KONVERSI DATA TRAINING TENTANG PEMILIHAN KELAS MENJADI BENTUK POHON KEPUTUSAN DENGAN TEKNIK KLASIFIKASI

Ni Luh Ratniasih
STMIK STIKOM Bali
Jl.Raya Puputan No. 86 Renon, Denpasar-Bali, Telp (0361)244445
e-mail: ratni@stikom-bali.ac.id

Abstrak

Penyajian data untuk menghasilkan nilai informasi sering kali ditampilkan dalam bentuk tabulasi. Apabila data yang ditampilkan memiliki kapasitas kecil, mungkin tidak terlalu sulit untuk mencerna kandungan informasi tersebut. Tetapi apabila data yang disajikan memiliki kapasitas yang sangat besar, dikawatirkan adanya kendala untuk menyerap informasi secara tepat dan cepat. Hal ini dikarenakan bahwa dibutuhkan waktu yang cukup lama untuk membaca data yang ditampilkan secara rinci hingga akhir data. Data yang akan dibahas dalam penelitian ini adalah data calon mahasiswa STMIK STIKOM Bali. Selama ini STMIK STIKOM Bali belum mampu melakukan klasifikasi pemilihan kelas yang sesuai untuk dapat direkomdasikan kepada mahasiswa baru, sehingga dari data historis yang ditampilkan akan dikonversi menjadi bentuk pohon keputusan. Variabel yang digunakan untuk klasifikasi adalah jurusan di SMA/K, status pekerjaan, pekerjaan orang tua, umur, status calon mahasiswa, dan status pernikahan. Dengan demikian penyerapan informasi akan menjadi lebih mudah untuk dilakukan. Penelitian ini mengimplementasikan disiplin ilmu data mining menggunakan teknik klasifikasi pohon keputusan serta diaplikasikan dengan tools Rapid Miner 4.1

Kata kunci: *Data Mining, Kelas, Klasifikasi*

Abstract

Presentation of data to generate the value of information is often presented in the form of tabulation. If the data displayed has a small capacity, it may not be too difficult to digest the information content. But if the data presented have a very large capacity, feared constraints to absorb information accurately and quickly. This is because that it takes a long time to read the data that is displayed in detail until the end of the data. The data will be discussed in this research is data STMIK STIKOM Bali prospective students. During STMIK STIKOM Bali has not been able to classify the corresponding class election to direkomdasikan to new students, so that the display of historical data will be converted into the form of a decision tree. Variable used for classification is majoring in high school, employment status, occupation of parents, age, status of students, and marital status. Thus the absorption of information would be easier to do. This study implements the discipline of data mining using decision tree classification techniques and tools applied to Rapid Miner 4.1

Keywords*: Data Mining, Class, Classification*

1. Pendahuluan

Teknologi komputasi dan media penyimpanan telah memungkinkan manusia untuk mengumpulkan dan menyimpan data dari berbagai sumber dengan jangkauan yang amat luas. Meskipun teknologi basis data modern telah menghasilkan media penyimpanan yang besar, teknologi untuk membantu menganalisis, memahami, atau bahkan memvisualisasikan data belum banyak tersedia. Hal inilah yang melatarbelakangi dikembangkannya konsep *data mining.*

STMIK STIKOM Bali merupakan perguruan tinggi pertama bidang komputer di Pulau Bali. Sebagai perguruan tinggi pertama, STMIK STIKOM Bali menjadi perguruan tinggi pavorit bidang komputer. STMIK STIKOM Bali sukses menarik banyak mahasiswa setiap periode dimana kenaikan persentase jumlah mahasiswa baru yang terdaftar mencapai rata-rata 20% - 30% setiap tahunnya.

STMIK STIKOM Bali menyediakan dua jenis kelas yang dapat dipilih oleh calon mahasiswa baru diantaranya kelas regular dan kelas karyawan. Pada saat melakukan pendaftaran atau pencarian informasi calon mahasiswa akan diminta untuk menentukan kelas atau akan dianjurkan oleh pihak STMIK STIKOM Bali dalam menentukan kelas yang akan dipilih. Namun akan muncul suatu masalah pada saat calon mahasiswa bersangkutan belum dapat menentukan kelas yang dipilih serta pihak STMIK STIKOM Bali salah dalam mengarahkan calon mahasiswa.

