PERBANDINGAN KINERJA ALGORITMA C4.5 DAN NAIVE BAYES UNTUK KETEPATAN PEMILIHAN KONSENTRASI MAHASISWA

Wiwit Supriyanti[1)], Kusrini[2)], Armadyah Amborowati[3)]

STMIK AMIKOM Yogyakarta[1),2),3)]
Email : wiwitsupriyanti13@gmail.com[1)], kusrini@amikom.ac.id[2)], armadyah.a@amikom.ac.id[3)]

Abstrak

Penentuan konsentrasi akan membantu mahasiswa lebih fokus terhadap apa yang diminati dan disesuaikan dengan nilai akademis yang dimilikinya. Banyak mahasiswa yang masih belum mengenal minat dan kemampuan yang dimilikinya. Hal tersebut membuat mahasiswa cenderung memilih dan menjalani konsentrasi yang tidak sesuai dengan minat dan kemampuannya. Perbandingan kinerja algoritma C4.5 dan naive bayes bertujuan untuk mengukur tingkat akurasi terbaik masing-masing algoritma untuk diterapkan dalam kasus pemilihan konsentrasi keahlian mahasiswa.

Penelitian ini mengambil sampel data alumni di Program Studi Informatika Universitas Muhammadiyah Surakarta. Variabel yang digunakan dalam penelitian ini antara lain: jurusan sekolah asal mahasiswa, gender, nilai akademik dari semester satu sampai dengan semester enam, konsentrasi keahlian yang dipilih serta lama studi yang ditempuh. Forward selection adalah salah satu metode seleksi fitur yang dapat digunakan untuk mengurangi atribut yang kurang relevan pada dataset. Penggunaan metode forward selection mampu menghasilkan tingkat akurasi yang lebih baik dibandingkan dengan tanpa penambahan seleksi fitur yang hanya mencapai tingkat akurasi pada algoritma C4.5 dari sebelumnya sebesar 84,43% meningkat menjadi 84,98%, sedangkan pada algoritma naive bayes sebelumnya sebesar 78,47% meningkat menjadi 82,01%.

Hasil dari komparasi algoritma klasifikasi antara decision tree C4.5, dan naive bayes yang digabungkan dengan metode seleksi fitur forward selection untuk kasus ketepatan pemilihan konsentrasi mahasiswa didapatkan tingkat akurasi tertinggi dengan algoritma terpilih C4.5 dengan nilai akurasi sebesar 84,98%.

Kata Kunci : konsentrasi keahlian, klasifikasi data mining, C4.5, naive bayes, forward selection

1. PENDAHULUAN

Penentuan konsentrasi akan membantu mahasiswa lebih fokus terhadap apa yang diminati dan disesuaikan dengan nilai akademis yang dimilikinya. Banyak mahasiswa yang masih belum mengenal minat dan kemampuan yang dimilikinya. Hal tersebut membuat mahasiswa cenderung memilih dan menjalani konsentrasi yang tidak sesuai dengan minat dan kemampuannya. Di sisi lain, jurusan (program studi) sebagai unit dalam perguruan tinggi yang terlibat langsung dengan transaksi akademik mahasiswa, pasti memiliki data akademik dan beberapa kebijakan terkait pengambilan konsentrasi. Penentuan konsentrasi seorang mahasiswa tentunya tidak terlepas dari penguasaan mahasiswa terhadap mata kuliah yang menjadi inti dari konsentrasi tersebut. Ketika seorang mahasiswa memilih suatu konsentrasi tertentu, harapan terbesar dari jurusan dan mahasiswa yang bersangkutan adalah dapat menyelesaikan studi dengan tepat waktu serta memiliki kompetensi sesuai dengan konsentrasi yang dipilihnya agar dapat diimplementasikan ke dalam dunia kerja yang sesuai dengan gelar sarjana yang disandang.

