OPTIMASI DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN C4.5 UNTUK KLASIFIKASI KELULUSAN MAHASISWA

Ni Luh Ratniasih
Program Studi Sistem Informasi
STMIK STIKOM Bali
ratni@stikom-bali.ac.id

ABSTRACT

Presentation of data to produce information values is often displayed in the form of tabulations. If the data displayed has a small capacity, it may not be difficult to process the information. But if the data presented has a very large capacity, it is feared there are obstacles to absorbing information accurately and quickly. This is because that it takes a long time to read the data displayed in detail until the end of the data. The data to be discussed in this study are data of STMIK STIKOM Bali students. Historical data displayed will be converted into a decision tree. Thus the absorption of information will become easier. This research implements data mining disciplines using the naïve bayes method comparison with C4.5 algorithm which is a method for performing classification techniques and applied with Rapid Miner tools.

***Keywords** : C4.5, KNN, Student Graduation*

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 mahasiswa STMIK STIKOM Bali. Data historis yang ditampilkan akan dikonversi menjadi bentuk pohon keputusan. Dengan demikian penyerapan informasi akan menjadi lebih mudah. Penelitian ini mengimplementasikan disiplin ilmu data mining menggunakan komparasi metode naïve bayes dengan algoritma C4.5 yang merupakan sebuah metode untuk melakukan teknik klasifikasi serta diaplikasikan dengan tools Rapid Miner.

Kata kunci : C4.5, KNN, Kelulusan Mahasiswa

PENDAHULUAN

Perkembangan teknologi informasi memberikan pengaruh yang sangat besar terhadap kehidupan manusia. Hal ini terlihat dari penggunaan komputer di segala bidang dalam aktivitas manusia, baik dalam bidang pendidikan, organisasi dan masyarakat umum. Penggunaan komputer dalam bidang pendidikan menghasilkan data yang melimpah terkait peserta didik serta hasil proses pembelajaran. Pada perguruan tinggi akan menghasilkan data mahasiswa berupa nilai IPK ser-ta jumlah lulusan yang terus bertambah setiap tahunnya. Dibalik data yang melimpah terdapat informasi baru yang tersebunyi. Infor-masi baru diperoleh dari sebuah pengolahan data sehingga dapat dimanfaatkan kembali.

Jumlah data yang terus meningkat ini memerlukan beberapa metode untuk mengolah dan mengambil kesimpulandan informasi dari data tersebut. Beberapa metode yang digunakan untuk mengolah data yang sifatnya besar untuk menemukan pola yang terdapat didalamnya diantaranya adalah : teknik klastering, analisis diskriminan, teorema bayes, *decision tree artificial neural networks, sup-port vector machine, regresi linear, support vector*

regresi. Setiap metode tersebut memiliki algoritma-algoritma yang digunakan untuk memproses data yang ada. Namun dalam penelitian ini akan dilakukan komparasi hasil klasifikasi dari dua buah metode untuk konversi data *training* kelulusan tepat waktu mahasiswa STMIK STIKOM Bali menggunakan metode naïve bayes dan algoritma C4.5

LANDASAN TEORI

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.

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

Algoritma Naïve Bayes

Algoritma ini menggunakan metode probabilias dan statistik yang dikemukakan oleh seorang ilmuan Inggris Thomas Bayes yaitu memprediksi probabilitas dimasa depan berdasarkan pengalaman dimasa sebelumnya. Serta alasan menggunakan metode ini adalah metode *naïve bayes* ini memiliki kelebihan sebagai berikut :

1. *Bayesian filter* memiliki komputasi yang mudah.
2. *Bayesian* memeriksa data secara keseluruhan yaitu memeriksa token di database spam maupun legitimate.
3. *Bayesian filtering* termasuk dalam *supervised learning* yaitu secara otomatis akan melakukan proses *learning* dari data yang masuk.
4. *Bayesian filtering* cocok diterapkan di level aplikasi *client*/individual *user*.
5. *Bayesian filtering* cocok diterapkan pada *binary class* yaitu klasifikasi ke dalam dua kelas.
6. Metode ini multilingual dan internasional. *Bayesian filtering* melakukan *generate token* dengan pengenalan karakter sehingga mampu diimplementasikan pada bahasa apapun.