Oleh karena itu untuk meningkatkan kualitas mahasiswa dalam pemilihan kelas pada perguruan tinggi STMIK STIKOM Bali, maka diperlukan suatu *rule* atau aturan klasifikasi yang dapat digunakan dalam menentukan dan memilih kelas yang cocok untuk calon mahasiswa. Metode yang tepat digunakan klasifikasi pemilihan kelas adalah pohon keputusan. Pohon keputusan merupakan salah satu metode yang tepat untuk membentuk pola-pola yang mungkin memberikan indikasi yang bermanfaat pada data mahasiswa yang dalam jumlah besar. Tujuan dari penelitian ini adalah untuk menentukan model dari training set yang membedakan *record* kedalam kategori atau kelas yang sesuai, model tersebut yang akan digunakan untuk mengklasifikasikan *record* yang kelasnya belum diketahui sebelumnya pada test set.

2. Landasan Teori

2.1 Definisi *Data Mining*

Menurut Gartner Group, *data mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [1]. *Data mining* bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dulu. Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu – ilmu lain, seperti database system, data warehousing, statistik, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing . Data mining didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar. Berawal dari beberapa disiplin ilmu, *data mining* bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani:

1. Jumlah data yang sangat besar
2. Dimensi data yang tinggi
3. Data yang heterogen dan berbeda bersifat

Data mining merupakan proses pencarian pola dan relasi-relasi yang tersembunyi dalam sejumlah data yang besar dengan tujuan untuk melakukan klasifikasi, estimasi, prediksi, association rule, clustering, deskripsi dan visualisasi [2]. Secara garis besar data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu [3]:

1. Descriptive mining, yaitu proses untuk menemukan karakteristik penting dari data dalam suatu basis data. Teknik data mining yang termasuk dalam descriptive mining adalah clustering, association, dan sequential mining.
2. Predictive, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu teknik yang terdapat dalam predictive mining adalah klasifikasi.

Berikut adalah tahapan dalam data mining

1. Pembersihan data (data cleaning)
   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa data mining yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.
2. Integrasi data (data integration)
   Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Tidak jarang data yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi

juga berasal dari beberapa database atau file teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. Seleksi Data (Data Selection)
   Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.
4. Transformasi data (Data Transformation)
   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.
5. Proses mining,
   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.
6. Evaluasi pola (pattern evaluation),
   Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses data mining, mencoba metode data mining lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat.
7. Presentasi pengetahuan (knowledge presentation),
   Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami data mining. Karenanya presentasi hasil data mining dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses data mining. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil data mining.

2.2 Decision Tree (Pohon Keputusan)

Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh 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. Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk mem-break down proses pengambilan keputusan yang kompleks menjadi lebih simpel sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan. Pohon Keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi. Decision tree (pohon keputusan) adalah sebuah diagram alir yang mirip dengan struktur pohon, dimana setiap internal node menotasikan atribut yang diuji, setiap cabangnya merepresentasikan hasil dari atribut tes tersebut, dan leaf node merepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas [2].

Klasifier pohon keputusan merupakan teknik klasifikasi yang sederhana yang banyak digunakan. Bagian ini membahas bagaimana pohon keputusan bekerja dan bagaimana pohon keputusan dibangun. Seringkali untuk mengklasifikasikan obyek, kita ajukan urutan pertanyaan sebelum bisa kita tentukan kelompoknya.

Walaupun banyak variasi model decision tree dengan tingkat kemampuan dan syarat yang berbeda, pada umumnya beberapa ciri kasus cocok untuk diterapkan decision tree [4] :

1. Data dinyatakan dengan pasangan atribut dan nilainya. Misalnya atribut satu data adalah temperatur dan nilainya adalah dingin. Biasanya untuk satu data nilai dari satu atribut tidak terlalu banyak

jenisnya. Dalam contoh atribut warna buah ada beberapa nilai yang mungkin yaitu hijau, kuning, merah.
2. Label/output data biasanya bernilai diskrit. Output ini bisa bernilai ya atau tidak, sakit atau tidak sakit, diterima atau ditolak. Dalam beberapa kasus mungkin saja outputnya tidak hanya dua kelas, tetapi penerapan decision tree lebih banyak untuk kasus binary.
3. Data mempunyai missing value. Misalkan untuk beberapa data, nilai dari suatu atributnya tidak diketahui. Dalam keadaan seperti ini decision tree masih mampu memberi solusi yang baik.

