan 6 *rule drop out.*

Daftar Pustaka

Al Fatta, Hanif. *Analisa dan Perancangan Sistem Informasi.* Andi Offset, Yogyakarta, 2007.

Bhardwaj, Brijesh Kumar & Saurabh Pal. *Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security.* 2011

Kadir, Abdul. *Pengantar Sistem Informasi edisi Revisi*. Andi Offset, Yogyakarta, 2013.

Kusrini & Luthfi Taufiq Emha. Algoritma Data Mining. Andi Offset, Yogyakarta, 2009.

Meilani, B.D. dan Susanti, Nofi. *Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naïve Bayes*. Jurnal link vol 21 no. 2, 2014.

Putra, Ade. Solusi Prediksi Mahasiswa Drop Out Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Bina Darma. Jurnal Simetris, Vol 8 No. 1. 2017.

Santoso, Sani dan Dedy Suryadi. *Pengantar Data Mining.* Andi Offset,Yogyakarta, 2010.