Klasifikasi–klasifikasi Bayes adalah klasifikasi statistik yang dapat memprediksi kelas suatu anggota probabilitas. Untuk klasifikasi Bayes sederhana yang lebih dikenal sebagai *Naïve Bayesian Classifier* dapat diasumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut–atribut lain. Asumsi ini disebut *class conditional independence* yang dibuat untuk memudahkan perhitungan-perhitungan pengertian ini dianggap *"naive",* dalam bahasa lebih sederhana *naïve* itu mengasumsikan bahwa kemunculan suatu *term* kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata-kata yang lain dalam kalimat padahal dalam kenyataannya bahwa kemungkinan kata dalam kalimat sangat dipengaruhi kemungkinan keberadaan kata-kata yang dalam kalimat. Dalam *Naïve Bayes* diasumsikan prediksi atribut adalah tidak tergantung pada kelas atau tidak dipengaruhi atribut laten.

Naïve bayes merupakan algoritma yang termasuk ke dalam *supervised learning*, maka dibutuhkan pengetahuan awal untuk mengambil keputusan. Langkah–langkah awalnya adalah

a. Step 1 : menghitung jumlah kategori setiap variabel pada setiap training
b. Step 2 : hitung probabilitas pada setiap kategori
c. Step 3 : tentukan frekuensi setiap kata pada setiap kategori

Pengklasifikasian :

a. Step 1 : hitung untuk setiap kategori 2.
b. Step 2 : tentukan kategori dengan nilai maksimal

Rumus probabilitas adalah

P (H | X) = P (X | H) P (H) / P (X)

Dalam hal :

X = data dengan class yang belum diketahui.

H = hipotesis data X merupakan suatu class spesifik

P(H|X) = probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)

P(H) = probabilitas hipotesis H (prior probability)

P(X|H) = probabilitas X berdasar kondisi pada hipotesis H

P(X) = probabilitas dari X

***Decision Tree* (Pohon Keputusan)**

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* [3]:

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 *output*-nya 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.

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 terkhir menyederhanakan rule [4].

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 []

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

Keterangan:

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

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 [5].

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.

PEMBAHASAN

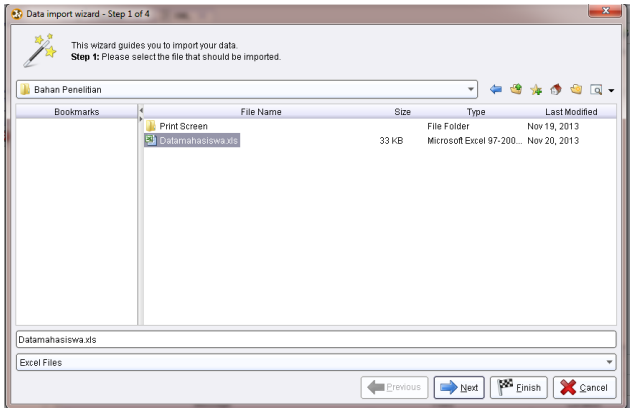
Pengolahan Data dengan Naïve Bayes

Import data dari MS. Excel. Data *training* yang telah ditransformasi akan di-*import* dari *tool* rapid miner. Pada gambar 1 adalah data training. Hasil data setelah di-*import* dapat dilihat pada gambar 2.