Penelitian lain yang mengkaji tentang perbandingan kinerja beberapa metode klasifikasi data mining sebelumnya telah dilakukan oleh A. K. Santra dan S. Jayasudha (2012), dalam penelitian ini menggunakan algoritma *Naive Bayes* untuk teknik klasifikasinya, dimana pada penelitian sebelumnya telah menggunakan algoritma C4.5. Hasil penelitian menunjukkan bahwa

algoritma Naïve Bayes lebih efisien kinerjanya bila dibandingkan dengan algoritma C4.5 dalam aplikasi *e-commerce* seperti *web caching*, *web page recommendation* dan *web personalization*. Khafiizh Hastuti (2012) membandingkan empat algoritma klasifikasi data mining yaitu logistic *regression, decision tree, naive bayes* dan *neural network* dengan menggunakan 3681 data set mahasiswa yang terdiri atas data demografi dan akademik mahasiswa sehingga dapat diketahui algoritma yang paling akurat untuk memprediksi mahasiswa non-aktif. George Dimitoglou, James A. Adams dan Carol M. Jim (2012) menguji kemampuan data mining dan mesin metode belajar untuk secara akurat memprediksi ketahanan hidup pasien yang didiagnosis menderita kanker paru-paru. Penelitian ini membandingkan efektivitas dari algoritma *naive bayes* dan *decision tree* C4.5 yang diimplementasikan untuk memprediksi ketahanan hidup seseorang akibat penyakit tertentu. Hasil yang diperoleh menunjukkan bahwa algoritma *naive bayes* lebih unggul dibandingkan dengan *decision tree* C4.5 untuk kasus tersebut. Tina R. Patil dan Mrs. S. S. Sherekar (2013) membuat evaluasi perbandingan antara teknik klasifikasi *naive bayes* dan J48 dalam konteks data set perbankan berdasarkan *true positive rate* dan *false positive rate* menggunakan tool WEKA. Hasil yang diperoleh menunjukkan bahwa tingkat akurasi dan efisiensi algoritma J48 lebih baik dibandingkan dengan *naive bayes*. Ahmad Ashari, Iman Paryudi dan A Min Tjoa (2013) mengusulkan sebuah metode baru dalam mencari alternatif desain yaitu dengan menggunakan metode klasifikasi. Metode yang digunakan dalam penelitian ini antara lain : *naive bayes*, *decision tree* dan *k-nearest neighbor*. Hasil percobaan menunjukkan bahwa decision tree unggul dalam proses kecepatan perhitungan diikuti oleh *naive bayes* dan *k-nearest neighbor*.

Pada penelitian ini penulis akan membandingkan kinerja dua metode klasifikasi dalam data mining untuk mendapatkan hasil pengujian paling akurat dalam mengolah informasi data mahasiswa sebagai dasar penentuan dalam pemilihan konsentrasi keahlian mahasiswa dengan sampel data dari Program Studi S1 Informatika Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta, sehingga mahasiswa mendapatkan solusi alternatif informasi untuk mengajukan konsentrasi keahlian secara tepat.

2. TINJAUAN PUSTAKA

2.1. Algoritma C4.5

Pohon keputusan merupakan metode yang umum digunakan untuk melakukan klasifikasi pada data mining. Seperti yang telah dijelaskan sebelumnya, klasifikasi merupakan suatu teknik menemukan kumpulan pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Metode ini popular karena mampu melakukan klasifikasi sekaligus menunjukkan hubungan antar atribut. Banyak algoritma yang dapat digunakan untuk membangun suatu decision tree, salah satunya ialah algoritma C4.5.

Algoritma C4.5 dapat menangani data numerik dan diskret. Algoritma C4.5 menggunakan rasio perolehan (gain ratio). Sebelum menghitung rasio perolehan, perlu dilakukan perhitungan nilai informasi dalam satuan bits dari suatu kumpulan objek, yaitu dengan menggunakan konsep entropi.

a. Konsep Entropi

Entropi (S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S. Entropi dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin kecil nilai entropi maka akan semakin entropy digunakan dalam mengekstrak suatu kelas. Entropi digunakan untuk mengukur ketidakaslian S.

Besarnya Entropi pada ruang sampel S didefinisikan dengan :

$$Entropy(S) \equiv -p_{\oplus}\, log_2\, p_{\oplus} - p_{\ominus}\, log_2 p_{\ominus}$$

b. Konsep Gain

Gain (S,A) merupakan Perolehan informasi dari atribut A relatif terhadap output data S. Perolehan informasi didapat dari output data atau variabel dependent S yang dikelompokkan berdasarkan atribut A, dinotasikan dengan gain (S,A).

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

Berikut merupakan cara membangun pohon keputusan dengan menggunakan algoritma C4.5 :

a. Pilih atribut sebagai akar. Sebuah akar didapat dari nilai gain tertinggi dari atribut-atribut yang ada.
b. Buat cabang untuk masing-masing nilai.
c. Bagi kasus dalam cabang.
d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

2.2. Naive Bayes

Naive Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam Naive Bayes model yang digunakan adalah "model fitur independen" (Prasetyo, 2012).

