PERBANDINGAN KINERJA METODE KLASIFIKASI DATA MINING MENGGUNAKAN NAÏVE BAYES DAN ALGORITMA C4.5 UNTUK PREDIKSI KETEPATAN WAKTU KELULUSAN MAHASISWA

Gian Fiastantyo A11.2009.04932
Program Studi Teknik Informatika –S1
Fakultas Ilmu Komputer
Universitas Dian Nuswantoro, Jl. Nakula 1 No. 5-11. Semarang
gian.fiastantyo@gmail.com

ABSTRAK

Perguruan tinggi adalah jenjang pendidikan yang dianggap sebagai gerbang terakhir bagi pelajar untuk menimba ilmu sebelum akhirnya melibatkan diri dalam persaingan kerja. Jumlah mahasiswa yang lulus tepat waktu menjadi indikator efektifitas dari sebuah perguruan tinggi baik negeri dan swasta. Penelitian dalam hal memprediksi tingkat kelulusan mahasiswa telah banyak dilakukan. Dalam penelitian ini dilakukan perbandingan metode data mining yaitu algoritma naïve bayes dan C4.5, yang diterapkan pada data mahasiswa strata 1 Fakultas Ilmu Komputer Universitas Dian Nuswantoro. Naïve bayes adalah metode yang menghitung probabilitas dari tingkat kemunculan data yang satu terhadap data yang lainnya. Algoritma C4.5 adalah satu dari sebagian algoritma dalam metode decision tree yang mengubah data menjadi pohon keputusan, untuk kemudian dapat disimpulkan menjadi rule-rule. Berdasarkan hasil pengujian dengan mengukur kinerja kedua metode tersebut menggunakan metode pengujian confusion matrix, kemudian diketahui bahwa C4.5 memiliki nilai akurasi yang lebih baik yakni sebesar 77,354% , sedangkan naïve bayes memiliki nilai akurasi mencapai 74,094%. Kemudian berdasarkan perbandingan kinerja kedua metode tersebut, metode dengan pencapaian nilai akurasi terbaik akan diimplementasikan dalam bentuk sebuah Decision Support System.

Kata Kunci : data mining, klasifikasi, kelulusan, algoritma C4.5, naïve bayes

I. Pendahuluan

Perguruan tinggi adalah jenjang pendidikan yang dianggap sebagai gerbang terakhir bagi pelajar untuk menimba ilmu sebelum akhirnya melibatkan diri dalam persaingan kerja. Saat ini institusi perguruan tinggi berada dalam lingkungan yang sangat kompetitif. Sehingga perguruan tinggi kini dituntut untuk memiliki keunggulan dalam bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Selain sumber daya manusia, sarana, serta prasarana, sistem informasi adalah contoh lain dari beberapa sumber daya yang dapat digunakan guna meningkatkan kemampuan dan daya saing perguruan tinggi. Sistem informasi dalam hal ini dapat digunakan guna memperoleh, mengolah serta menyebarkan informasi yang telah diolah, agar dapat menunjang berbagai kegiatan operasional, sekaligus dapat berperan serta dalam mendukung pengambilan keputusan strategis yang akan dilakukan.

Institusi perguruan tinggi kini diwajibkan meningkatkan kualitas layanan dan memuaskan para mahasiswa serta

ruang publik disekitar mereka. Perguruan tinggi menganggap mahasiswa dan dosen sebagai *resource* utama dan mereka ingin terus menggunakan *resource* tersebut dengan cara yang lebih efektif [18]. Dalam struktur pendidikan saat ini, mahasiswa memiliki peran penting bagi sebuah institusi pendidikan. Oleh karena itu perlu ditinjau ulang mengenai tingkat kelulusan mahasiswa tepat pada waktunya.

Kelulusan tepat waktu merupakan isu penting yang perlu disikapi dengan bijak oleh institusi pendidikan. Tingkat kelulusan dianggap sebagai salah satu parameter efektifitas institusi pendidikan [18]. Sehingga saat ini memperhatikan tingkat kelulusan tepat waktu suatu perguruan tinggi menjadi hal penting. Penurunan tingkat kelulusan mahasiswa akan berpengaruh terhadap akreditasi perguruan tinggi tersebut. Oleh karena itu perlu adanya *monitoring* serta *evaluasi* terhadap kecenderungan kelulusan mahasiswa, tepat waktu atau tidak.