Sheet1 | Sheet2 | Sheet3

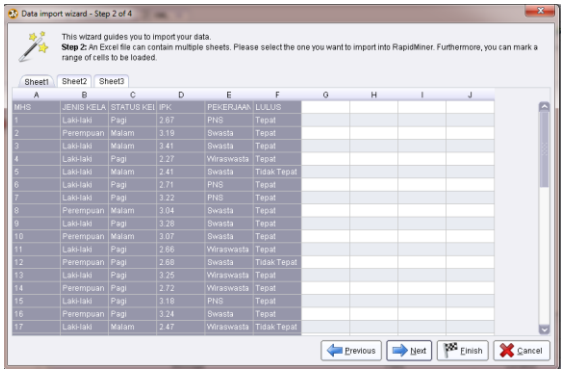
	A	B	C	D	E	F
1	JNS KELAMI	IPK	SKS	PRODI	THN LULUS	KET
2	L	3.620	148	SK	2016	YES
3	L	3.320	144	SK	2017	NO
4	L	3.320	152	SK	2015	YES
5	L	3.340	144	SK	2017	NO
6	L	3.100	148	SK	2017	NO
7	L	3.360	148	SK	2016	YES
8	L	3.260	144	SK	2017	NO
9	L	3.690	142	SK	0	NO
10	L	2.880	140	SK	0	NO
11	L	3.230	142	SK	0	NO
12	L	3.670	142	SK	0	NO
13	L	3.170	142	SK	0	NO
14	L	3.290	146	SK	2016	YES
15	L	3.330	148	SK	2016	YES
16	L	3.200	148	SK	2017	NO
17	L	3.290	148	SK	2016	YES
18	L	3.440	148	SK	2016	YES
19	L	3.350	142	SK	0	NO
20	L	3.430	146	SK	0	NO
21	L	2.480	142	SK	0	NO
22	L	2.330	90	SK	0	NO
23	L	3.610	146	SK	2016	YES
24	L	3.010	144	SK	2017	NO
25	L	2.990	144	SK	2017	NO
26	L	3.230	142	SK	0	NO
27	L	3.110	142	SK	0	NO

Gambar 1. Data Training



Gambar 2. Import Data Training

Pemilihan data training. Data yang telah di-import dari *MS. Excel* didipilih area yang akan digunakan sebagai data training. Pada gambar 3 adalah gambar pada saat data *training* di-pilih.

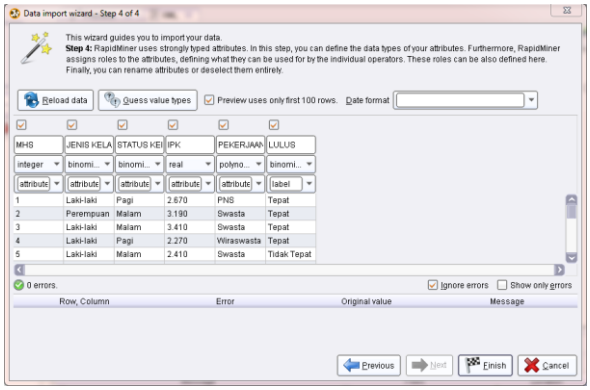


Gambar 3. Pemilihan Data Training

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:

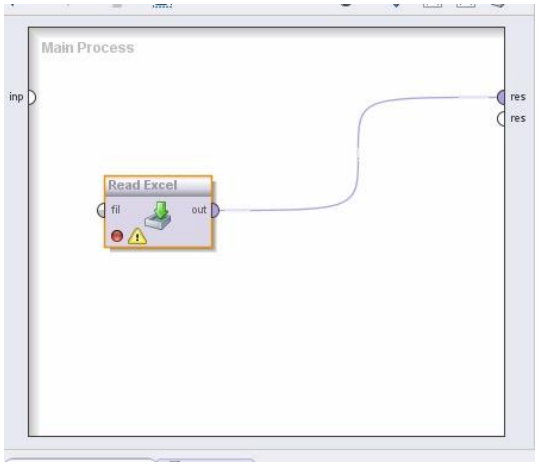
- Polynom = tipe data ini untuk karakter baik angka ataupun huruf (sama seperti varchar/text)
- Binominom = tipe data ini untuk 2 kategori (Y/T, L,P, Besar/Kecil, dll)
- Atribut = digunakan sebagai *variable predictor*/prediksi
- Label = digunakan sebagai variable tujuan

Pada gambar 4 adalah gambar pada saat menentukan label dan tipe dari data training.