Dalam sebuah aturan yang mudah, sebuah klasifikasi Naive Bayes diasumsikan bahwa ada atau tidaknya ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Untuk contohnya, buah akan dianggap sebagai sebuah apel jika berwarna merah, berbentuk bulat dan berdiameter sekitar 6 cm. Walaupun jika ciri-ciri tersebut bergantung satu sama lainnya, dalam Bayes hal tersebut tidak dipandang sehingga masing-masing fitur seolah tidak memiliki hubungan apapun. Berdasarkan ciri alami dari sebuah model probabilitas, klasifikasi Naive Bayes bisa dibuat lebih efisien dalam bentuk pembelajaran. Dalam beberapa bentuk praktiknya, parameter untuk perhitungan model Naive Bayes menggunakan metode maximum likehood, atau kemiripan tertinggi.

Prediksi Naive Bayes didasarkan pada teorema Bayes dengan formula untuk klasifikasi sebagai berikut :

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

Sedangkan *Naive Bayes* dengan fitur kontinyu memiliki formula :

$$P(X|Y) = \frac{1}{\sqrt{2\pi}\,\sigma} exp^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

2.3. Forward Selection

Seleksi fitur atau yang lebih dikenal dengan feature selection, subset selection, attribute selection or variable selection adalah proses memilih fitur yang tepat untuk digunakan dalam proses klasifikasi atau clustering. Tujuan dari seleksi fitur ini adalah untuk mengurangi tingkat kompleksitas dari sebuah algoritma klasifikasi, meningkatkan akurasi dari algoritma klasifikasi tersebut, dan mampu mengetahui fitur-fitur yang paling berpengaruh terhadap tingkat akurasi.

Metode forward selection adalah pemodelan dimulai dari nol peubah (empty model), kemudian satu persatu peubah dimasukan sampai kriteria tertentu dipenuhi. Langkah-langkah metode forward selection adalah sebagai berikut (Draper dan Smith, 1992) :

a. Membuat model dengan meregresikan variabel respon Y dengan setiap variabel prediktor. Kemudian dipilih model yang mempunyai nilai R^2 tertinggi. Misal model tersebut adalah yang memuat prediktor X_a, yaitu:
$$\hat{Y} = b_0 + b_a X_a$$
b. Meregresikan variabel respon Y, dengan prediktor X_a, ditambah dengan setiap prediktor selain X_a dan prediktor lain. Kemudian dipilih model yang nilai R^2 nya tertinggi, misal mengandung tambahan prediktor X_b, yaitu model
$$\hat{Y} = b_0 + b_a X_a + b_b X_b$$
Prediktor terpilih X_b berarti mempunyai $F_{sequensial}$ tertinggi. Formula $F_{sequensial}$ untuk X_b adalah
$$F_{seq} = R(\beta_b \mid \beta_0, \beta_a) / MSE / db$$
Nilai $F_{sequensial}$ untuk X_b juga dapat diperoleh dengan cara mengkuadratkan nilai statistik uji T prediktor X_b.
c. Proses diulang sampai didapatkan $F_{sequensial} > F_{in}$. Nilai $F_{in} = F(1,v,\alpha_{in})$, sehingga model terbaik yang dipilih adalah model yang tidak mempunyai prediktor dengan $F_{sequensial} < F_{in}$.

2.4. RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk *input, output, data*

preprocessing dan visualisasi. RapidMiner merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan Bahasa pemrograman java sehingga dapat bekerja di semua sistem operasi.

3. METODE PENELITIAN

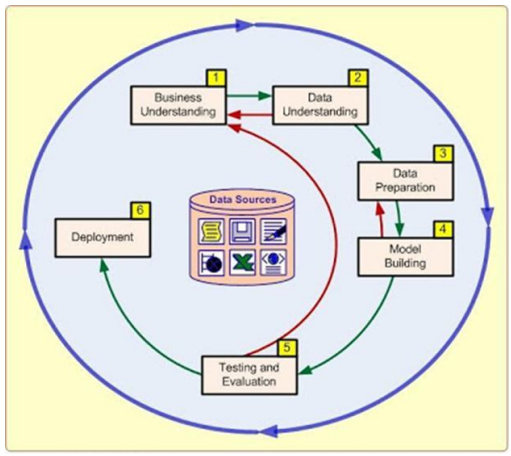
Penelitian yang dilaksanakan adalah jenis penelitian eksperimen, yaitu melakukan pengujian tingkat akurasi terbaik antara algoritma C4.5 dan *Naive Bayes* dalam pemilihan konsentrasi keahlian berdasarkan jurusan sekolah asal, *gender*, indeks prestasi, konsentrasi yang dipilih dan lama studi. Data eksperimen diambil dari data mahasiswa yang telah lulus pada Program Studi S1 Informatika Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta.

