IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS: UNIVERSITAS DEHASEN BENGKULU)

Siska Haryati, Aji Sudarsono, Eko Suryana

Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Dehasen Bengkulu
Jl. Meranti Raya No. 32 Kota Bengkulu 38228 Telp. (0736) 22027, 26957 Fax. (0736) 341139

ABSTRAK

The purpose of this study is to use the C4.5 decision tree algorithm-based and implemented into an application that RapidMiner is expected to improve the accuracy of the analysis of the study period the student. This research was conducted at the Dehasen University of Bengkulu. In the study discussed by monitoring the results of studies at the university in the form of the GPA and the number of credits that have not been accurate for determining a student to graduate on time or not. In this study was to classify the grading of student used data mining techniques with C4.5 algorithms and implemented into Rapid Miner, it aims to see the results of the development can graduate on time or not. From the research results prove that the algorithm C4.5 is more accurate than analysis conducted by analyst's students. This is evidenced by the results of the evaluation study found C4.5 algorithms capable of analyzing the punctuality of students completing their study.
Keyword: Data mining, C4.5 Algorithm

INTISARI

Tujuan dari penelitian ini adalah dengan menggunakan pohon keputusan berbasis algoritma C4.5 dan dimplementasikan ke suatu aplikasi yaitu RapidMiner diharapkan dapat meningkatkan keakuratan analisa masa studi mahasiswa. Pada penelitian dibahas dengan memantau hasil belajar di universitas berupa nilai IPK dan Jumlah SKS yang belum akurat untuk menentukan seorang mahasiswa lulus tepat waktu atau tidak. Pada Penelitian ini untuk mengklasifikasikan kelululan mahasiswa digunakan teknik data mining dengan algoritma C4.5 dan diImplementasikan ke Rapid Miner, hal tersebut bertujuan untuk melihat hasil perkembangan mahasiswa apakah dapat lulus tepat waktu atau tidak. Dari hasil penelitian terbukti bahwa algoritma C4.5 lebih akurat dibandingkan analisa yang dilakukan oleh analis mahasiswa. Hal ini dibuktikan dengan hasil evaluasi penelitian bahwa algoritma C4.5 mampu menganalisa tingkat ketepatan waktu mahasiswa menyelesaikan masa studinya.
Kata Kunci: Data mining, Algoritma C4.5

I. PENDAHULUAN

Kemajuan teknologi pengetahuan dan teknologi beserta aplikasinya disegala bidang tidak bisa lepas dari perangkat komputer. Penggunaan komputer sudah menjangkau hampir segala bidang dalam aktivitas kehidupan manusia, baik dalam lingkungan pendidikan, organisasi, perusahaan maupun masyarakat umum. Penggunaan komputer terbukti banyak membantu kita dalam melakukan pekerjaan dengan lebih baik. Didalam suatu Universitas penggunaan komputer.

Kebutuhan akan layanan informasi sangatlah penting, melalui aplikasi yang menggunakan data mining dapat mempermudah dalam proses memprediksi masa studi mahasiswa.

Semakin ketatnya persaingan mahasiswa dalam mendapatkan lapangan pekerjaann menuntut ilmu di perguruan tinggi menghasilkan sarjana yang berkualitas dam memiliki daya saing. Untuk itu, setiap perguruan tinggi selalu melakukan evaluasi performansi mahasiswa. Hasil evaluasi tersebut disimpan dalam basis data akademik. Data tersebut dapat digunakan untuk sebagai pendukung keputusan oleh manajemen perguruan tinggi. Salah satu variabel indikator efisiensi proses pendidikan adalah informsasi mengenai lama masa studi mahasiswa.

Pertumbuhan yang sangat pesat dari akumulasi data telah menciptakan kondisi kaya akan data tapi minim informasi. Data Mining merupakan penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data dalam jumlah besar yang diharapkan dapat mengatasi kondisi tersebut. Data Mining sendiri memiliki beberapa teknik salah satunya klasifikasi. Teknik klasifikasi terdiri beberapa metode, dan dicision tree adalah bagian dari metode klasifikasi. Kemudian metode dicision tree memiliki algoritma, algoritma C4.5 adalah salah satu dari algoritma yang memiliki dicision tree.

Program Sarjana (S1) Fakultas Ilmu Komputer Universitas Dehasen Bengkulu adalah program pendidikan akademik setelah pendidikan menengah,

yang memiliki beban sekurang-kurangnya 144 (sesuaikan pada kampus) (seratus empat puluh empat) sks (satuan kredit semester) yang dijadwalkan untuk 8 (delapan) semester dan dapat ditempuh dalam waktu kurang dari 8 (delapan) semester paling lama 14 (empat belas) semester. Hal ini menunjukan bahwa masih banyak mahasiswa Program Sarjana (S1) regular Fakultas Ilmu Komputer yang menempuh lama studi dari 8 semester dari yang dijadwalkan 8 semester.

