ALGORITMA NAIVE BAYES DENGAN FITUR SELEKSI UNTUK MENGETAHUI HUBUNGAN VARIABEL NILAI DAN LATAR BELAKANG PENDIDIKAN

Yani Parti Astuti
Fakultas Ilmu Komputer, Program Studi Teknik Informatika
Universitas Dian Nuswantoro
Email: yanipartiastuti@dsn.dinus.ac.id

Usman Sudibyo
Fakultas Ilmu Komputer, Program Studi Teknik Informatika
Universitas Dian Nuswantoro
Email: usmansudibyo@gmail.com

Achmad Wahid Kurniawan
Fakultas Ilmu Komputer, Program Studi Teknik Informatika
Universitas Dian Nuswantoro
Email: wahid@dsn.dinus.ac.id

Yuniarsi Rahayu
Fakultas Ilmu Komputer, Program Studi Teknik Informatika
Universitas Dian Nuswantoro
Email: yuniarsi.rahayu@dsn.dinus.ac.id

ABSTRAK

Setiap Perguruan Tinggi mempunyai mahasiswa baru yang berasal dari berbagai sekolah menengah atas dan juga sekolah menengah kejuruan. Seperti halnya pada program studi teknik informatika fakultas ilmu komputer di Universitas Dian Nuswantoro. Program studi ini mempunyai mahasiswa terbanyak di Udinus, sehingga perlu selalu diadakan evaluasi. Dalam hal ini evaluasi yang dipilih adalah tentang asal jurusan sekolah mahasiswa dengan variabel nilai mata kuliah. Dengan mengambil mahasiswa dari angkatan tahun 2010 sampai 2012 sebanyak 10030 mahasiswa, hanya 489 mahasiswa yang mengisi asal jurusan sekolah. Dari sejumlah mahasiswa tersebut dilakukan preposisi dengan mengambil nilai mata kuliah wajib sebanyak 25 mata kuliah dan asal jurusan sekolah. Teknik data mining berupa algoritma *naive bayes* dioptimasi dengan fitur selesi *forward selection* telah meningkatkan akurasi dalam penemuan pola klasifikasi. Peningkatan akurasi dari *naive bayes* 64,77% menjadi 78,08% setelah dioptimasi dengan *forward selection*. Dengan demikian hasil klasifikasi tersebut bisa digunakan sebagai informasi dalam metode pembelajaran yang bisa diterapkan.

Kata kunci: *data mining, forward selection, naïve bayes.*

ABSTRACT

Each university has new students coming from various high schools as well as vocational high schools. As well as in the study program of informatics engineering faculty of computer science at Dian Nuswantoro University. This study program has the highest number of students in Udinus, so it is always necessary to evaluate. In this case the evaluation that choose about the origin of the student department with the value of the course variable. By taking students from 2010 to 2012 as many as 10030 students, only 489 students who fill the origin of the school. Of the number of students are done preposition by taking the compulsory courses as much as 25 courses and the origin of school majors. Data mining techniques in the form of naive bayes algorithm optimized with forward selection feature have improved accuracy. Increased accuracy of 64.77% naive bayes to 78.08% after optimization with forward selection. Thus the classification results can be used as information in the methods of learning that can be applied.

Keywords: *data mining, classification, naïve bayes.*

1. PENDAHULUAN

Pada setiap perguruan tinggi harus bisa mengolah database dengan baik agar bisa digunakan sebagai peningkatan mutu pendidikan dengan mempelajari data tersebut. Salah satu caranya yaitu dengan mengetahui dan mempelajari atribut – atribut yang mempengaruhi kinerja mahasiswa [1]. Atribut latar belakang atau asal jurusan sekolah mahasiswa adalah atribut penting karena bisa berpengaruh terhadap nilai mata kuliah [2]. Dengan demikian penting untuk dipelajari seberapa besar peranan latar belakang sekolah dengan nilai mata kuliah khususnya mata kuliah wajib. Hal itu diperuntukkan agar setiap perguruan tinggi mencermati atribut mahasiswa yang bisa dijadikan analisa untuk meningkatkan performa mahasiswa melalui keunikan atribut mahasiswa [3].