3.1. Metode Pengumpulan Data

a. Metode Observasi
   Melakukan pengamatan langsung ke Universitas Muhammadiyah Surakarta untuk memperoleh data yang dibutuhkan.
b. Metode Wawancara
   Mengadakan wawancara dengan pihak-pihak yang berkaitan langsung dengan permasalahan yang sedang dibahas pada penelitian ini untuk memperoleh gambaran dan penjelasan secara mendasar.
c. Metode Studi Pustaka
   Metode ini dengan mengumpulkan referensi dari literatur-literatur yang bisa mendukung penelitian sebagai landasan teori dan dasar pedoman dalam pembuatan laporan.

3.2. Metode Analisis Data

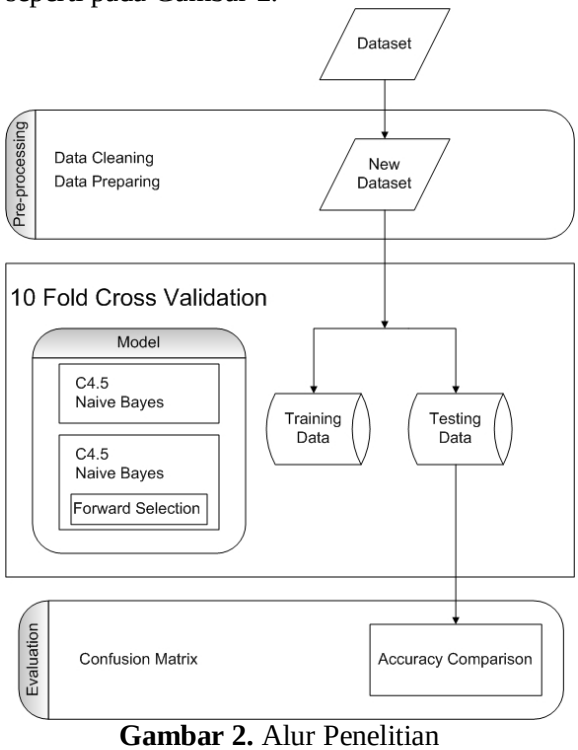
Metode analisis data dalam penelitian ini mengacu pada tahapan proses CRISP-DM. CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standar dalam *data mining* yang dapat diaplikasikan di berbagai sektor industri. Gambar 1 menjelaskan tentang siklus hidup pengembangan *data mining* yang telah ditetapkan dalam CRISP-DM.



Gambar 1. Enam Tahap Proses CRISP-DM dalam Data Mining

3.3. Alur Penelitian

Secara umum alur penelitian yang dilakukan mengacu pada kerangka penelitian seperti pada Gambar 2.



Gambar 2. Alur Penelitian

4. HASIL DAN PEMBAHASAN

4.1. Penentuan Dataset Mahasiswa

Langkah pertama proses klasifikasi diawali dengan penentuan dataset yang disimpan dalam format excel (*.xls) seperti yang terlihat pada Gambar 3.