Kelebihan dari metode pohon keputusan adalah:

1 Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik.
2 Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sample diuji hanya berdasarkan kriteria atau kelas tertentu.
3 Fleksibel untuk memilih fitur dari internal node yang berbeda, fitur yang terpilih akan membedakan suatu kriteria dibandingkan kriteria yang lain dalam node yang sama. Kefleksibelan metode pohon keputusan ini meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang lebih konvensional
4 Dalam analisis multivariat, 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.

Kekurangan Pohon Keputusan

1 Terjadi overlap terutama ketika kelas-kelas dan criteria yang digunakan jumlahnya sangat banyak. Hal tersebut juga dapat menyebabkan meningkatnya waktu pengambilan keputusan dan jumlah memori yang diperlukan.
2 Pengakumulasian jumlah eror dari setiap tingkat dalam sebuah pohon keputusan yang besar.
3 Kesulitan dalam mendesain pohon keputusan yang optimal.
4 Hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat tergantung pada bagaimana pohon tersebut didesain.

Kefleksibelan membuat metode ini atraktif, khususnya karena memberikan keuntungan berupa visualisasi saran (dalam bentuk pohon keputusan) yang membuat prosedur prediksinya dapat diamati. Decision tree banyak digunakan untuk menyelesaikan kasus penentuan keputusan seperti dibidang kedokteran (diagnosis penyakit pasien), ilmu komputer (struktur data), psikologi (teori pengambilan keputusan), dan sebagainya.

Karakteristik dari decision tree dibentuk dari sejumlah elemen sebagai berikut [2]:

1 Node akar, tidak mempunyai lengan masukan dan mempunyai nol atau lebih lengan keluaran
2 Node internal, setiap node yang bukan daun (non-terminal) yang mempunyai tepat satu lengan masukan dan dua atau lebih lengan keluaran. Node ini menyatakan pengujian yang didasarkan pada nilai fitur.
3 Lengan, setiap cabang menyatakan nilai hasil pengujian di node bukan daun
4 Node daun (terminal), node yang mempunyai tepat satu lengan masukan dan tidak mempunyai lengan keluaran. Node ini menyatakan label kelas (keputusan)

Decision tree mempunyai tiga pendekatan klasik:

1. Pohon klasifikasi, digunakan untuk melakukan prediksi ketika ada data baru yang belum diletahui label kelasnya. Pendekatan ini yang paling banyak digunakan.
2. Pohon regresi, ketika hasil prediksi dianggap sebagai nilai nyata yang mungkin akan didapatkan. Misalnya, kasus hanya minyak, kenaikan harga rumah, prediksi inflasi tiap tahun, dan sebagainya.
3. CART (atau C&RT), ketika masalah klasifikasi dan regresi digunakan bersama-sama.

Jika memperhatikan kriteria pemilihan cabang pemecah, maka algoritma penginduksi pohon keputusan mempunyai beberapa macam:

1 GINI (*impurity*) index
2 entropy (*impurity*)

3 misclassification
4 Chi-square
5 G-square

Dapat dilihat bahwa menggunakan pohon keputusan sebagai support tool dalam menganalis suatu masalah pengambilan keputusan dapat sangat membantu kita dalam melakukan pengambilan keputusan. Kegunaan pohon keputusan yang dapat melihat berbagai macam alternatif keputusan-keputusan yang dapat kita ambil serta mampu memperhitungkan nilai-nilai dari faktor-faktor yang mempengaruhi alternatif-alternatif keputusan tersebut adalah sangat penting dan berguna, karena membuat kita dapat mengetahui alternatif mana yang paling menguntungkan untuk kita ambil.