Pada penilitian ini dibuat suatu Aplikasi data mining dengan algoritma C4.5 guna untuk menganalisis kemungkinan mahasiswa lulus lebih dari 8 semester dengan melakukan klasifikasi dari kumpulan data mahasiswa yang telah lulus.

Universitas Dehasen Bengkulu merupakan salah satu Universitas yang ada dibengkulu yang bergerak pada bidang pendidikan. Untuk memprediksi masa studi mahasiswa selama ini masih berdasarkan perkiraan saja, belum diperhitungkan secara tepat.

II. TINJAUAN PUSTAKA

A) Pengertian Implementasi

Browne dan Wildavsky (dalam Nurdin dan Usman, 2004:70) mengemukakan bahwa ”implementasi adalah perluasan aktivitas yang saling menyesuaikan”. Pengertian implementasi sebagai aktivitas yang saling menyesuaikan.

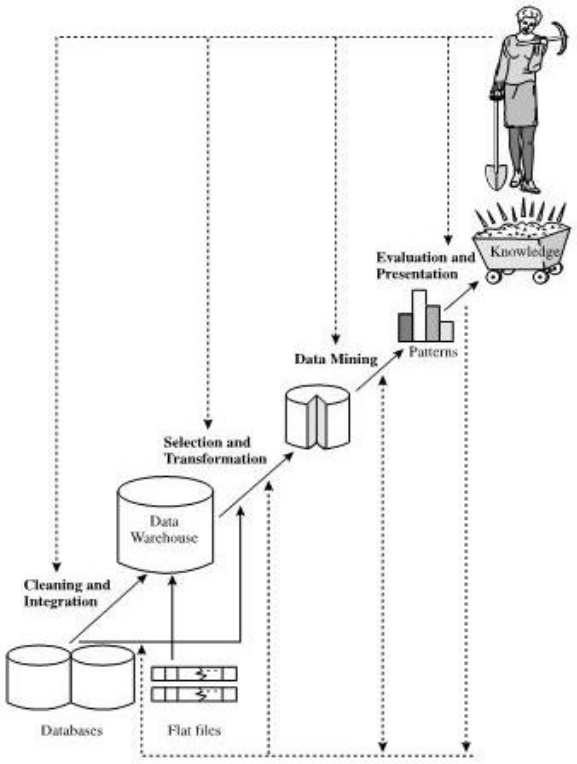
Setelah sistem informasi yang baru dirancang, sistem tersebut harus diimplementasikan sebagai sistem kerja, dan dipelihara agar dapat berrjalan dengan baik. Proses implementasi yang akan kita bahas dalam bagian ini adalah kelanjutan dari tahap investigasi, analis, dan desain siklus pengembangan sistem yang kita bahas. Implementasi adalah langkah yang vital dalam pengembangan teknologi informasi untuk mendukung karyawan, pelanggan, dan pihak-pihak yang berkepentingan lainnya.

B) Pengertian Data Mining

Tan (2006:2) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah data mining kadang disebut juga *knowledge discovery*. (Eko Prasetyo, 2012:2)

Salah satu teknik yang dibuat dalam *data mining* adalah adalah bagaimanana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Dalam *data mining*, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil. Semua hal tersebut bertujuan mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan diharapkan dapat tercapai.

Data mining merupakan bagian dari proses Knowledge Discovery from Data (KDD). Dibawah ini digambarkan skema dari proses KDD.



Gambar 1. Data mining sebagai dari proses *knowledge discovery* (sumber gambar: Data mining concept and tehniques, Han & Kamber).

Gambar 1 menunjukkan proses penjelajahan pengetahuan dimulai dari beberapa *database* dilakukan proses *cleaning* dan *integration* sehingga menghasilkan *data warehouse*. Dilakukan proses *selection* dan *transformation* yang kemudian disebut sebagai *data mining* hingga menemukan pola dan memperoleh pengetahuan dari data (*knowledge*).

Terdapat beberapa teknik data mining yang sering disebut-sebut dalam literatur. Namun ada 3 teknik data mining yang populer (Santosa 1999), yaitu:

1) Association Rule Mining

Association Rule mining adalah teknik mining untuk menemukan asosiatif antara kombinasi atribut. Contoh dari aturan asosiatif dari analisa pembelian di

suatu pasar swalayan dapat mengatur penempatan barangnya atau merancang strategi pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

2) Clustering

Berbeda dengan *association rule mining* dan klasifikasi dimana kelas data telah ditentukan sebelumnya, clustering dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*. Prinsip *clustring* adalah memaksimalkan kesamaan antar *cluster*. *Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi.