Berdasar deskripsi di atas, jelas bahwa memprediksi kelulusan adalah hal yang penting bagi institusi dan potensi besar bagi institusi untuk menyikapi serta menentukan kebijaksaan strategis perihal kelulusan tepat waktu. Setelah institusi melakukan identifikasi mahasiswa yang beresiko, kemudian dilanjutkan dengan mekanisme evaluasi dan *monitoring*, guna meningkatkan kegigihan mahasiswa, dengan harapan meningkatnya tingkat kelulusan. Tugas prediksi ini dapat dianggap membagi mahasiswa menjadi dua kelas yaitu "tepat" bagi mahasiswa yang lulus tepat waktu dan "terlambat" bagi mahasiswa yang lulus terlambat.

Penelitian dalam hal pengolahan data siswa atau mahasiswa telah dilakukan dengan beberapa metode yaitu [7], [14], [19]. Tetapi belum ada yang melakukan perbandingan kinerja antara metode *naïve bayes* dan algoritma C4.5 sehingga belum diketahui metode yang paling akurat.

Oleh sebab itu dalam penelitian ini akan dilakukan perbandingan metode *Naïve Bayes*, dan Algoritma C4.5 sehingga dapat diperoleh metode dengan akurasi prediksi ketepatan kelulusan mahasiswa yang terbaik berdasar model data yang ada.

II. Latar Belakang

A. Naïve Bayes

Bayes merupakan teknik prediksi berbasis probabilistic sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam Naïve Bayes, model yang digunakan adalah "model fitur independen".

Dalam Naïve Bayes, maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Contohnya pada kasus klasifikasi hewan dengan fitur penutup kulit, melahirkan, berat, dan menyusui. Dalam dunia nyata, hewan yang berkembang biak dengan cara melahirkan dapat dipastikan juga menyusui. Di sini ada ketergantungan

pada fitur menyusui karena hewan menyusui biasanya melahirkan, atau hewan yang bertelur biasanya tidak menyusui. Dalam Bayes, hal tersebut tidak dipandang sehingga masing-masing fitur seolahtidak memiliki hubungan apa pun.

Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut:

$$P(H|E) = \frac{P(E|H)xP(H)}{P(E)} \quad (6)$$

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peritiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu:

1. Sebuah probabilitas awal/priori H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Kaitan antara Naïve Bayes dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi. Jika X adalah vector masukan yang berisi fitur dan Y adalah label kelas. Naïve Bayes dituliskan dengan P(Y|X). Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y.

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikan dengan mecari nilai Y' dengan memaksimalkan nilai P(Y'|X') yang didapat.

Formulasi Naïve Bayes untuk klasifikasi adalah

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)}$$

P(Y|X) adalah probabilitas data dengan vector X pada kelas Y. P(Y) adalah probabilitas awal kelas Y. $\prod_{i=1}^{q} P(X_i|Y)$ adalah probabilitas independen kelas Y dari semua fitur dalam vector X. Nilai P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya kita tinggal menghitung bagian $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ dengan memilih yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi. Sementara probabilitas independen $\prod_{i=1}^{q} P(X_i|Y)$ tersebut merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan:

$$P(X|Y=y) = \prod_{i=1}^{q} P(X_i|Y=y)$$

Dan setiap fitur X= $\{X_1, X_2, X_3, \ldots, X_q\}$ terdiri atas q atribut.

B. Algoritma C4.5

Tree atau pohon banyak dikenal sebagai bagian dari *Graph,* yan termasuk dalam irisan bidang ilmu otomata dan teori bahasa serta matematika diskrit. *Tree* sendiri merupakan graf tak-berarah yang terhubung, serta tidak mengandung sirkuit. [15] Dalam sebuah *tree,* setiap pasang simpul terhubung hanya oleh satu lintasan, dan sebuah *tree* terdiri dari [25]:

a) Root/akar, yang merupakan simpul tertinggi.
b) Leaf/daun, yang berupa simpul tanpa anak lagi .
c) Branch/cabang, yang merupakan simpul-simpul selain daun.

Decision tree merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode *decision tree* mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Selain itu aturan juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* (SQL) untuk mencari *record* pada kategori tertentu.