Pohon keputusan juga dapat dipergunakan untuk memperhitungkan dan melakukan analisa terhadap resiko-resiko yang mungkin muncul dalam suatu alternatif pemilihan keputusan. Selain itu, pohon keputusan juga dapat dipakai untuk memperhitungkan berapa nilai suatu informasi tambahan yang mungkin kita perlukan agar kita dapat lebih mampu dalam membuat suatu pilihan keputusan dari suatu alternatif-alternatif keputusan yang ada.

Dengan melihat kegunaan pohon keputusan dan kemampuannya dalam memperhitungkan berbagai alternatif pemecahan masalah termasuk faktor-faktor yang mempengaruhinya serta nilai resiko dan nilai informasi dalam alternatif keputusan itu, maka jelaslah bahwa pohon keputusan ini dapat menjadi alat bantu yang sangat berguna dalam pengambilan keputusan.

2.3 Algoritma C4.5

Algoritma C4.5 adalah salah satu algoritma untuk mengubah fakta yang besar menjadi pohon keputusan (*decision tree*) yang merepresentasikan aturan (*rule*). Tujuan dari pembentukan pohon keputusan dalam algoritma C4.5 adalah untuk mempermudah dalam penyelesaian permasalahan.

Dalam menggunakan algoritma C4.5 terdapat beberapa tahapan yang umum yaitu pertama mengubah bentuk data dalam tabel menjadi model pohon kemudian mengubah model pohon menjadi aturan (*rule*) dan terakhir menyederhanakan rule [5].

Secara umum, algoritma C4.5 untuk membangun sebuah pohon keputusan adalah sebagai berikut:

a. Hitung jumlah data, jumlah data berdasarkan anggota atribut hasil dengan syarat tertentu. Untuk proses pertama syaratnya masih kosong.
b. Pilih atribut sebagai Node.
c. Buat cabang untuk tiap-tiap anggota dari Node.
d. Periksa apakah nilai entropy dari anggota Node ada yang bernilai nol. Jika ada, tentukan daun yang terbentuk. Jika seluruh nilai entropy anggota Node adalah nol, maka proses pun berhenti.
e. Jika ada anggota Node yang memiliki nilai entropy lebih besar dari nol, ulangi lagi proses dari awal dengan Node sebagai syarat sampai semua anggota dari Node bernilai nol.

Node adalah atribut yang mempunyai nilai gain tertinggi dari atribut-aribut yang ada. Untuk menghitung nilai gain suatu atribut digunakan rumus seperti yang tertera dalam persamaan berikut:

Berikut adalah bentuk umum dari C4.5 :

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

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

penghitungan nilai entropi dapat dilihat pada persamaan berikut (kusrini, Emha Taufiq Lut :2009).

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

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

2.4 Rapid Miner

Rapid miner adalah aplikasi data mining yang berbasis *open source*. *Open source* rapid miner berlisensi AGPL (*GNU Affero General Public License*) versi 3. Penelitian mengenai *tools* ini dimulai sejak tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund yang kemudian diambil alih oleh SourceForge sejak tahun 2004. Rapid miner memperoleh peringkat satu sebagai *tools* data mining untuk proyek nyata pada poll oleh KDnuggets, sebuah koran data-mining pada 2010-2011.

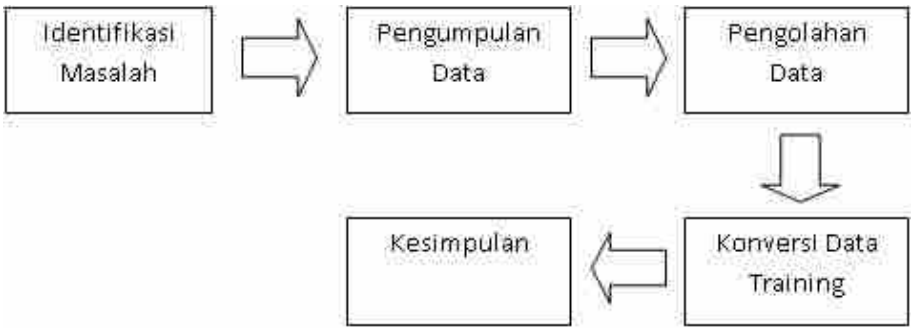
Dalam penerapannya, rapid miner menyediakan prosedur *data mining* dan *machine* learning termasuk : ETL (*extraction, transformation, loading*), data preprocessing, visualisasi, modelling dan evaluasi. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI. *Tools* rapid miner ditulis dalam bahasar pemrograman Java dan juga mengintegrasikan proyek data mining Weka dan statistika [6].