Untuk mengetahui seberapa besar pengaruh asal jurusan sekolah dengan variable nilai tersebut dapat digunakan dengan menggunakan data-data mahasiswa tentang asal jurusan sekolah dengan nilai mata kuliah yang telah didapatkan kemudian diproses yang akhirnya menemukan suatu pola yang mengandung arti berupa mode simpan menggunakan teknik pengenalan pola yaitu *data mining* [4].

Data mining merupakan suatu cara untuk pengenalan pola berdasarkan teknik matematika dan juga statistika dalam mengolah dan menggali sejumlah data [5]. Salah satu teknik data mining adalah klasifikasi. Proses yang digunakan untuk menghasilkan model serta fungsi yang membedakan kelas data disebut klasifikasi [6].

Mohammed M. Abu Tair [7] menggunakan Algoritma Naïve bayes dan decision tree untuk mengetahui performa mahasiswa pascasarjana dalam mengatasi nilai rendah mahasiswa pascasarjana.

Khafizh Hastuti [8] melakukan penelitian menggunakan algoritma naïve bayes, neural network dan decision tree untuk meneliti pada performa siswa dengan menggunakan 300 siswa yang didapat dari 5 perguruan tinggi yang sederajat dan berbeda pada Bachelor of Computer Application (BCA). Klasifikasi yang digunakan adalah metode Bayesian dengan menggunakan 17 atribut dan didapatkan faktor misalnya pada hasil ujian SLTA, tempat tinggal mahasiswa , media pembelajaran, kualifikasi ibu, aktivitas kebiasaan – kebiasaan mahasiswa, pendapatan keluarga setiap tahunnya, dan tentang status keluarga mahasiswa. Semua itu dianggap sangat mempengaruhi prestasi akademik mahasiswa.,

Berdasarkan penelitian – penelitian di atas maka penulis akan menggunakan Algoritma Naïve Bayes sebagai teknik klasifikasi. Algoritma Naive Bayes digunakan dengan pertimbangan bahwa Naïve Bayes adalah salah satu teknik prediksi berbasis probabilistik sederhana berdasarkan pada penggunaan aturan atau teorema Bayes. Algoritma Naïve Bayes juga dikenal dengan penggunaan asumsi kebebasan fitur yang kuat, berarti bahwa fitur pada data tidak ada keterkaitan dengan ada maupun tidaknya fitur lain dalam data yang tidak berbeda [9]. Salah satu algoritma klasifikasi yang mempunyai kecepatan komputasi sangat tinggi adalah Naïve Bayes . Naïve Bayes juga bisa memecahkan masalah dataset mempunyai dimensi besar [10]. Namun Naïve Bayes memiliki kelemahan yaitu tidak adanya keterkaitan antar atribut sehingga kurang meningkatkan kinerja dari Naïve Bayes tersebut [9][10].

Untuk meningkatkan kinerja Naïve Bayes maka diperlukan adanya optimasi. Optimasi yang digunakan berupa fitur seleksi. Contoh teknik fitur seleksi yaitu Backward Elimination dan Forward Selection. Dalam fitur seleksi forward selection fitur akan diseleksi satu per satu dan akan dihilangkan fitur yang tidak relevan [5].

2. METODOLOGI PENELITIAN

Data yang berkualitas dapat didapat dengan melakukan teknik [11]:

a. Pengumpulan Data
   Data akan diproses pada tahap ini. Data yang diperoleh mempunyai beberapa atribut dan terdiri dari *record - record*. Data tersebut akan diintegrasikan dan dijadikan menjadi sebuah dataset yang siap diproses.
b. Pengolahan awal data
   Data akan diolah pada tahap ini. Data akan diseleksi dan dibersihkan dari *noise*. Data – data yang kosong akan dihilangkan dan selanjutnya ditransformasi untuk mendapatkan model. Pada tahap ini juga akan dilakukan untuk mempersiapkan data yang betul - betuk valid sebelum dilakukan proses berikutnya.
c. Eksperimen dan pengujian model
   Model yang digunakan akan diuji pada tahap ini yang digunakan untuk melihat hasil yang bisa digunakan sebagai pengambilan keputusan.
d. Evaluasi dan validasi hasil
   Evaluasi pada model dilakukan pada tahap ini sebagai hasil untuk mengetahui tingkat performa (akurasi) model.