3) Klasifikasi

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

C) Pohon Keputusan (Decision Tree)

Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah di interperensi manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Misalkan untuk menentukan main tenis, kriteria yang digunakan adalah cuaca, angin, iklim dan temperatur.

Manfaat utama menggunakan pohon keputusan adalah kemampuannya untuk membreak downproses pengambilan keputusan yang kompleks menjadi lebih simpel sehingga pengambilan keputusan akan menjadi lebih menginterprestasikan solusi permasalahan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuat variabel target. Pohon keputusan memadukan antara eksplorasi data dan pemodelan sehingga sangat bagus sebagai langkah awal pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

D) Algoritma C4.5

Pohon keputusan mirip sebuah struktur pohon dimna terdapat node internal (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Pohon keputusan bekerja mulai dari akar paling atas, jika diberikan sejumlah data uji, misalnya X dimana kelas dari data X belum diketahui, maka pohon keputusan akan menelusuri mulai dari akar sampai node dan setiap nilai dari atribut sesuai data X diuji apakah sesuai dengan aturan pohon keputusan, kemudian pohon keputusan akan memprediksi kelas dari tupel X.

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5.Di akhir tahun 1970 hingga di awal tahun 1980-an, J.Ross Quinlann seorang peniliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (Iterative Dichotomiser), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B.Hunt, J. Marin, dan P.T.Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning.*

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 (Kusrini & Lutfhi, 2009), yaitu:

1) Menyiapkan data training. Data training biasnya dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2) Menentukan akar dari pohon . akar akan diambil dari atibut yang terpilih dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, hitung dahulu nilai entropy yaitu:

$$\text{Entropy}(S) = \sum_{i=1}^{n} - \text{pi} * \log_2 \text{pi}$$

Keterangan:
S : himpunan kasus
A : atribut
n : jumlah partisi S
pi : proporsi dari Si terhadap S

3) Kemudian hitung nilai Gain dengan metode *information gain:*

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|si|}{|s|} * \text{Entropy}$$

Keterangan:
S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

4) Ulangi langkah ke-2 hingga semua semua tupel terpartisi.
5) Proses partisi pohon keputusan akan berhenti saat:
   a) Semua tupel dalma node N mendapat kelas yang sama.
   b) Tidak ada atribut di dalam tupel yang dipartisi lagi.
   c) Tidak ada tupel di dalam cabang yang kosong

E) Klasifikasi Rule Based

Rule based atau algoritma berbasis aturan merupakan cara terbaik untuk merepresentasikan sejumlah bit data atau pengetahuan (Han & Kamber, 2006). *Rule based* biasanya dituliskan dalam bentuk logika *IF-THEN* atau jika dibuat persamaannya yaitu: *IF condition THEN conlusion* contoh sebuah *rule* yaitu:
IF age = youth AND student = yes THEN buys_computer = yes

Pernyataan IF dari persamaan di atas dikenal sebagai rule *antecedent* atau *precondition* sedangkan pernyataan THEN disebut sebagai *rule consequent*.

Dalam *rule antecedent* biasanya menyertakan satu atau lebih atribut (misalnya atribut *age* dan *student*) dan menggunakan logika AND jika menggunakan lebih dari satu atribut. Rule consequentmerupakan prediksi kelas, dalam contoh di atas prediksinya yaitu membeli komputer atau buys_computer = yes (Han & Kamber, 2006).

Aturan-aturan dalam rule based dapat diturunkan dari pohon keputusan yang telah terbentuk. Karena pohon keputuan yang besar, terkadang sulit untuk menginterpretasikan pohon bentuk keputusan (Han & Kamber, 2006). Agar pohon keputusan ini dapat lebih mudah dipahami oleh manusia, maka perlu diinterpretasikan dalam bentuk aturan-aturan atau *rule based*.

Dalam kasus ini tidak digunakan logika OR, karenaaturan-aturan diekstraksi langsung dari pohon keputusan yang disebut mutually *exclusive* dan *exhaustive*. Dengan *mutually exclusive* artinya tidak ada aturan yang berbenturan atau konflik karena tidak boleh ada dua aturan dalamdalam tupel yang sama. Sedangkan exhaustive artinya dalam satu set aturan merupakan kombinasi nilai yang mungkin, artinya setiap aturan pasti menggambarkan kombinasi atribut dan nilai yang mungkin (Han & Kamber, 2006).