Decision tree juga berguna dalam mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Karena *decision tree* memadukan antara eksplorasi data dan pemodelan. *Decision tree* digunakan untuk kasus-kasus dimana outputnya bernilai diskrit [10].

Sebuah *decison tree* adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip dengan yang lain [4]

Proses pada *decision tree* adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule* [3].

Sebuah model *decision tree* terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih homogen dengan memperhatikan pada variabel tujuannya. Variabel tujuan biasanya dikelompokkan dengan pasti dan lebih mengarah pada perhitungan probabilitas dari tiap-tiap record terhadap kategori-kategori tersebut atau untuk mengklasifikasi record dengan mengelompokkannya dalam satu kelas.

Data dalam decision tree biasanya dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut

menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Atribut ini juga memiliki nilai yang terkandung didalamnya yang disebut instance. Dalam decision tree setiap atribut akan menempati posisi simpul. Selanjutnya setiap simpul akan memiliki jawaban yang dibentuk dalam cabang-cabang, jawaban ini adalah instance dari atribut (simpul) yang ditanyakan. Pada saat penelusuran, pertanyaan pertama akan ditanyakan pada simpul akar. Selanjutnya akan dilakukan penelusuran ke cabang-cabang simpul akar dan simpul-simpul berikutnya. Penelusuran setiap simpul ke cabang-cabangnya akan berakhir ketika suatu cabang telah menemukan simpul kelas atau obyek yang dicari.

Saat menyusun sebuah decision tree pertama yang harus dilakukan adalah menentukan atribut mana yang akan menjadi simpul akar dan atribut mana yang akan menjadi simpul selanjutnya. Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan decision tree yang paling kecil ukurannya. Atau atribut yang bisa memisahkan obyek menurut kelasnya. Secara heuristik atribut yang dipilih adalah atribut yang menghasilkan simpul yang paling ”purest” (paling bersih). Ukuran purity dinyatakan dengan tingkat impurity, dan untuk menghitungnya, dapat dilakukan dengan menggunakan konsep Entropy, Entropy menyatakan impurity suatu kumpulan objek. Jika diberikan sekumpulan objek dengan label/output y yang terdiri dari objek berlabel 1, 2 sampai n, Entropy dari objek dengan n kelas ini dapat dihitung dengan rumus berikut.

$$Entropy(y) = -p_1 log_2 p_1 - p_2 log_2 p_2 \ldots . - p_n log_n p_n \qquad (1)$$

Kemudian setelah itu ada beberapa kriteria yang dibahas, yakni Information Gain, Gain Ratio, Indeks Gini.

1. Information Gain

   Information gain adalah kriteria yang paling populer untuk pemilihan atribut. Information gain dapat dihitung dari output data atau variabel dependent y yang dikelompokkan berdasarkan atribut A, dinotasikan dengan gain (y,A). Information gain, gain (y,A), dari atribut A relatif terhadap output data y adalah :

$$gain(y, A) = entropy(y) - \sum_{c \in nilai(A)} \frac{y_c}{y} entropy(y_c) \qquad (2)$$

Dimana nilai(A) adalah semua nilai yang mungkin dari atribut A, dan y_c adalah subset dari y dimana A mempunyai nilai c.

2. Gain Ratio

Untuk menghitung gain ratio diperlukan suatu term

SplitInformation. SplitInformation dapat dapat dihitung dengan formula sebagai berikut :

Dimana sampai adalah c subset yang dihasilkan dari pemecahan S demgan menggunakan atribut A yang mempunyai sebanyak c nilai. Selanjutnya gain ratio dihitung dengan cara :

3. Indeks Gini

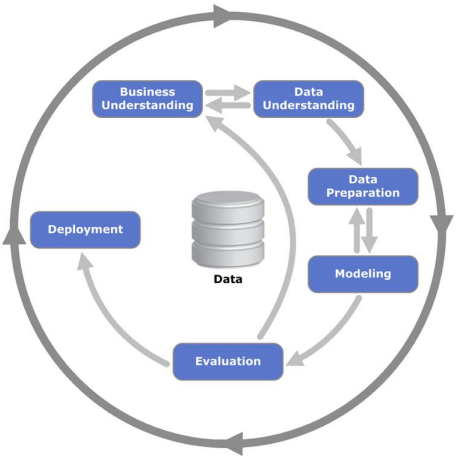
Jika kelas obyek dinyatakan dengan k, k-1,2, ...C, dimana C adalah jumlah kelas untuk variabel/output dependent y, Indeks Gini untuk suatu cabang atau kotak A dihitung sebagai berikut :

Dimana pk adalah ratio observasi dalam kotak A yang masuk dalam kelas k. Jika IG(A) = 0 berarti semua data dalam kotak A berasal dari kelas yang sama. Nilai IG(A) mencapai maksimum jika dalam kelas A proporsi data dari masing-masing kelas yang ada mencapai nilai yang sama.