2.1 Naïve Bayes

Naïve Bayes ialah suatu penerapan dari teorema *Bayesian*. Algoritma *Naïve Bayes* juga didasarkan pada suatu asumsi yang digunakan untuk menyederhanakan atribut dengan mengkondisikan setiap atribut tidak terkait satu sama lain [11]. *Naïve Bayes* juga bisa digolongkan pengklasifikasian menggunakan aturan statistika yang dapat bermanfaat bagi prediksi probabilitas dalam anggota kelas [5]. *Naïve bayes* memiliki tingkat akurasi yang tergolong tinggi dan kecepatan komputasi yang relative cepat saat digunakan data berdimensi besar.

Dalam *Naïve Bayes* diartikan bahwa tidak ada keterkaitan antar fitur. Pada perhitungan *Naïve bayes* didasari dengan menggunakan teorema *Bayesian*: [9]

$$P(X|H) = \frac{P(X|H)*P(H)}{P(X)} \quad (1)$$

Keterangan:
X: data di mana kelasnya belum diketahui
H: Hipotesis data *x* menjadi kelas spesifik
P(X |H): Probabilitas bahwa *X* yang terjadi akan mempengaruhi nilai *H*
P(H): Probabilitas(peluang) awal pada hipotesis *H* tanpa memandang suatu bukti lain
P(X): Probabilitas(peluang) awal *X* yang terjadi dengan tidak melihat hipotesis lain

Formula *Naïve Bayes* yang digunakan sebagai klasifikasi adalah

$$P(Y|X) = \frac{P(Y)\Pi_{i=1}^{q} P(X_i|Y)}{P(X)} \quad (2)$$

P(Y|X) adalah suatu probabilitas dengan data yang menggunakan variabel *X* dengan label *Y*. *P(Y)* adalah probabilitas awal label *Y*. $\Pi_{i=1}^{q} P(X_i|Y)$ adalah probabilitas variable terikat atau *independen* dalam label *Y* dari fitur variabel *X*. Nilai *P(X)* akan selalu tetap di dalam hitungan prediksi dan akhirnya hanya dihitung bagian $P(Y)\Pi_{i=1}^{q} P(X_i|Y)$ dengan memilih bagian yang paling besar sebagai kelas hasil prediksi. Sedangkan probabilitas terikat atau independen $\Pi_{i=1}^{q} P(X_i|Y)$ adalah pengaruh dari setiap fitur data pada setiap kelas *Y*, dan dinotasikan sebagai

$$P(X|Y=y) = \Pi_{i=1}^{q} P(X_i|Y = y) \quad (3)$$

Semua kelompok fitur $X = \{X_1, X_2, X_3, \ldots, Xq\}$terdiri dari *q* atribut atau *q* dimensi

2.2 Forward Selection

Metode *Sequential Forward Selection* atau metode seleksi maju adalah algoritma pencarian paling sederhana. *Forward Selection* didasarkan pada model *Regresi Linear*. *Forward Selection* adalah salah satu teknik untuk mereduksi dimensi dataset untuk menghapus atribut-atribut yang tidak relevan [17]. Metode *Forward Selection* merupakan model yang diawali dengan nol variable, untuk selanjutnya variable dimasukkan satu persatu sampai pada kriterianya terpenuhi. Prosedur *Forward Selection* dimulai dengan tidak ada variabel dalam model [23]. Dimulai dari himpunan kosong, dilakukan berurutan dengan menambahkan fitur *x +* yang menghasilkan tertinggi obyektif fungsi *J* $(Y_k + x+)$ bila dikombinasikan dengan fitur Y_k yang telah dipilih [21].
Metode *Forward Selection:*

a. Dimulai dari himpunan kosong $Y_0 = \{?\}$
b. Pilih fitur terbaik berikutnya $x+ = argmax\ [J\ (y_k + x)]; x \notin y_k$
c. *Update* $y_k + 1 = y_k + x+, k = k + 1$
d. Kembali ke langkah ke-2