F) Rapid Miner

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technologi Blanchardstown dan Ralf Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat *open* source dan dibuat dengan menggunakan program Java di bawah lisensi *GNU Public Licence* dan *Rapid Miner* dapat dijalankan di sistem operasi manapun. Dengan menggunakan Rapid Miner, tidak dibutuhkan kemampuan koding khusus, karena semua fasilits sudah disediakan. Rapid Miner dikhususkan untuk penggunaan data mining. Model yang disediakan juga cukup banyak dan lengkap, seperti Model Bayesian, Modelling, Tree Induction, Neural Network dan lain-lain.

Banyak metode yang disediakan oleh Rapid Miner mulai dari kllasifikasi, klustering, asosiassi dan lain-lain. Jika tidak ada model atau model alogoritma yang tidak ada dalam Weka, pengguna boleh menambahkan modul lain, karena weka bersifat *open source*, jadi siapapun dapat ikut mengembangkan perangkat lunak ini.

III. METODOLOGI PENELITIAN

A) Hardware dan Software

Hardware atau lebih dikenal dengan perangkat keras adalah peralatan fisik dari komputer yang dapat dilihat, dipegan, ataupun dipindahkan. (Supriyanto, 2008: 87). Hardware yang digunakan untuk membuat penelitian ini adalah:
1) Laptop dengan processor intel Core i3
2) Memory RAM 2 GB
3) Hardisk
4) Mouse
5) Modem
6) Printer Cannon IP 2700

Software adalah suatu program yang berisi instruksi-intruksi yang ditulis dalam bahasa komputer yang dimengerti oleh hardware komputer. (Supriyanto, 2008:2). Software yang digunakan untuk membuat laporan ini adalah:
1) Sistem Operasi *Windows Seven Ultimate*
2) *Rapid Miner5*
3) *Microsoft Office 2007*

B) Analisis Sistem

Analisa sistem adalah penguraian dari sistem informasi kedalam bagian komponennya dengan maksud untuk mengidentifikasi dan mengevaluasi permasalahan-permasalahan yang terjadi dan kebutuhan yang diharapkan sehingga dapat diusulkan perbaikan-perbaikannya. Analisa sistem merupakan tahap awal dala perancangan dan pengembangan sebuah sistem yang akan dirancang, karena tahap inilah akan diukur dan dieavakuasi tentang kinerja dari sistem yang dirancang. Identifikasi terhadap masalah yang ada dan langkah-langkah untuk kebutuhan perancangan yang diharapkan.

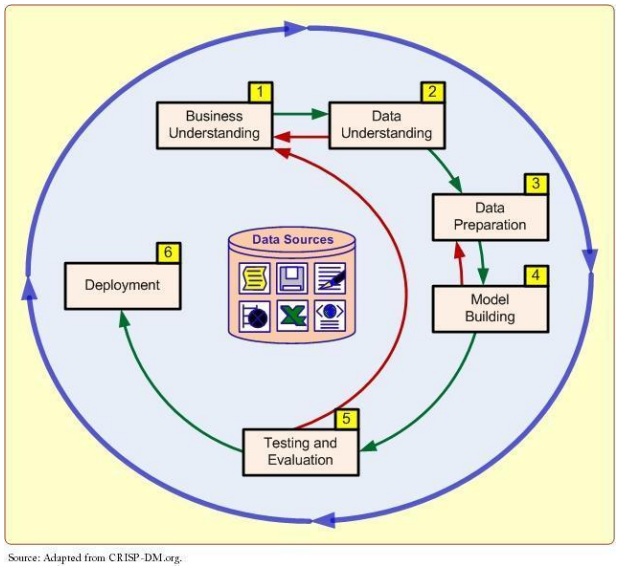
Dalam melakukan analisis sistem terlebih dahulu harus mengetahui dan memahami sistem, untuk menganalisa sistem diperlukan data dari sistem untuk dianalisa. Data yang diperlukan adalah hal-hal yang diperlukan untuk definisi data. Analisa data merupakan tahap untuk melakukan penganalisaan terhadap data-data yang dibutuhkan untuk perancangan sistem yang akan dibuat, dalam hal ini penulis mengambil data dari survei yang berhubungan dengan tema penelitian, untuk mencari informasi menyusun teori-teori yang berhubungan dengan dengan pembahasan sehingga terjadi perpaduan komplek antara satu dengan yang lainnya.

Dari hasil pra penelitian yang dilakukan diketahui bahwa sistem yang sudah berjalan dan digunakan saat ini masih manual. Pengelompok-kan dalam memprediksi masa studi mahasiswa tepat waktu atau tidak tepat waktu guna untuk menambahkan tingkat kualitas mahasiswa tersebut.