NO	GENDER	JURUSAN SEKOLAH ASAL	IPS 1	IPS 2	IPS 3	IPS 4	IPS 5	IPS 6	KONSENTRASI KEAHLIAN	LAMA STUDI	KESIMPULAN
1	Pria	IPA	2,76	2,8	2,3	2,47	2,65	2,8	SIE	Terlambat	Pilihan Kurang Tepat
2	Pria	SMK	3,12	3,38	3,14	3,1	3,2	3,4	RPLA	Tepat Waktu	Pilihan Tepat
3	Pria	IPS	2,8	2,7	3	2,9	3	2,8	SIE	Terlambat	Pilihan Kurang Tepat
4	Pria	IPA	2,6	2,8	2,5	2,9	2,8	2,9	SIE	Terlambat	Pilihan Kurang Tepat
5	Wanita	IPA	2,9	2,89	2,96	2,85	3,2	3,1	SIE	Terlambat	Pilihan Tepat
6	Pria	IPS	2,95	2,82	3,4	3	3,2	3,1	RPLA	Terlambat	Pilihan Tepat
7	Pria	IPS	2,76	2,87	2,6	2,95	2,7	2,9	SIE	Terlambat	Pilihan Kurang Tepat
8	Pria	IPA	2,62	2,89	2,32	2,5	2,43	2,7	SIE	Terlambat	Pilihan Kurang Tepat
9	Wanita	IPA	3	3,17	3	2,98	3,3	3,2	RPLA	Tepat Waktu	Pilihan Tepat
10	Wanita	IPA	3,4	3,56	3,47	3,21	3,4	3,6	SIE	Tepat Waktu	Pilihan Tepat
11	Pria	IPA	2,1	2,3	2,1	2,8	2,71	2,9	SIE	Terlambat	Pilihan Kurang Tepat
12	Pria	IPA	3,1	3,3	3,2	3	3,2	3,4	SIE	Terlambat	Pilihan Tepat
13	Pria	IPA	2,57	2,82	2,8	3	2,8	3,1	SIM	Terlambat	Pilihan Kurang Tepat
14	Wanita	IPA	3,6	3,5	3,1	3,4	3,4	3,6	SIE	Tepat Waktu	Pilihan Tepat
15	Pria	IPS	2,8	2,7	2,1	2	2,3	1,9	SIE	Terlambat	Pilihan Kurang Tepat
16	Pria	IPS	2,86	2,86	2,45	2,6	2,4	2,6	SIE	Terlambat	Pilihan Kurang Tepat
17	Pria	SMK	2,71	3,9	2,54	3,36	2,9	2,7	SIM	Terlambat	Pilihan Kurang Tepat
18	Pria	SMK	2,67	2,2	3,7	3,74	3,5	3,6	SIE	Tepat Waktu	Pilihan Tepat
19	Pria	IPS	2,67	2,68	2,95	2,61	2,5	2,4	SIE	Terlambat	Pilihan Kurang Tepat
20	Pria	IPS	3,1	3,71	2,96	3,4	3,1	2,9	SIE	Terlambat	Pilihan Kurang Tepat
21	Pria	SMK	3,8	3,9	3,9	4	4	4	RPLA	Tepat Waktu	Pilihan Tepat
22	Pria	IPA	3,12	3,23	2,96	3	3,1	3,3	SIE	Tepat Waktu	Pilihan Tepat
23	Pria	IPS	2,9	3,32	2,83	2,91	2,9	3	SIE	Terlambat	Pilihan Kurang Tepat
24	Wanita	IPA	2,95	3,41	3,48	3,24	3,25	3,47	SIE	Tepat Waktu	Pilihan Tepat
25	Pria	IPS	3,05	2,92	2,27	2,65	2,68	2,71	SIE	Terlambat	Pilihan Kurang Tepat
26	Pria	SMK	2,64	3	2,9	2,77	2,43	2,79	SIE	Terlambat	Pilihan Kurang Tepat
27	Wanita	IPA	2,45	2,45	2,14	2,95	3,05	3,28	SIE	Tepat Waktu	Pilihan Tepat
28	Wanita	IPA	3,3	2,9	3,16	2,35	2,98	3,38	SIE	Terlambat	Pilihan Tepat
29	Pria	IPS	2,33	2,63	3,29	3,29	2,95	2,87	SIE	Tepat Waktu	Pilihan Tepat
30	Pria	IPS	2,52	2,7	2,84	2,77	2,69	2,94	SIE	Terlambat	Pilihan Kurang Tepat
31	Pria	IPS	3,06	3,38	3,21	2,98	3,08	2,87	SIM	Terlambat	Pilihan Kurang Tepat
32	Pria	IPS	2,81	3,09	2,63	3,36	3,17	2,98	SIE	Terlambat	Pilihan Tepat
33	Pria	IPA	2,52	2,75	2,25	2,5	2,87	3,15	SIE	Tepat Waktu	Pilihan Tepat
34	Pria	IPS	3,12	3,1	2,73	2,8	2,33	1,97	SIE	Terlambat	Pilihan Kurang Tepat
35	Pria	IPS	2,86	3,05	2,54	2,36	2,78	3,31	SIM	Terlambat	Pilihan Tepat
36	Pria	SMK	3,05	3,27	2,31	3,05	2,97	3,41	SIM	Terlambat	Pilihan Tepat
37	Pria	IPA	2,71	3,16	2,44	2,91	3,11	3,24	SIE	Terlambat	Pilihan Tepat
38	Pria	IPS	2,88	3,32	2,48	3,27	2,53	2,41	SIE	Terlambat	Pilihan Kurang Tepat
39	Pria	IPA	2,76	3,25	2,98	2,91	2,92	3,13	SIM	Terlambat	Pilihan Tepat
40	Pria	IPS	2,69	2,59	3,2	3	3,16	2,95	SIE	Terlambat	Pilihan Tepat
41	Pria	IPS	2,29	3,09	3,2	2,74	3,14	3,25	SIE	Terlambat	Pilihan Tepat
42	Pria	IPS	2,71	2,59	3,09	2,76	2,81	2,75	SIM	Terlambat	Pilihan Kurang Tepat
43	Pria	IPS	2,43	3,02	3	2,79	2,56	2,91	SIE	Terlambat	Pilihan Kurang Tepat
44	Wanita	IPA	2,88	3,59	2,82	3,05	3,42	3,67	SIE	Terlambat	Pilihan Tepat
45	Wanita	IPA	2,07	2,59	2,59	2,48	2,88	3,21	SIE	Terlambat	Pilihan Tepat
46	Pria	SMK	3,1	3	2,98	3,1	2,83	3,27	SIE	Tepat Waktu	Pilihan Tepat
47	Pria	IPA	2,57	2,55	2,57	2,91	3,22	3,42	SIE	Terlambat	Pilihan Tepat
48	Wanita	SMK	3,26	3,73	3,02	3,5	3,75	3,65	SIE	Tepat Waktu	Pilihan Tepat
49	Pria	IPS	2,71	2,89	2,75	3,18	2,74	2,81	SIM	Terlambat	Pilihan Kurang Tepat
50	Pria	IPA	2,9	2,91	2,59	2,28	3,11	2,95	SIM	Tepat Waktu	Pilihan Tepat