Algortima C4.5 sendiri merupakan pengembangan dari algortima ID3, dimana pengembangan dilakukan dalam hal: bisa mengatasi missing data, bisa mengatasi data kontiyu, pruning. Secara umum langkah algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

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

C. Tahap-Tahap Data Mining



Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan *knowledge base*.

a. *Business Understanding* atau pemahaman domain (penelitian). Pada fase ini dibutuhkan pemahaman tentang substansi dari kegiatan data mining yang akan dilakukan, kebutuhan dari perspektif bisnis. Kegiatannya antara lain: menentukan sasaran atau tujuan bisnis, memahami

situasi bisnis, menentukan tujuan data mining dan membuat perencanaan strategi serta jadwal penelitian.

b. *Data Understanding* atau pemahaman data adalah fase mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai. Fase ini mencoba mengidentifikasikan masalah yang berkaitan dengan kualitas data, mendeteksi subset yang menarik dari data untuk membuat hipotesa awal.
c. *Data preparation* atau persiapan data. Fase ini sering disebut sebagai fase yang padat karya. Aktivitas yang dilakukan antara lain memilih *table* dan *field* yang akan ditransformasikan ke dalam *database* baru untuk bahan *data mining* (set data mentah).
d. *Modeling* adalah fase menentukan teknik *data mining* yang digunakan, menentukan *tools data mining*, teknik *data mining*, algoritma data mining, menentukan parameter dengan nilai yang optimal.
e. *Evaluation* adalah fase interpretasi terhadap hasil *data mining* yang ditunjukan dalam proses pemodelan pada fase sebelumnya. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam fase pertama.
f. *Deployment* atau penyebaran adalah fase penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses *data mining* [11].

D. Confusion Matrix

Confusion Matrix adalah alat (tools) visualisasi yang biasa digunakan pada supervised learning. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya (Gorunescu, 2011).

Confusion matrix berisi informasi aktual (actual) dan prediksi (predicted) pada sisitem klasifikasi. Tabel 2.2 adalah contoh tabel confusion matrix yang menunjukan klasifikasi dua kelas.

		Prediksi	
Aktual	Negaif	A	C
	Positif	B	D

Keterangan:

A = jumlah prediksi yang tepat bahwa instance bersifat negatif

B =jumlah prediksi yang salah bahwa instance bersifat positif

C =jumlah prediksi yang salah bahwa instance bersifat negatif

D =jumlah prediksi yang tepat bahwa instance bersifat positif.

Beberapa persyaratan standar yang telah didefiniskan untuk matrik klasifikasi dua kelas:

a. Keakuratan (AC) adalah proposi jumlah prediksi benar. Rumus persamaannya:

$$AC = A + D/A + B + C + D$$

b. Penarikan kembali (recall) atau tingkat positif benar (TP) adalah

proporsi kasus positif yang diidentifikasi dengan benar, yang dihitung dengan persamaan:

$$TP = D/C + D \quad (12)$$

c. Tingkat positif salah (FP) adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, yang dihitung dengan menggunakan persamaan:

$$FP = B/A+B \quad (13)$$

d. Tingkat negatif sejati (TN) didefinisikan sebagai proporsi kasus negative yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan:

$$TN = A/A + B \quad (14)$$

e. Tingkat negatif palsu (FN) adalah proporsi kasus positif yang salah diklasifikasikan sebagai negatif, yang dihitung dengan menggunakan persamaan:

$$FN = C/C + D$$

f. Presisi (P) adalah proporsi prediksi kasus positif yang benar, yang dihitung dengan menggunakan persamaan:

$$P = D/B + D$$

III. Desain Penelitian

Metode yang digunakan dalam penelitian ini adalah model CRISP-DM (*Cross Industry Standard Process for Data Mining*), dengan langkah-langkah sebagai berikut:

a. Pemahaman Bisnis (*Bussiness Understanding*)

Saat ini institusi perguruan tinggi berada dalam lingkungan yang sangat kompetitif. Sehingga perguruan tinggi kini dituntut untuk memiliki keunggulan dalam bersaing dan diwajibkan meningkatkan kualitas layanan serta memuaskan para mahasiswa serta ruang publik disekitar mereka. Dalam struktur pendidikan saat ini, mahasiswa memiliki peran penting bagi sebuah institusi pendidikan. Karena itu perlu ditinjau ulang mengenai tingkat kelulusan mahasiswa tepat pada waktunya.