Ada beberapa tahap yang dilakukan dalam melakukan eksprimen ini, penulis menggunakan model *Cross-Standard Indudtry For Data Mining* (CRISP-DM) (LAROSE, 2005) yang terdiri dari 6 tahap, yaitu:

1) *Bussines/Research Understanding Phase*
2) *Data Understanding Phase* (Fase Pemahaman Data)
3) *Data Preparation Phase* (Fase Pengolahan Data)
4) *Modeling Phase* (Fase Permodelan)
5) *Evaluation Phase* (Fase Evaluasi)
6) *Deplooyment Phase* (Fase Penyebaran)

Keenam tahap tersebut disajikan pada Gambar 2.



Gambar 2. Tahap CRISP-DM (*Cross Industry standard process for data mining*)

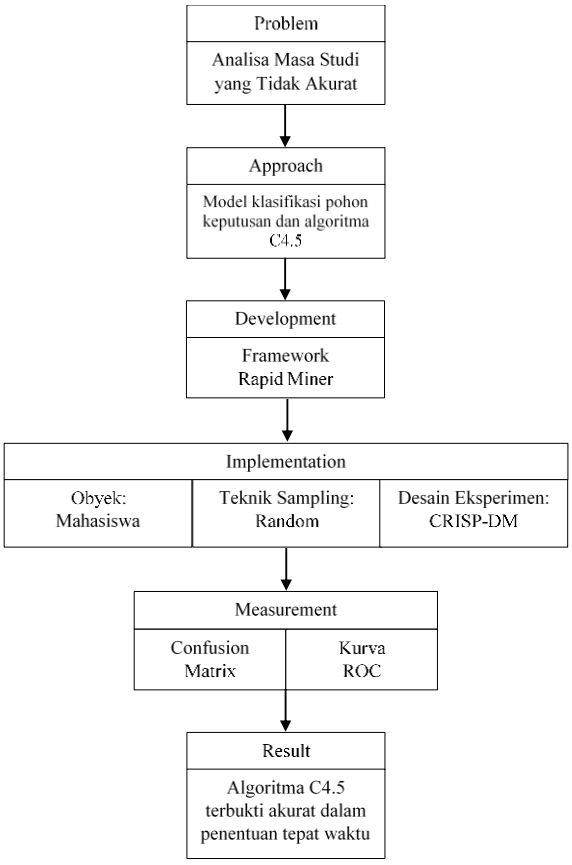
Ada beberapa tahap dalam menggunakan Rapid Miner:

1) Untuk menganalisa, dibutuhkan data training. Data tarining yang akan dimasukkan kedalam Rapid Miner bisa dalam format .csv, .xls, .mdb dan laini-lain. Data yang digunakan penulis adalah data dalam format .csv.
2) Buka program Rapid Miner, kemudian akan muncul tampilan awal. Untuk memasukkan data training yang telah dibuat sebelumnya, pilih menu File – Import Data – Import CSV file.
3) Tampilan jendela Data wizard dengan total 5 langkah. Pada langkah ke-1 ini tentukan nama file yang berisi data training dalam direktori kemudian pilih Next
4) Data training yang sebelumnya disimpan, akan tersimpan otomatis ke dalam Repositories. Pilih tab Repositories – NewLocalRepository – data_training. Geser data_training ke area Main Process. Untuk menambahkan model, pilih tab Operators- Modelling-Classification and Regression-Tree Induction-Decision Tree. Geser Decision Tree ke area Main Process dan hubungkan
5) Untuk melihat hasilnya, pilih process – Run maka akan tampil hasil berupa pohon keputusan.

C) Kerangka Pemikiran

Kerangka pemikiran dari penelitian ini, dimulai dari *problem* (permasalahan) analisa masa studi yang tidak akurat kemudian dibbuat *approach* (model) yaitu algoritma C4.5 untuk memecahkan permasalahan. Untuk mengem-bangkan aplikasi (*develovment*) berdasarkan model yang dibuat, digunakan Rapid Miner. Tahap berikutnya yaitu

implementation (*implementasi*), pada tahap ini objek implementasi dilakukan pada mahasiswa, teknik sampling menggunakan *random sample*, dan desain eksprimennya digunakan CRISP-DM. Kerangka pemikiran dalam penelitian ini disajikan pada Gambar 3.



Gambar 3. Kerangka Pemikiran

D) Pengijian Sistem

Pengujian Sistem digunakan untuk memproses kebenaran data. Untuk pengujian data maka dibuat perancangan pengujian menggunakan aplikasi Rapid Miner 5 dan database mahasiswa. Pengujian tersebut dilakukan secara langsung di Universitas Dehasen Bengkulu.