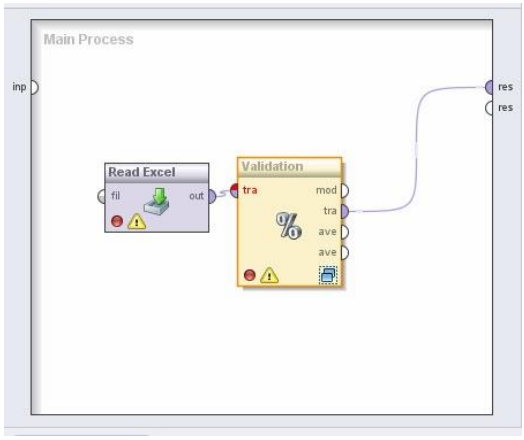
Gambar 4. Penentuan Label dan Tipe Data

Ambil data hasil import. Setelah data selesai di-import, akan dilakukan proses pengambilan data melalui wizard di tab operator. Pada gambar 5 adalah gambar pengambilan data hasil import.

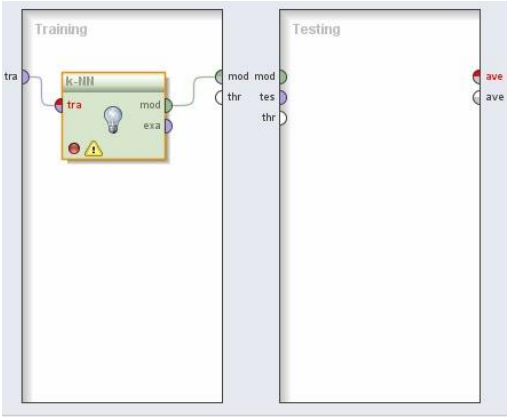


Gambar 5. Hasil Import Data dari Excel

Training dan akurasi metode naïve bayes. Masukkan *tool validation*, digunakan untuk *training* dan akurasi sehingga menghasilkan *validation* berada pada *main process* seperti yang terlihat pada gambar 6.

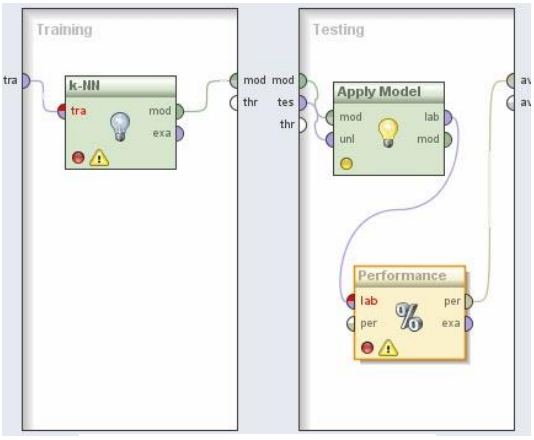


Gambar 6. Tool Validation



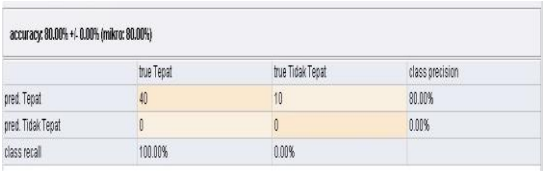
Gambar 7. Tool Naïve Bayes

kemudian tambahkan *tool Apply Model* dan Performance di kotak *Testing* seperti yang terlihat pada gambar 8.



Gambar 8. *Tool Apply Model* dan *Performance*

Kemudian kembali ke *main proses* dengan menekan tombol panah biru yang menghadap keatas Klik tool read excel, klik tombol *configuration wizard* pada sebelah kanan atas. Pilih file excel, klik next, next, next, pilih salah satu kolom menjadi label dan dirubah menjadi nominal klik finish. Kemudian di-RUN dengan menekan tombol panah warna biru, maka akan muncul hasil akurasinya sebesar 80.00% seperti terlihat pada gambar 9 serta hasil kurva ROC pada gambar 10



accuracy: 80.00% +/- 0.00% (mikro: 80.00%)