Gambar 3. Potongan Data Lulusan Mahasiswa

Tabel 1 merupakan pembagian variabel dan kelas data yang digunakan dalam analisis data mining.

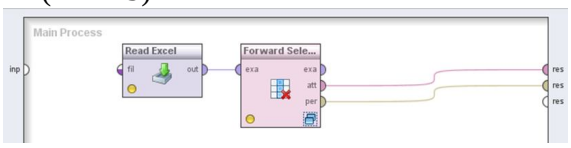
Tabel 1. Pembagian Variabel dan Kelas Data

Variabel	Nama Field	Jenis Kelas Data	Kelas Data yang Digunakan
Y	Kesimpulan	Binomial	Pilihan Tepat ; Pilihan Kurang Tepat
X1	Gender	Binomial	Pria ; Wanita
X2	Jurusan Sekolah Asal	Polynomial	IPA ; IPS ; SMK
X3	IPS 1	Numeric	0 s/d 4,00
X4	IPS 2	Numeric	0 s/d 4,00
X5	IPS 3	Numeric	0 s/d 4,00
X6	IPS 4	Numeric	0 s/d 4,00
X7	IPS 5	Numeric	0 s/d 4,00
X8	IPS 6	Numeric	0 s/d 4,00
X9	Lama Studi	Binomial	Tepat Waktu ; Terlambat
X10	Konsentrasi Keahlian	Polynomial	SIE ; RPLA ; SJM

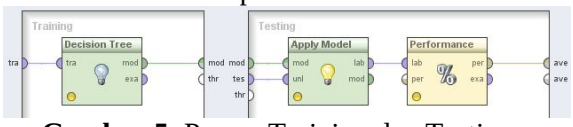
4.2. Implementasi dengan RapidMiner

Kombinasi metode seleksi fitur dilakukan untuk mendapatkan akurasi yang tinggi. Kombinasi metode yang dilakukan pada penelitian ini yaitu :

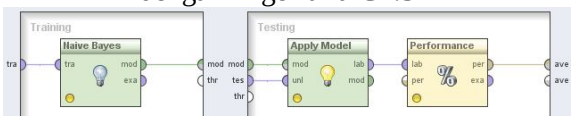
a. Klasifikasi dataset menggunakan algoritma C4.5 dan metode forward selection (DT+FS)
b. Klasifikasi dataset menggunakan algoritma naive bayes dan metode forward selection (NB+FS)