Penelitian ini mengunakan pengujian *White box testing*. *White box* testing adalah pengujian yang didasarkan pada pengecekan terhadap detail perancangan, menggunakan struktur kontrol dari desain program secara procedural untuk membagi pengujian ke dalam beberapa kasus pengujian. Secara sekilas dapat diambil kesimpulan white box testing merupakan petunjuk untuk mendapatkan program yang benar secara 100%.

Pengujian white box bertujuan untuk:

1) mengetahui cara kerja suatu perangkat lunak secara internal.
2) menjamin operasi-operasi internal sesuai dengan spesifikasi yang telah ditetapkan dengan menggunakan struktur kendali dari prosedur yang dirancang.

Pelaksanaan pengujian white box:

1. Menjamim seluruh independent path dieksekusi paling sedikit satu kali. Independent path adalah jalur dalam program yang menunjukkan paling sedikit satu kumpulan proses ataupun kondisi baru.
2. Menjalani logical decision pada sisi dan false.
3. Mengeksekusi pengulangan (looping) dalam batas-batas yang ditentukan.
4. Menguji struktur data internal.

Berdasarkan konsep pengujian; *White box* (*structural*) *testing*/*glass box testing*: memeriksa kalkulasi internal path untuk mengidentifikasi kesalahan.

Langkah-langkah white box:

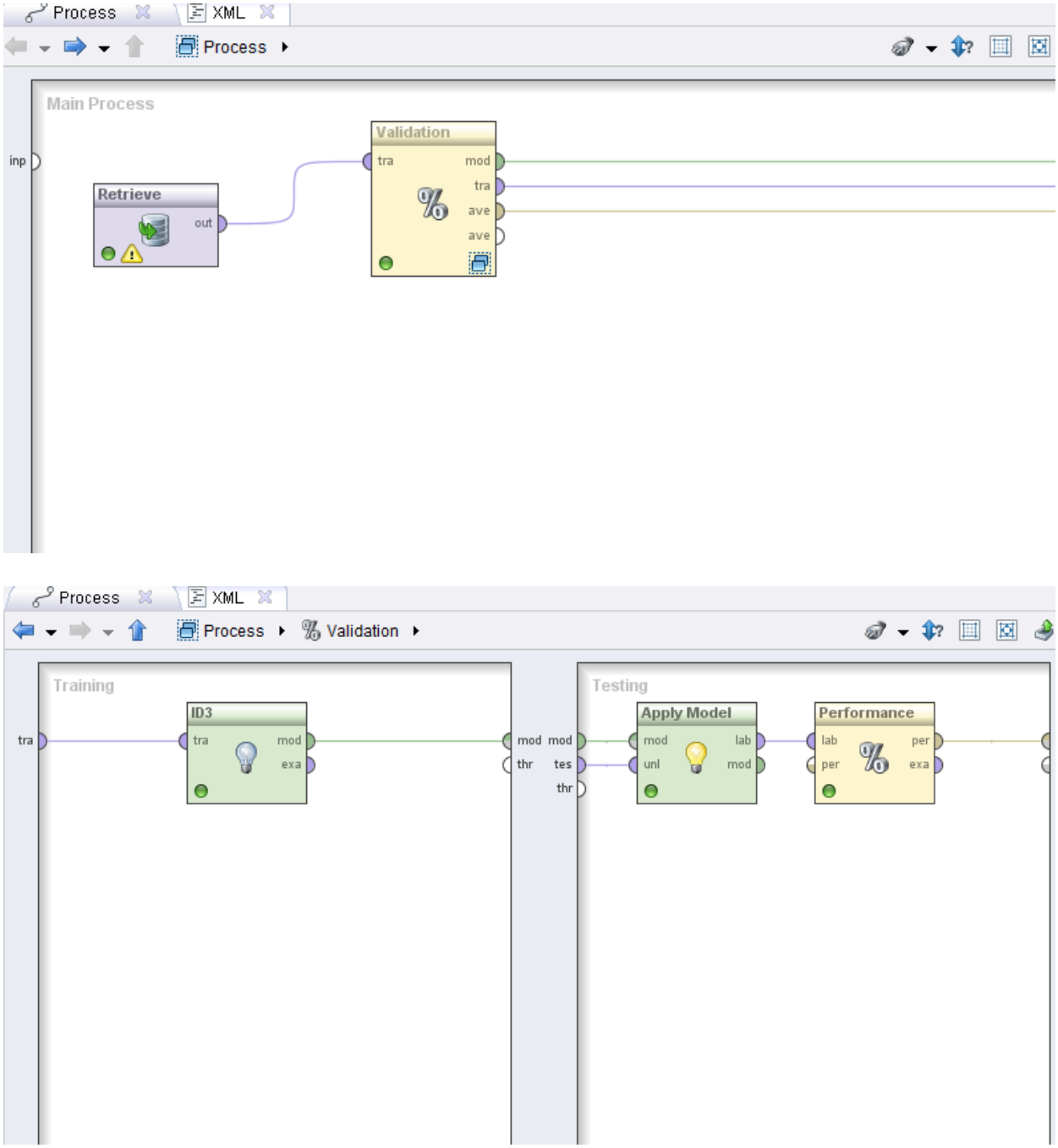
1. Mendefinisikan semua alur logika
2. Membangun kasus untuk digunakan dalam pengujian
3. Melakukan pengujian