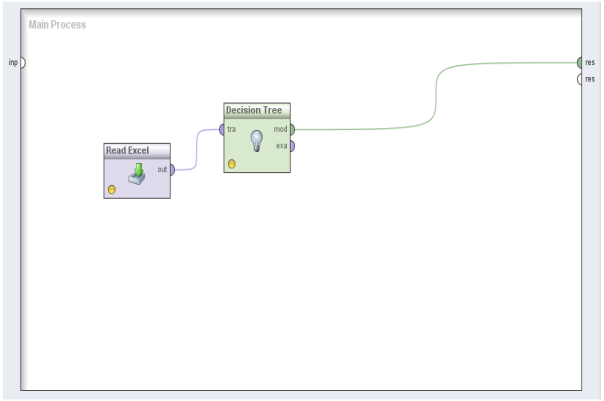
	true Tepat	true Tidak Tepat	class precision
pred. Tepat	40	10	80.00%
pred. Tidak Tepat	0	0	0.00%
class recall	100.00%	0.00%	

Gambar 9. Hasil Akurasi

Gambar 10. Kurva ROC

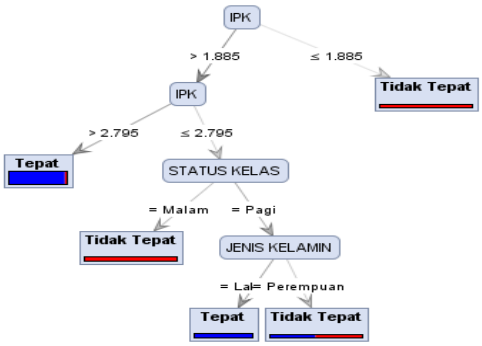
Pengolahan Data dengan Algoritma C4.5

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.



Gambar 11. Main Proses

Pohon keputusan C4.5. Tujuan dari menganalisis data dengan menggunakan algoritma *decision tree* adalah ingin mendapatkan *rule* yang akan dimanfaatkan untuk pengambilan keputusan pada data baru. [6]



Gambar 12. Pohon Keputusan C4.5

Aturan-aturan yang muncul adalah sebagai berikut :
"JIKA IPK = < 1.9 MAKA class = Tidak Tepat"
"JIKA IPK = >2.8 MAKA class = Tepat"
"JIKA IPK = < 2.8 DAN Status Kelas = Malam MAKA class = Tidak Tepat"
"JIKA IPK = < 2.8 DAN Status Kelas = Pagi DAN Jenis Kelamin = Laki-laki MAKA class = Tepat"
"JIKA IPK = < 2.8 DAN Status Kelas = Pagi DAN Jenis Kelamin = Perempuan MAKA class = Tidak Tepat"
Hasil klasifikasi pada data training adalah atribut IPK sebagai root pada *decision tree*, sedangkan atribut lainnya sebagai child node. Dari data training dengan jumlah 50 data dihasilkan 5 aturan. Aturan-aturan yang telah dari data training diperoleh dapat digunakan sebagai aturan untuk menentukan kelulusan tepat waktu atau tidak pada mahasiswa STMIK STIKOM Bali.

SIMPULAN

Hasil analisa data training menggunakan metode C4.5 diperoleh hasil bahwa deng-an pemilihan data training 50 record, 4 atribut predictor dan 1 atribut target menghasilkan 5 aturan dalam pohon ke-putusan sehingga aturan tersebut dapat digunakan dalam menentukan kelulusan tepat waktu pada mahasiswa STMIK STIKOM Bali.
Hasil analisa menggunakan metode Naïve Bayes diperoleh hasil akurasinya sebesar 89.27% dimana hasil *perfor-mance* akurasi menunjukkan kelulusan tepat waktu sebanyak 40 dan tidak tepat 10.

DAFTAR PUSTAKA

[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] Santosa, Budi. 2007. Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis, Teori dan Aplikasi. Graha Ilmu Yogyakarta.

[4] Basuki, Achmad dan Syarif, Iwan. 2003. Modul Ajar Decision Tree. Surabaya : PENS-ITS.

[5] Rapid-I GmbH. (2008). Rapidminer-4.2-tutorial. Germany: Rapid-I.

[6] Ian H. Witten, frank Eibe, and Mark A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Asma Stephan and Burlington, Eds. United States of America: Morgan Kaufmann,.