Beberapa solusi yang diusung oleh rapid miner antara lain :

a. Integrasi data, Analitis ETL, Data Analisis, dan pelaporan dalam satu suite tunggal.
b. Powerfull tetapi memiliki antarmuka pengguna grafis yang intuitif untuk desain anakisis proses.
c. Repositori untuk prose, data dan penanganan meta data.
d. Hanya solusi dengan transformasi meta data: lupakan trail and arror dan memeriksa hasil yang telah diinspeksi selama desain.
e. Hanya solusi yang mendukung *on-the-fly* kesalahan dan dapat melakukan perbaikan dengan cepat. Lengkap dan fleksibel: ratusan loading data, transformasi data, pemodelan data dan metode visualisasi data.

3. Metode Penelitian

Metode penelitian yang akan dilakukan dalam penelitian ini dapat digambarkan pada gambar 1 berikut ini :



Gambar 1. Tahap Penelitian

a. Tahap Identifikasi Masalah

Tahap identifikasi permasalahan adalah tahap awal dari penelitian ini. Permasalahan yang dicoba untuk diselesaikan yaitu: bagaimana mengkonversi data training tentang pemilihan kelas menjadi bentuk pohon keputusan dengan teknik klasifikasi.

b. Tahap Pengumpulan Data

Pada tahapan ini akan dilakukan pengumpulan data-data yang diperlukan untuk mendukung penelitian ini. Hal yang paling terpenting dalam penyelesaian kasus pohon keputusan adalah ketersediaan data training atau data histori. Data yang dikumpulkan antara lain data mahasiswa yang dapat menentukan tingkat pemilihan kelas. Data yang digunakan adalah data sekunder yang didapat dari database sistem informasi penerimaan mahasiswa baru (PMB) STMIK STIKOM Bali.

c. Tahap Pengolahan Data

Pada tahap pengolahan data ini akan dilakukan pre-processing. Data yang diperoleh dari database penerimaan mahasiswa baru (PMB) STMIK STIKOM Bali sebagai bahan awal penelitian ini harus dilakukan tiga tahapan. Tahapan pertama yaitu menghapus beberapa data yang dianggap dapat menimbukan noise. Tahapan kedua yaitu merubah beberapa nilai variabel menjadi bentuk simbol untuk

mempermudah proses pengolahan data. Tahapan ketiga yaitu membagi data menjadi data training dan testing. Dimana data training digunakan untuk membangun rule sedangkan data testing digunakan untuk menilai performansi dari rule yang telah dibangun.

d. Tahap Konversi Data Training

Pada tahapan ini akan dilakukan proses konversi data training yang diperolah dari hasil tahap pengolahan data dengan menggunakan tools rapid miner sehingga dihasilkan suatu rule (aturan) penentuan kelas calon mahasiswa.

e. Tahap Penarikan Kesimpulan dan Saran

Tahapan penarikan kesimpulan merupakan tahap akhir dari penelitian ini. Pada tahap ini akan dapat ditarik kesimpulan untuk menjawab tujuan penelitian. Saran-saran perbaikan juga akan diberikan pada tahap ini yang berguna bagi penelitian ke depannya.

4. Hasil dan Pembahasan

3.1. Analisis Data *Record* dan Variabel/Atribut

Data yang digunakan dalam penelitian ini adalah data mahasiswa STMIK STIKOM Bali angkatan 2013/2014 semester ganjil yang diperoleh dari *database* sistem informasi penerimaan mahasiswa baru (PMB). Data mahasiswa yang terdapat pada *database* akan dilakukan seleksi data dan pembersihan data. Proses pembersihan data mencakup antara lain membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data seperti kesalahan cetak. Jika terdapat *record* tertentu pada salah satu atribut kehilangan nilai maka *record* tersebut akan dihapus karena *record* tersebut dinilai *missing value* (kehilangan data). Terdapat beberapa cara untuk mengatasi *missing value* [4] yaitu jika jumlah datanya besar, maka pengamatan dengan *missing value* dapat diabaikan atau dihapus. Jika jumlah pengamatan terbatas atau kecil, maka missing value bisa diganti dengan nilai rata-rata dari variabel atau atribut bersangkutan.