IV. PEMBAHASAN

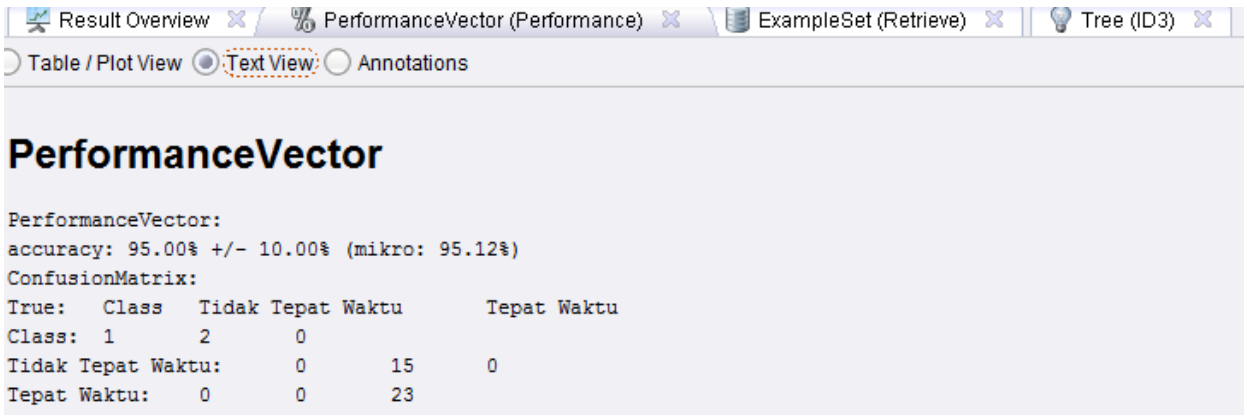
A) Validasi dan Evaluasi

Tujuan dari penelitian adalah untuk menganalisa prediksi kelulusan mahasiswa dengan menerapkan teknik klasifikasi data mining dengan algoritma *decision tree C4.5*. Pada tahap pengujian model ini data yang digunakan telah melewati tahap *preprocessing.* Desain model yang akan digunakan ditunjukkan pada Gambar 4.

1) *Retreiving data*: Operator ini digunakan untuk mengimport dataset yang akan digunakan, pada penelitian ini data diimport dari File csv.
2) *Validation*: Metode *validation* yang digunakan dalam penelitian ini adalah teknik sampling.
3) *ID3:* Metode algoritma yang digunakan
4) Decision tree: Metode klasifikasi yang digunakan dalam penelitian ini.
5) *Apply Model:* Operator yang digunakan metoden yang digunakan dalam penelitian ini C4.5.
6) *Performance*: Operator yang digunakan untuk mengukur performance akurasi dari model.



Gambar 4. Desain Model Metode C4.5



Gambar 5. Model Confusion Matrix

Pengujian akan dilakukan dari populasi data training. Jumlah data trainig yang akan digunakan sebanyak 83 dengan tingkat kesalahan 5% baik prediksi tepat waktu dan data prediksi tidak tepat waktu secara acak (*simple random sampling).*

Dillihat dari hasil masing–masing metode penyeleksian atribut, hasilnya menunjukkan kesamaan. Dari hasil pengujian diatas akan dievaluasi tingkat akurasinya menggunakkan model yaitu menggunkan *confusion matrix*.

B) Evaluasi model confusion matrix

Evaluasi ini menggunakan tabel seperti matrix di bawah ini:

	C1	C2
C1	true positives	False negatives
C2	false positives	Truenegatives

Sedangkan untuk model *confusion matrix* ditunjukkan pada gambar 5.

True positives merupakan tupel positif didata set yang diklasifikasikan positif. *True negatives* merupakan tupel negatif di data set yang diklasifikasikan negatif. *False positives* adalah tupel positif didata set yang diklasifikasikan negatif sedangkan *false negatives* merupakan jumlah tupel negatif yang diklasifikasikan positif. Kemudian masukan data uji yang ada kedalam model *confusion matrix*.