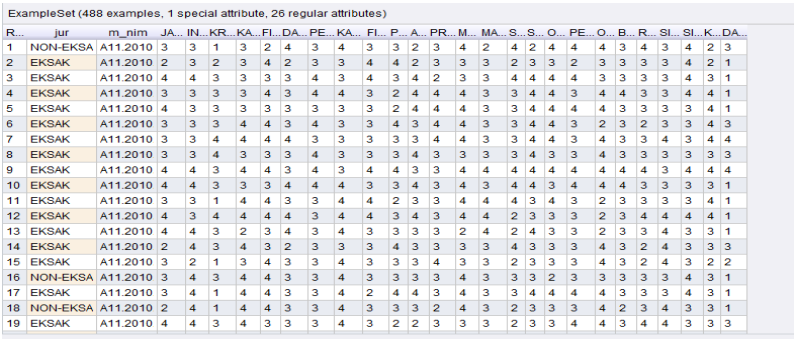
Tahapan metode *Forward Selection – Naïve Bayes* adalah sebagai berikut [4]:

a. Untuk variabel pertama yang memasuki model, pilih predictor yang paling berkorelasi dengan target. Jika model yang didapatkan tidak signifikan, maka berhenti dan melaporkan bahwa tidak ada predictor yang penting. Jika tidak lanjutkan ke langkah 2
b. Untuk setiap variabel yang tersisa, hitung nilai F-statistika untuk variabel yang sudah siap diberikan pada model. Misalkan pada langkah satu berhasil, tentukan F(x2/x1), F(x3/x1), F(x4/x1). Pada keberhasilan langkah kedua, mungkin bisa F(x3/x1, x2), F(x4/x1, x2). Urutkan variable dengan F-statistika terbesar

c. Untuk variabel yang terpilih pada langkah 2, F-statistika yang signifikan. Jika tidak ada yang signifikan berhenti dan melaporkan bahwa tidak ada penambahan variable pada model dari langkah 2. Jika tidak, maka tambahkan variable yang didapat dari langkah 2 dan kembali ke langkah 2.

3. HASIL DAN PEMBAHASAN

Data yang didapat dari pengolahan awal data selanjutnya akan diuji dengan kevalidan data dengan menggunakan cross validasi k = 10. Sebelum data dicross validasi maka atribut akan direduksi menggunakan fitur seleksi *forward selection*. Data sebelum dilakukan fitur seleksi adalah :

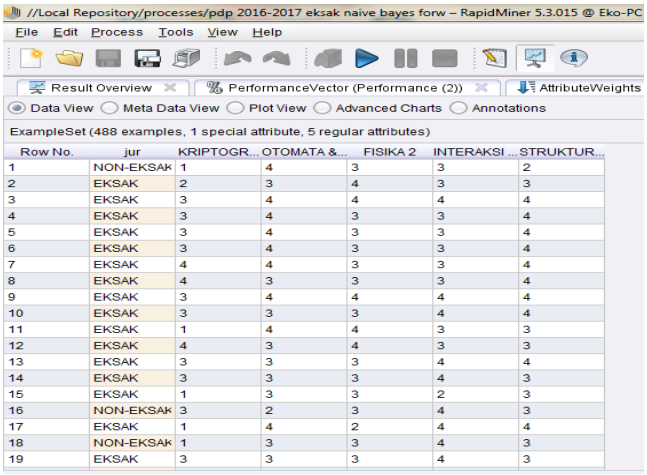


ExampleSet (488 examples, 1 special attribute, 26 regular attributes)