Dalam data mahasiswa yang diperoleh terdapat missing value, sehingga record yang kehilangan data atau missing value dapat diabaikan. Tahap seleksi ini disebut juga dengan tahap perbersihan data (data cleaning) yang bertujuan mendapatkan data yang bersih, sehingga data tersebut dapat digunakan untuk tahap selanjutnya yaitu seleksi data sample. Pembersihan data juga memeriksa data yang tidak konsisten. Data mahasiswa yang diambil dari database adalah data yang konsisten dan relevan. Dalam penelitian ini data yang digunakan dalam proses data mining hanya data sample sejumlah 50 record data.

3.2 Data

Data yang digunakan dalam proses data mining adalah data yang telah dianalisa sebelumnya dan data yang dipilih sebagai data sample. Akan dilakukan transformasi data pada data yang telah dipilih sehingga sesuai dengan proses data mining. Data ditansformasi dari database ke dalam bentuk MS. Excel karena data akan digunakan dalam tool rapid miner.

3.3 Pengujian Data dan Hasil

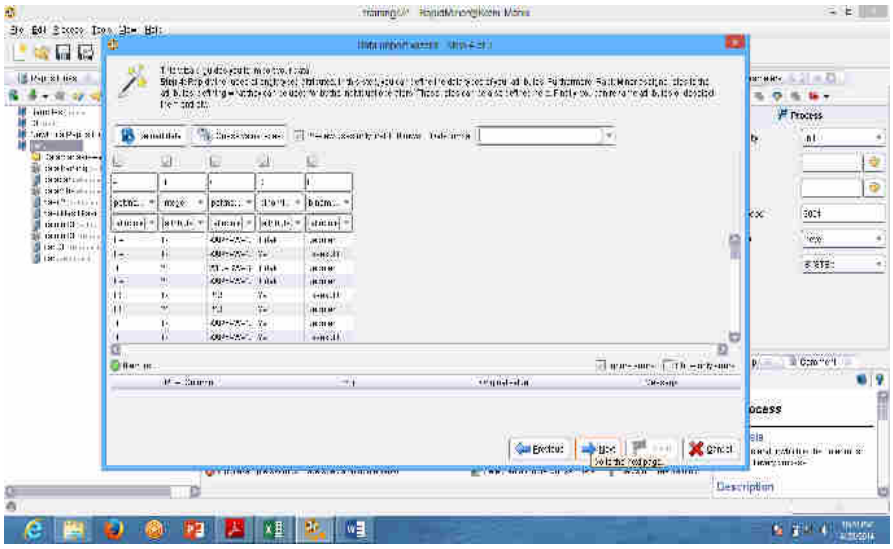
Pada tahap ini dilakukan pengujian data *training* sesuai tujuan penelitian yaitu untuk menerapkan teknik klasifikasi menggunakan metode *decision tree* yaitu dengan konsep algoritma C4.5 dan *tool* rapid miner. Dari data *training* akan dibentuk suatu model pohon yang nanti akan menghasilkan sejumlah aturan dalam pohon tersebut. Model pohon akan terbentuk dengan menggunakan *tool* rapid miner. Berikut ini adalah proses pengolahan data penggunakan *tool* rapid miner :

a. *Import* data dari MS. Excel. Data *training* yang telah diditransformasi akan di*import* dari *tool* rapid miner.
b. Pemilihan data *training*. Data yang telah di*import* dari MS. Excel didipilih area yang akan digunakan sebagai data *training*. Pada gambar 1 adalah gambar pada saat data *training* dipilih.