Setelah data diuji dimasukan kedalam *confusion matrix*, hitung nilai-nilai yang telah dimasukan tersebut untuk dihitung jumlah *sensitivity, specificity, precison* dan *accuracy*. *Sensitivity* digunakan untuk membandingkan jumlah *true positives* terhadap jumlah tupel yang *positives* sedangkan *specificity* adalah perbandinagn jumlah *true negatives* terhadap jumlah tupel yang *negatives*.

Terlihat pada gambar di bawah ini yang menunjukan niali *accuracy*, *recall*, dan *precision* yang dihasilkan oleh Rapid Miner menggunakan model *confusion matrix:*

C) Hasil Pengujian

Berdasarkan hasil pengujian sistem yang telah dilakukan di Universitas Dehasen Bengkulu yang dilakukan penulis dapat diketahui bahwa terdapat kelebihan dan kekurangan dari sistem lama dan sistem baru tersebut.

Dengan menggunakan sistem lama, masih menggunakan sistem perkiraan saja sehingga tingkat kesalahan untuk memprediksi tingkat kelulusan mahasiswa tepat waktu atau tidak masih besar. Sedangkan dengan menggunakan teknik data mining ini tingkat kesalahan dalam memprediksi masa studi mahasiswa tersebut lulus tepat waktu atau tidak dapat dikurangi dengan tingkat kesalahan 5 % .

Dengan menggunakan sistem lama, pengambilan keputusannya masih kompleks dan global. Sedangkan setelah menggunakan sistem baru daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapatndiubah menjadi lebih simpel dan spesifik.

Dengan menggunakan sistem lama, seorang penguji masih kesulitan dalam menganalisis untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Sedangkan dengan sistem baru, dalam analisis, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Metode pohon keputusan dapat menghindari munculnya permasalahan ini dengan menggunakan criteria yang jumlahnya lebih sedikit pada setiap node internal tanpa banyak mengurangi kualitas keputusan yang dihasilkan.

Dengan menggunakan sistem lama, membutuhkan waktu lama dalam penentuan keputusan. Sedangkan dengan sistem baru waktu yang diperlukan dalam penentuan keputusan lebih cepat dengan menggunakan teknik data mining tersebut.

V. PENUTUP

Dalam penelitian ini yang berjudul "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algotitma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu), penulis menerapkan algoritma C4.5 dengan menggunakan sepuluh parameter yaitu NPM, Nama Mahasiswa, Semester, Prodi, Jenjang Pendidikan, Jenis Kelamin, IPK, dan Jumlah SKS. Kemudian untuk pengolahan data penulis menggunakan aplikasi pada Rapid Miner 5. Dari hasil penelitian terbukti bahwa algoritma C4.5 lebih akurat dibandingkan analisa yang dilakukan oleh analis mahasiswa. Hal ini dibuktikan dengan hasil evaluasi penelitian bahwa algoritma C4.5 mampu menganalisa tingkat ketepatan waktu mahasiswa menyelesaikan masa studinya.

Walaupun model algoritma C4.5 sudah diterapkan dan berjalan dengan baik di dalam sistem, namun ada hal yang harus ditambahkan untuk menambah akurasi algoritma C4.5, yaitu:

1) Melakukan *pruning* terhadap algoritma C4.5 sehingga pohon yang terbentuk tidak terlalu besar bahkan mungkin untuk jumlah data yang besar sekalipun. Ini dilakukan untuk mengefisienkan

kinerja dari algoritma C4.5 tanpa mengurangi keakuratannya.
2) Pada riset selanjutnya dapat digunakan metode seleksi atribut yang lain seperti Chi-Square untuk ketepatan penyeleksian atribut.
3) Menerapkan algoritma C4.5 kedalam data yang lebih besar untuk menguji akurasi algoritma

DAFTAR PUSTAKA

Firmansyah, (2011). *Penerapan Algoritma Klasifikasi C4.5 Untuk Penentuan Kelayakan Pemberian Kredit Koperasi*. Jakarta. 56 Halaman

Han & Kamber, (2006). *Data Mining Concept and Thenique.* Morgan Kaufffman.San Fransisco.

Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining.* Yogyakarta: Andi Publishing.

Prasetyo, Eko, (2012). *Data Mining*, Andi Yogyakarta, 356 Halaman.

Romi, Satriyono, (2012). Rapid Miner. Udinus. Jakarta.

Sugiyanto. (2010). *Metode Penelitian.* Intan Permata. Jogjakarta.

Sutanta, Edhy, 2005, *Pengantar Teknologi Informasi*, Graha Ilmu.Yogyakarta,. 611 halaman.

Supriyanto, (2011). *Pengenalan Komputer*. Graha Ilmu. Yogyakarta