Kelulusan tepat waktu merupakan isu penting yang perlu disikapi dengan bijak oleh institusi pendidikan. Tingkat kelulusan dianggap sebagai salah satu parameter efektifitas institusi pendidikan (Qudri & Kalyankar, 2010). Sehingga saat ini memerhatikan tingkat kelulusan tepat waktu suatu perguruan tinggi menjadi hal penting. Penurunan tingkat kelulusan mahasiswa akan berpengaruh terhadap akreditasi perguruan tinggi tersebut. Oleh karena itu perlu adanya monitoring serta evaluasi terhadap kecenderungan kelulusan mahasiswa, tepat waktu atau tidak.

b. Pemahaman Data (*Data Understanding*)

Data yang digunakan dalam penelitian ini tidak diperoleh dari sumber data, dalam hal ini mahasiswa, secara langsung. Melainkan data ini diperoleh dari

database mahasiswa yang dimiliki oleh Universitas Dian Nuswantoro Semarang, yaitu melalui bagian ruang data yang dimiliki oleh fakultas Ilmu Komputer. Data yang dikumpulkan adalah data mahasiswa fakultas Ilmu Komputer dengan program studi strata satu (S1) untuk tahun angkatan 2008 dan 2009. Data terkumpul sebanyak 1919 data, dengan atribut nim (nomor induk mahasiswa), nama, program studi, umur, jenis kelamin, status marital, status pekerjaan, ip (indeks prestasi) semester 1 sampai dengan ip semester 8, dengan label keterangan tepat atau terlambat.

c. Pengolahan Data (*Data Preparation*)

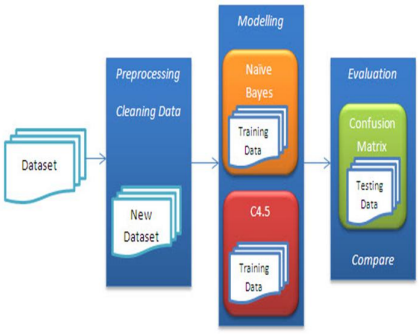
Pada tahap ini atribut data yang akan digunakan adalah sebagai berikut

Atribut	Detail Penggunaan	
NIM	√	ID
Nama	×	No
Program Studi	√	Nilai Model
Jenis Kelamin	√	Nilai Model
Umur	√	Nilai Model
Status Marital	√	Nilai Model
Status	√	Nilai Model
IPS 1	√	Nilai Model
IPS 2	√	Nilai Model
IPS 3	√	Nilai Model
IPS 4	√	Nilai Model
IPS 5	×	No
IPS 6	×	No
IPS 7	×	No
IPS 8	×	No
Keterangan	√	Label Target

Tabel diatas menjelaskan mengenai atribut yang akan digunakan dalam penelitian, indikator yes (√) menandakan bahwa atribut bersangkutan akan digunakan dalam penelitian, sedangkan indikator no (×) menandakan bahwa atribut tersebut akan dieliminasi pada tahap *data preparation.*

d. Pemodelan (*Modelling*)

Terdapat dua metode yang akan digunakan dalam penelitian ini, yaitu *Naïve Bayes* dan Algoritma C4.5. Untuk melakukan pengukuran serta perbandingan akurasi dalam penelitian ini akan menggunakan *framework* RapidMiner versi 6.



e. Validasi dan Evaluasi

Dalam tahapan ini akan dilakukan validasi serta pengukuran keakuratan hasil yang dicapai oleh model menggunakan beberapa teknik yang terdapat dalam *framework* RapidMiner versi. 5.13 yaitu *Confusion Matrix* untuk pengukuran tingkat akurasi model, dan *Split Validation* untuk validasi.