Gambar 2. Gambar Pemilihan Data Training

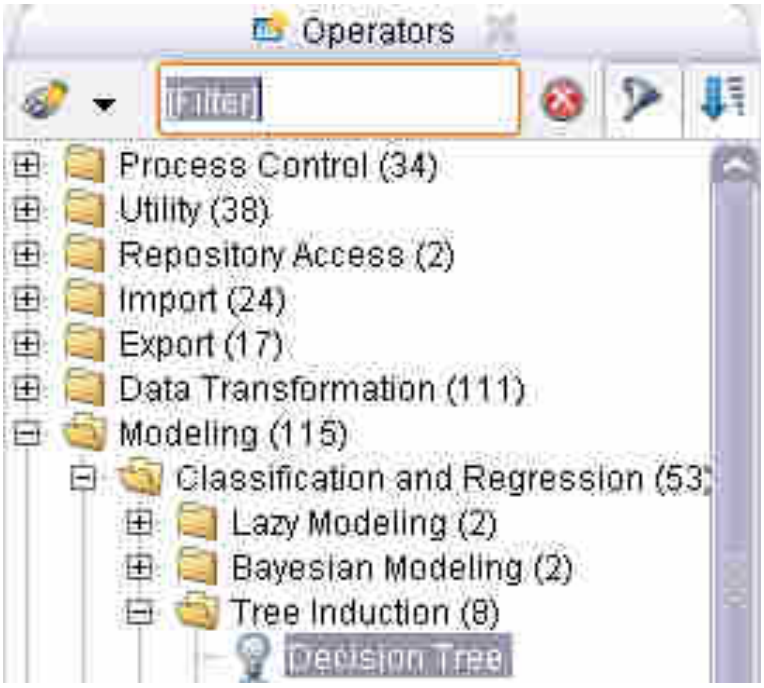
c. Menentukan label dan tipe data. Pada tahap ini setiap data harus ditentukan label serta tipe datanya. Untuk menentukan tipe label dan atribut harus disesuaikan dengan ketentuan berikut :
   - Polynom = tipe data ini untuk karakter baik angka ataupun huruf (sama seperti varchar/text)
   - Binominom = tipe data ini untuk 2 kategori (Y/T, L,P, Besar/Kecil, dll)
   - Atribut = digunakan sebagai variable predictor/prediksi
   - Label = digunakan sebagai variable tujuan

Pada gambar 3 adalah gambar pada saat menentukan label dan tipe dari data *training.*



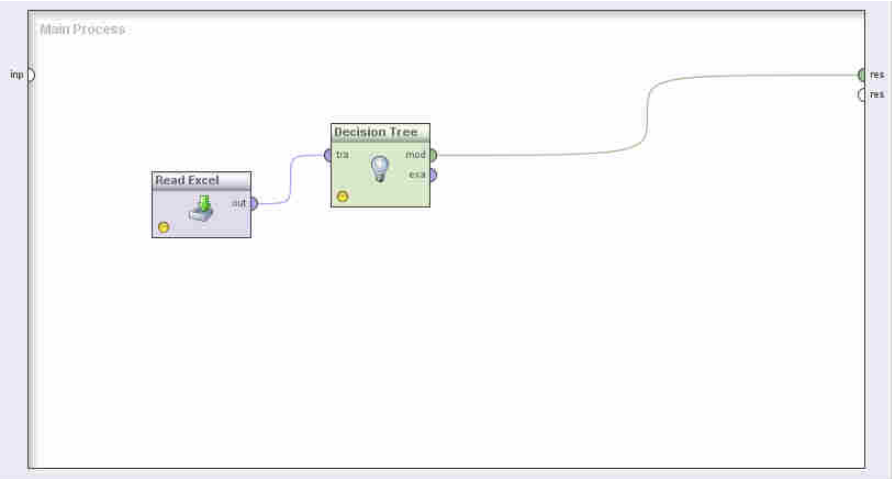
Gambar 3. Penentuan Label dan Tipe Data

d. Ambil data hasil import. Setelah data selesai diimport, akan dilakukan proses pengambilan data melalui wizard di tab operator, ketik *decision tree* pada bagian filter. Pada gambar 4 adalah gambar pengambilan data hasil *import.*