R...	jur	m_nim	JA...	IN...	KR...	KA...	FI...	DA...	PE...	KA...	FI...	P...	A...	PR...	M...	MA...	S...	S...	O...	PE...	O...	B...	R...	SI...	SI...	K...	DA...
1	NON-EKSA	A11.2010	3	3	1	3	2	4	3	4	3	3	2	3	4	2	4	2	4	4	4	3	4	3	4	2	3
2	EKSAK	A11.2010	2	3	2	3	4	2	3	3	4	4	2	3	3	3	2	3	3	2	3	3	3	3	4	2	1
3	EKSAK	A11.2010	4	4	3	3	3	3	4	3	4	3	4	2	3	3	4	4	4	4	3	3	3	3	4	3	1
4	EKSAK	A11.2010	3	3	3	3	4	3	4	4	3	2	4	4	4	3	3	4	4	3	4	4	3	3	4	4	1
5	EKSAK	A11.2010	4	3	3	3	3	3	3	3	3	2	4	4	4	3	3	4	4	4	4	3	3	3	3	4	1
6	EKSAK	A11.2010	3	3	3	4	4	3	4	3	3	4	3	4	4	3	3	4	4	3	2	3	2	3	3	4	3
7	EKSAK	A11.2010	3	3	4	4	4	4	3	3	3	3	3	4	4	3	3	4	4	3	4	3	3	4	3	4	4
8	EKSAK	A11.2010	3	3	4	3	3	3	4	3	3	3	4	3	3	3	3	4	3	3	4	3	3	3	3	3	3
9	EKSAK	A11.2010	4	4	3	4	4	3	4	4	4	3	3	4	4	4	4	4	4	4	4	4	4	3	4	4	4
10	EKSAK	A11.2010	4	4	3	3	3	4	4	4	3	3	4	3	4	3	4	4	3	4	4	4	3	3	3	3	1
11	EKSAK	A11.2010	3	3	1	4	4	3	3	4	4	2	3	3	4	4	4	3	4	3	2	3	3	3	3	4	1
12	EKSAK	A11.2010	4	3	4	4	4	4	3	4	4	3	4	3	4	4	2	3	3	3	2	3	4	4	4	4	1
13	EKSAK	A11.2010	4	4	3	2	3	4	3	4	3	3	3	3	2	4	2	4	3	3	2	3	3	4	3	3	1
14	EKSAK	A11.2010	2	4	3	4	3	2	3	3	3	4	3	3	3	3	4	3	3	3	4	3	2	4	3	3	3
15	EKSAK	A11.2010	3	2	1	3	4	3	3	4	3	3	3	4	3	3	2	3	3	3	4	3	2	4	3	2	2
16	NON-EKSA	A11.2010	3	4	3	4	4	3	3	4	3	3	3	3	4	3	3	3	2	3	3	3	3	3	4	3	1
17	EKSAK	A11.2010	3	4	1	4	4	3	3	4	2	4	4	3	4	3	3	4	4	4	4	3	3	3	4	3	1
18	NON-EKSA	A11.2010	2	4	1	4	4	3	3	4	3	3	3	3	2	4	3	2	3	3	3	4	2	3	4	3	3	1
19	EKSAK	A11.2010	4	4	3	4	3	3	3	4	3	2	2	3	3	3	2	3	3	4	4	3	4	4	3	3	3

Gambar 1. Data Mahasiswa Sebelum Dilakukan Fitur Seleksi (Hasil Pengolahan Data)

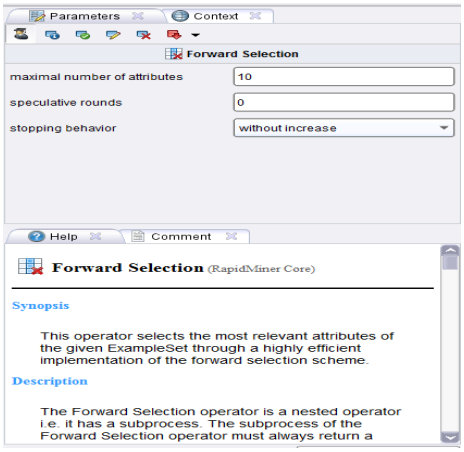
Selanjutnya data direduksi dengan fitur seleksi *forward selection* yang menghasilkan data reduksi sebagai berikut :