f. Penyebaran (Deployment)

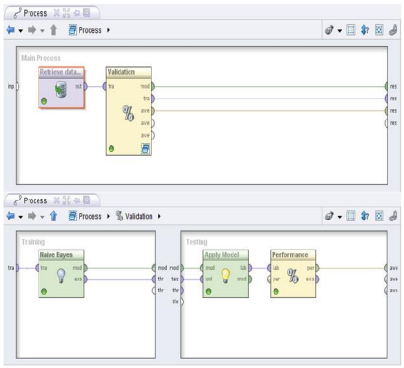
Hasil dari penelitian ini berupa analisa yang mengarah ke DSS (*Decision Support System*), yang diharapkan dapat digunakan oleh

institusi perguruan tinggi sebagai bahan pertimbangan dalam menentukan langkah guna mengatasi permasalahan ketepatan kelulusan mahasiswa, dan juga dapat digunakan sebagai bahan rujukan untuk penelitian selanjutnya. Selain itu hasil analisa ini juga akan digunakan sebagai dasar perancangan sebuah sistem pengambilan keputusan guna melakukan idemtifikasi ketepatan kelulusan mahasiswa.

IV. Hasil Pengujian

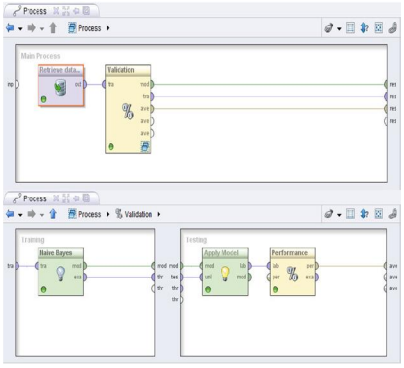
a. Pembahasan model Naïve Bayes

Pengaturan dan penggunaan operator serta parameter dalam framework RapidMiner sangat berpengaruh terhadap akurasi dan model yang terbentuk, sebagai contoh dalam penggunaan operator metode *naïve bayes* dibawah ini.



b. Pembahasan model Algoritma C4.5

Serupa dengan proses dalam membangun model *naïve bayes*, pengaturan dan penggunaan operator serta parameter pada model C4.5 juga sangat berpengaruh terhadap akurasi yang dihasilkan. Berikut desain model C4.5 yang akan digunakan.



c. Hasil Komparasi

Berdasarkan sembilan kali pengujian yang dilakukan dengan metode *sampling* dan ratio perbandingan yang berbeda-beda dari *data training* dan *data testing* dengan menggunakan kedua metode diatas, didapatkan hasil pengujian sebagai berikut:

	Metode	Sampling Type	Ratio Data Training (%)									$\bar{X}$
			10	20	30	40	50	60	70	80	90	
Akurasi	Naïve Bayes	Linear	70.41	68.01	67.31	67.07	65.90	65.80	66.32	64.06	62.50	66.15
		Shuffled	73.89	73.42	73.12	73.41	73.72	73.05	72.57	72.92	73.44	73.28
		Stratified	75.39	73.89	73.12	73.15	73.93	72.27	74.13	73.37	77.60	74.09
	C4.5	Linear	76.55	76.16	78.41	82.10	82.38	79.82	87.85	89.06	89.58	82.43
		Shuffled	74.64	70.42	77.81	77.67	77.37	78.26	78.30	78.65	80.21	77.03
		Stratified	77.88	76.11	76.02	76.54	78.94	78.39	77.95	76.76	77.60	77.34

Tabel perbandingan diatas menampilkan hasil pengujian dari metode *naïve bayes* dan C4.5 dengan metode *sampling* yang bervariasi, serta ratio penggunaan *data training* yang bertahap, mulai dari 10% hingga 90% dari keseluruhan 1919 data mahasiswa yang tersedia di *dataset*. Dan dapat disimpulkan bahwa metode *sampling Linear* serta metode algortima C4.5 memiliki tingkat akurasi yang lebih baik dalam melakukan prediksi ketepatan kelulusan mahasiswa.