Gambar 4. Tampilan Menu Preprocessing di RapidMiner



Gambar 5. Proses Training dan Testing dengan Algoritma C4.5



Gambar 6. Proses Training dan Testing dengan Algoritma Naive Bayes

Gambar 7. Proses Validasi Menggunakan 10-Fold Cross Validation

4.3. Analisis Hasil

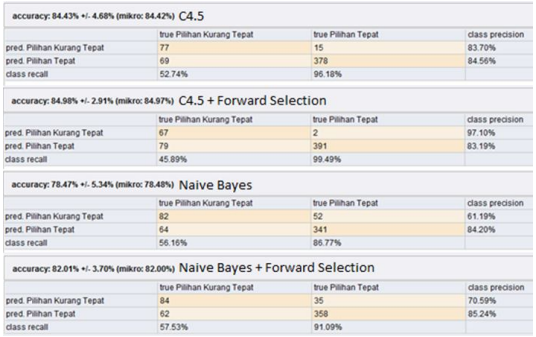
Percobaan dilakukan untuk mengetahui tingkat akurasi dari algoritma C4.5 dan algoritma *naive bayes* dengan menambahkan metode *forward selection* yang dilakukan pada dataset mahasiswa sebanyak 539 alumni. Pada Tabel 2 menunjukkan bahwa penggunaan *forward selection* meningkatkan akurasi algoritma C4.5 mencapai 84,98% sedangkan pada algoritma *naive bayes* mencapai 82,01%.

Tabel 2. Hasil Komparasi Algoritma C4.5 dan Naive Bayes

Model	Accuracy

C4.5	84,43%
C4.5 + Forward Selection	84,98%
Naive Bayes	78,47%
Naive Bayes + Forward Selection	82,01%

Penambahan seleksi fitur *forward selection* menghasilkan tingkat akurasi yang lebih baik dibandingkan dengan tanpa penambahan seleksi fitur yang hanya mencapai tingkat akurasi sebesar 84,43% pada algoritma C4.5 dan pada algoritma *naive bayes* sebesar 78,47%.



accuracy: 84.43% +/- 4.68% (mikro: 84.42%) C4.5

	true Pilihan Kurang Tepat	true Pilihan Tepat	class precision
pred. Pilihan Kurang Tepat	77	15	83.70%
pred. Pilihan Tepat	69	378	84.56%
class recall	52.74%	96.18%	

accuracy: 84.98% +/- 2.91% (mikro: 84.97%) C4.5 + Forward Selection

	true Pilihan Kurang Tepat	true Pilihan Tepat	class precision
pred. Pilihan Kurang Tepat	67	2	97.10%
pred. Pilihan Tepat	79	391	83.19%
class recall	45.89%	99.49%	

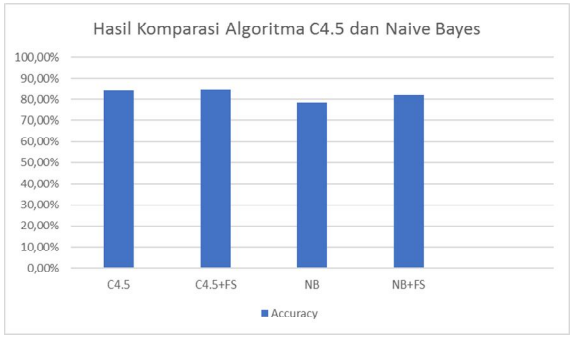
accuracy: 78.47% +/- 5.34% (mikro: 78.48%) Naive Bayes

	true Pilihan Kurang Tepat	true Pilihan Tepat	class precision
pred. Pilihan Kurang Tepat	82	52	61.19%
pred. Pilihan Tepat	64	341	84.20%
class recall	56.16%	86.77%	

accuracy: 82.01% +/- 3.70% (mikro: 82.00%) Naive Bayes + Forward Selection

	true Pilihan Kurang Tepat	true Pilihan Tepat	class precision
pred. Pilihan Kurang Tepat	84	35	70.59%
pred. Pilihan Tepat	62	358	85.24%
class recall	57.53%	91.09%	