Row No.	jur	KRIPTOGR...	OTOMATA &...	FISIKA 2	INTERAKSI ...	STRUKTUR...
1	NON-EKSAK	1	4	3	3	2
2	EKSAK	2	3	4	3	3
3	EKSAK	3	4	4	4	4
4	EKSAK	3	4	3	3	4
5	EKSAK	3	4	3	3	4
6	EKSAK	3	4	3	3	4
7	EKSAK	4	4	3	3	4
8	EKSAK	4	3	3	3	4
9	EKSAK	3	4	4	4	4
10	EKSAK	3	3	3	4	4
11	EKSAK	1	4	4	3	3
12	EKSAK	4	3	4	3	3
13	EKSAK	3	3	3	4	4
14	EKSAK	3	3	3	4	3
15	EKSAK	1	3	3	2	3
16	NON-EKSAK	3	2	3	4	3
17	EKSAK	1	4	2	4	4
18	NON-EKSAK	1	3	3	4	3
19	EKSAK	3	3	3	4	3

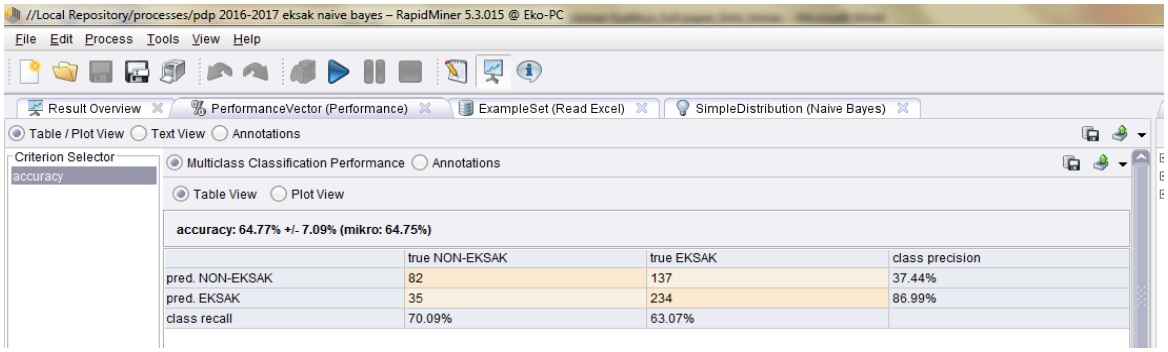
Gambar 2. Data Setelah Direduksi Dengan *Forward Selection* (Hasil Reduksi Dengan Aplikasi Rapidminer)

Berdasarkan gambar 1 dan 2 dapat dilihat bahwa ada pengurangan atribut dari 25 menjadi 6 atribut. Hal tersebut berarti 19 atribut yang hilang dikatakan tidak relevan. Keenam variable yang tersisa terdiri dari satu atribut kelas (*predictor*) dan lima atribut berupa variabel nilai yaitu Kriptografi, Teori Bahasa dan Otomata, Fisika 2, Interaksi dengan Komputer, struktur Data. Selanjutnya data diuji kevalidan dengan cross validasi k = 10 sebagai berikut :



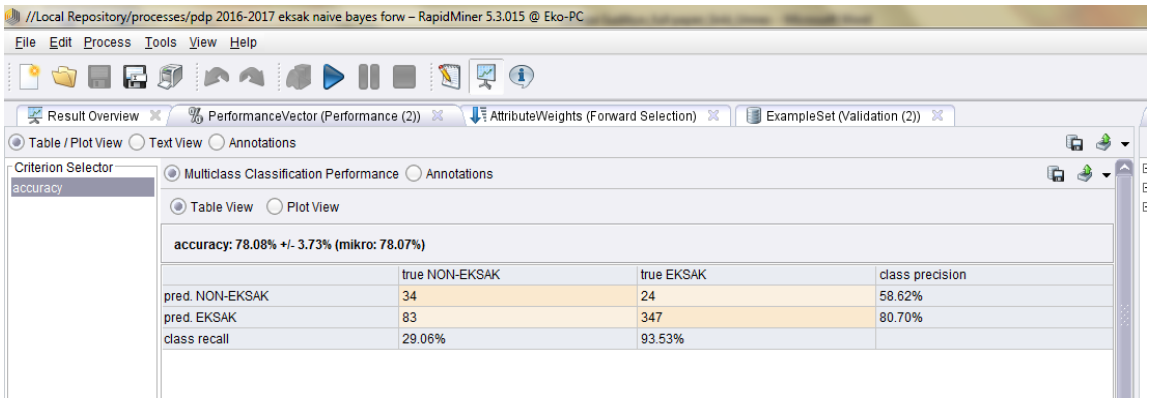
Gambar 3. Penggunaan Cross Validasi Dengan K = 10 (Hasil Aplikasi Ripedminer)