Gambar 4. Pengambilan Data Hasil Import

e. *Main Process*. Pada proses ini akan dilakukan proses *decision tree* dari data yang telah diimport dengan menyeret "*decision tree*" ke *main process*, kemudian klik and drag "out" (pada Read Excel dihubungkan ke "tra"(pada *decision tree*). Kemudian klik and drag "mod" (pada *decision tree*) dihubungkan ke "res" disebelah kanan *main process*. Pada gambar 4 adalah gambar *main proses* yang telah dihubungkan. Setelah "Read Excel" dan "Decision Tree" dibungkan kemudian klik run maka akan tampil *decision tree*.



Gambar 5. Main Proses

Setelah dilakukan pengujian dengan menggunakan konsep algoritma C4.5 dengan *tool* rapid miner ternyata tidak menghasilkan bentuk pohon keputusan dan aturan-aturan klasifikasi. Dari hasil analisis ditemukan bahwa adanya ketidaksesuaian pemilihan komponen variable yang dapat menentukan klasifikasi kelas untuk calon mahasiswa. Hal ini disebabkan adanya variable jurusan SMA/K yang tidak mempengaruhi pemilihan kelas. Pada saat variable jurusan SMA/K diganti dengan status perkawinan, baru menghasilkan bentuk pohon keputusan dan aturan-aturan klasifikasi.

4. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, maka dapat disimpulkan adalah algoritma data mining decision tree dapatdigunakan untuk mengklasifikasikan kelas yang sesuai untuk mahasiswa baru namun hal tersebut ditentukan oleh variable yang digunakan sebagai data training. Pada penelitan ini komponen variable dan pemilihan data *training* kurang tepat sehingga pada hasil analisis tidak membentuk pohon keputusan dan menghasilkan aturan (*rule*) dalam pohon keputusan sehingga klasifikasi tidak berhasil dilakukan.

Dapat dilihat bahwa menggunakan pohon keputusan sebagai support tool dalam menganalis suatu masalah pengambilan keputusan dapat sangat membantu kita dalam melakukan pengambilan keputusan. Kegunaan pohon keputusan yang dapat melihat berbagai macam alternatif keputusan-keputusan yang dapat kita ambil serta mampu memperhitungkan nilai-nilai dari faktor-faktor yang mempengaruhi alternatif-alternatif keputusan tersebut adalah sangat penting dan berguna, karena membuat kita dapat mengetahui alternatif mana yang paling menguntungkan untuk kita ambil.

Pohon keputusan juga dapat dipergunakan untuk memperhitungkan dan melakukan analisa terhadap resiko-resiko yang mungkin muncul dalam suatu alternatif pemilihan keputusan. Selain itu, pohon keputusan juga dapat dipakai untuk memperhitungkan berapa nilai suatu informasi tambahan yang mungkin kita perlukan agar kita dapat lebih mampu dalam membuat suatu pilihan keputusan dari suatu alternatif-alternatif keputusan yang ada.

Dengan melihat kegunaan pohon keputusan dan kemampuannya dalam memperhitungkan berbagai alternatif pemecahan masalah termasuk faktor-faktor yang mempengaruhinya serta nilai resiko dan nilai informasi dalam alternatif keputusan itu, maka jelaslah bahwa pohon keputusan ini dapat menjadi alat bantu yang sangat berguna dalam pengambilan keputusan.

References

[1] Larose, D.T, 2006. *Discovering Knowledge in Data: An Introduction to Data mining*. John Willey & Sons, Inc.
[2] Han J, Kamber M. 2001. *Data Mining : Concepts and Techniques*. Simon Fraser University, Morgan Kaufmann Publishers.
[3] Tan S, Kumar P, Steinbach M. 2005. Introduction To Data Mining. Addison Wesley.
[4] Santosa, Budi. 2007. Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis, Teori dan Aplikasi. Graha Ilmu Yogyakarta.
[5] Basuki, Achmad dan Syarif, Iwan. 2003. Modul Ajar Decision Tree. Surabaya : PENS-ITS.
[6] Rapid-I GmbH. (2008). *Rapidminer-4.2-tutorial.* Germany: Rapid-I.