Gambar 8. Screenshot Hasil Uji Algoritma C4.5 dan Naive Bayes



Gambar 9. Grafik Hasil Komparasi Algoritma C4.5 dan Naive Bayes

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil implementasi algoritma C4.5 dan Naive Bayes pada kasus ketepatan pemilihan konsentrasi mahasiswa dapat diambil beberapa kesimpulan sebagai berikut :

a. Dengan dataset yang sama, penggunaan seleksi fitur forward selection pada algoritma C4.5 dan Naive Bayes pada kasus ketepatan pemilihan konsentrasi mahasiswa dengan melibatkan atribut-atribut : jurusan sekolah asal, gender dan nilai indeks prestasi semester, lama studi mampu meningkatkan hasil akurasi.
b. Hasil uji kinerja algoritma klasifikasi untuk kasus ketepatan pemilihan konsentrasi mahasiswa untuk algoritma C4.5 tanpa penambahan seleksi fitur forward selection diperoleh nilai akurasi sebesar 84,43%, kemudian setelah ditambahkan seleksi fitur forward selection meningkat menjadi 84,98%. Sedangkan pada algoritma Naive Bayes tanpa penambahan seleksi fitur forward selection diperoleh nilai akurasi sebesar 78,47%, setelah ditambahkan seleksi fitur forward selection meningkat menjadi 82,01%.
c. Kinerja antara algoritma C4.5 tanpa penambahan seleksi fitur forward selection dengan algoritma C4.5 ditambah seleksi fitur forward selection lebih unggul bila dibandingkan dengan algoritma Naive Bayes pada kasus ketepatan pemilihan konsentrasi mahasiswa.

5.2. Saran

Berdasarkan kesimpulan yang di dapat, maka dapat diberikan saran sebagai berikut :

a. Hasil penelitian ini perlu diimplementasikan menjadi perangkat lunak yang dapat digunakan oleh pihak pengelola program studi terkait untuk membantu mahasiswa dalam mengambil keputusan pada pemilihan konsentrasi keahlian yang ditawarkan yang bertujuan agar lebih terarah sehingga mahasiswa dapat lebih memaksimalkan potensi diri serta dapat menyelesaikan studi dengan tepat waktu.
b. Untuk penelitian mendatang, pemilihan algoritma klasifikasi selain C4.5 dan Naive Bayes, seperti k-NN, Support Vector Machine, serta penggunaan teknik optimasi dengan metode seleksi fitur selain forward selection baik pada dataset maupun pada algoritma klasifikasi perlu diteliti agar diperoleh hasil akurasi yang lebih baik daripada penelitian ini. Penggunaan variabel yang lain juga memungkinkan untuk kasus yang sama, misalnya atribut status pekerjaan bagi mahasiswa kelas karyawan.

6. DAFTAR PUSTAKA

Anonim, 2012, Panduan Akademik Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta 2013/2014, Surakarta

Chapman, P., 2000, CRISP-DM 1.0: Step-by-step Data Mining Guide, SPSS Inc

Draper, N., Smith, H., 1992, Analisis Regresi Terapan Edisi Kedua, PT. Gramedia Pustaka Utama, Jakarta

Domingos, P., Pazzani, M., 1997, On The Optimality of The Simple Bayesian Classifier Under Zero-One Loss

Han, J., Kamber, M., 2006, Data Mining: Concepts and Techniques Second Edition, Morgan Kaufmann, New York

Larose, D. T., 2005, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons Inc

Olson, D., Delen, D., 2008, Advanced Data Mining Techniques, Springer, USA

Prasetyo, E., 2012, Data Mining Konsep dan Aplikasi Menggunakan Matlab, Penerbit Andi, Yogyakarta

Ashari, A., Paryudi, I., Tjoa, A. M., 2013, Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 4 No. 11

Da Rocha, Timoteo, R., 2010, Identifying Bank Frauds Using CRISP-DM and Decision Tree, International Journal of Computer Science & Information Technology pp.162-169

Dimitoglou, G., Adams, J. A., Jim, C. M., 2012, Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability, Journal of Computing Press, ISSN (online): 2151-9617, Vol. 4 Issue 8 August 2012, New York USA

Hastuti, K., 2012, Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif, Seminar Nasional Teknologi Informasi & Komunikasi Terapan (Semantik), Semarang

Patil, T. R., Sherekar, S. S., 2013, Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal Of Computer Science And Applications, ISSN: 0974-1011 (Open Access), Vol. 6 No.2 Apr 2013

Santra, A. K., Jayasudha, S., 2012, Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification, International Journal of Computer Science Issues (IJCSI), ISSN (online) : 1694-0814, Vol. 9 Issue.1 No. 2, January 2012

Susanto, H., Sudiyanto, 2014, Data Mining untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan dan Prestasi Masa Lalu, Jurnal Pendidikan Vokasi, Vol. 4 No. 2 Juni 2014

RapidMiner, 16 Maret 2016, RapidMiner Documentation, http://docs.rapidminer.com/

Wahono, R., S., 20 Januari 2016, Data Mining, http://romisatriawahono.net/dm