Data selanjutnya diuji dengan *naïve bayes* menggunakan tabel *confussion matrik*s dengan uraian sebagai berikut :



Gambar 4. *Confussion* Matriks Algoritma *Naïve Bayes* (Hasil Akurasi Dari Aplikasi Ripedminer)

Dari gambar 4 dapat dilihat hasil akurasi 64,77%. Ini berarti penggunaan algoritma naïve bayes belum optimal. Sehingga perlu dioptimasi dengan menggunakan fitur seleksi. Fitur seleksi yang digunakan adalah *forward selection* dengan hasil sebagai berikut :



Gambar 5. *Confussion* Matriks Algoritma *Naïve Bayes – Forward Selection*

Dari gambar 5 dapat dilihat hasil akurasi dari optimasi *forward selection* pada algoritma *naïve bayes* sebesar 78,08%. Artinya ada peningkatan performa dengan melakukan optimasi sebesar 13,31%. Dari hasil yang telah didapatkan maka dapat dibuat tabel sebagai berikut :

Tabel 1. Perbandingan hasil pengujian

Algoritma	*Waktu*	*Akurasi*
Naïve Bayes	1 detik	64,77%
Naïve Bayes – forward selection	6 detik	78,08%

4. KESIMPULAN

Dari semua hasil yang telah diuraikan, maka dapat diambil suatu kesimpulan bahwa tidak semua nilai mata kuliah dipengaruhi oleh asal jurusan sekolah atau latar belakang pendidikan mahasiswa. Hal ini dapat ditunjukkan dengan adanya kenaikan performa dari pengujian algoritma *naïve bayes* dengan pengujian optimasi *forward selection* terhadap algoritma *naïve bayes*. Kenaikan performa yang didapat sebesar 13,31%.

DAFTAR PUSTAKA

[1] Kurniawan, W Wicaksono, YP Astuti, 2016, *Educational Data Mining* (EDM) Untuk Memprediksi Keterlambatan Masa Studi Mahasiswa Menggunakan Algoritma C4.5, Momentum, , Hal. 48-52
[2] P Andriani, 2010, Pengaruh dan Asal Jurusan terhadap Hasil belajar Pengantar Dasar Matematika, Beta, hal. 118 – 133
[3] Fauzan, A, 2010, *Ide-ide penelitian pendidikan matematika*. Makalah disampaikan dalam Seminar Nasional Pendidikan Matematika, UIN Syarif Hidayatullah Jakarta.
[4] D. T. Larose, 2006, *Discovering Knowledge In Data*. United States of America: John Wiley & Sons, Inc.,.
[5] Larose, D. T., 2006, *Data Mining Methods and Models.* Hoboken, New Jersey, United State of America: John Wiley & Sons, Inc.,.
[6] Han, J., & Kamber, 2007, M., *Data Mining Concepts and Techniques* (2nd ed.). San Francisco, United State America: Morgan Kaufmann Publishers.
[7] Mohammed M. Abu Tair, 2012, Alaa M. El-Halees, *Mining Educational Data to Improve Students' Performance, International Journal of Information and Communication Technology Research*
[8] Khafiizh Hastuti , 2012, Analisis komparasi algoritma klasifikasi data mining Untuk prediksi mahasiswa non aktif, Seminar Nasional Teknologi Informasi & Komunikasi Terapan
[9] E. Prasetyo, 2012, *Data Mining: Konsep dan Aplikasi menggunakan MATLAB*, 1st ed. Yogyakarta, Indonesia: Andi.
[10] Bose, I. and Chen, X. ,2009, Quantitative models for direct marketing: A review from systems perspective, *European Journal of Operational Research*, Elsevier B.V., pp. 1–16.
[11] Budi Santoso, 2007, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, 1st ed. Yogyakarta, Indonesia.