V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan maka dapat diambil beberapa kesimpulan, antara lain:

1. Dalam melakukan prediksi tingkat ketepatan kelulusan mahasiswa, dengan menggunakan pemodelan metode *Decision Tree* didapatkan tingkat akurasi tertinggi sebesar 82.43%, dengan menggunakan parameter *Split Relative* dan *Sampling Type Linear,* sedangkan pada pemodelan metode Naïve Bayes memperoleh rata-rata tingkat akurasi tertinggi sebesar 74.09%, dengan menggunakan parameter *Split Relative* dan *Sampling Type Stratified*. Oleh karena itu dapat disimpulkan berdasarkan tingkat akurasi, bahwa pemodelan metode *Decision Tree* lebih baik dalam melakukan prediksi ketepatan kelulusan mahasiswa pada data penelitian mahasiswa strata 1 (S1) Fakultas Ilmu Komputer Universitas Dian Nuswantoro angkatan 2008 dan 2009.
2. Aplikasi yang dibangun berdasarkan hasil analisa dengan menggunakan RapidMiner, dapat digunakan sebagai *Decision Support System* (DSS) atau alat bantu pengambilan keputusan bagi pihak Fakultas Ilmu Komputer Universitas Dian Nuswantoro, guna merancang serta mempersiapkan langkah-langkah strategis dalam menyikapi permasalahan ketepatan kelulusan mahasiswa.

VI. DAFTAR PUSTAKA

[1] Azwar, S. (2004). *Penyusunan Skala Psikologi.* Yogyakarta: Pustaka pelajar.

[2] Balagatabi, Z. N. (2012). Comparison of Decision Tree and Naïve Bayes Methods in Classification of Researcher's Cognitive Styles in Academic Environment. *Journal of Advances in Computer Research*.

[3] Basuki, A., & Syarif, I. (2004). *Modul Ajar Decision Tree.* Surabaya: PENS-ITS.

[4] Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques.* New Jersey: John Willey and Sons Inc.

[5] Darmawan, A. (2012). Pembuatan Aplikasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighborhood. *Digilab Unikom*.

[6] Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques.* Verlag Berlin Heidelberg: Springer.

[7] Hamidah, I. (2012). Aplikasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5. *Digilab Unikom*.

[8] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques.* San Francisco: Mofgan Kaufann Publishers.

[9] Karamouzis, T. S., & Vrettos, A. (2008). An Artificial Neural Network for Predicting Student Graduation Outcomes. *Preceeding of World Congress on Engineering and Computer Science*.

[10] Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining.* Yogyakarta: Penerbit ANDI.

[11] Larose, D. T. (2005). *Discovering Knowledge in Databases.* New Jersey: John Willey and Sons Inc.

[12] Larose, D. T. (2006). *Data Mining Methods and Models*. New Jersey: John Wiley and Sons.

[13] Latifah, E. (2013). Perancangan Sistem Klasifikasi Masa Studi Mahasiswa Menggunakan Data Mining Berbasis Algoritma ID3. *Digilab Unikom*.

[14] Meinanda, M. H., Annisa, M., Muhandri, N., & Suryadi, K. (2009). Prediksi Masa Studi Sarjana dengan Artificial Neural Network. *Internetworking Indonesia Journal*, 31-35.

[15] Munir, R. (2010). *Matematika Diskrit*. Bandung: Informatika Bandung.

[16] Nuswantoro, U. D. (2006). Peraturan Akademik.

[17] Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Penerbit ANDI.

[18] Quadril, M. N., & Kalyankar, N. V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science*.

[19] Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data MiningUntuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Jurnal EECCIS*, 59-63.

[20] Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Surabaya: Graha Ilmu.

[21] Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data. *Journal of Data Warehousing*.

[22] Shereker, S. S., & Patil, T. R. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications*.

[23] Siregar, A. R. (2006). Motivasi Belajar Mahasiswa ditinjau dari Pola Asuh. *USU Repository*.

[24] Sivakumari, Priyadarsini, & Amudha. (2009). Accuracy Evaluation of C4.5 and Naïve Bayes Classifiers Using Atribute Ranking Method.

[25] Utdirartatmo, F. (2005). *Teori Bahasa dan Otomata*. Yogyakarta: Graha Ilmu.

[26] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Machine Learning Tools and Techniques*. Burlington: Morgan Kaufmann Publishers.

[27] Yingkuachat, J., Praneetpolgrang, P., & Kijsirikul, B. (2007). An Application of the Probabilistic Model to the Prediction of Student Graduation Using Bayesian Belief Networks. *ECTI Transaction on Computer and Technology*.