PENERAPAN *DATA MINING* DENGAN METODE KLASIFIKASI *NAÏVE BAYES* UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DALAM MENGIKUTI *ENGLISH PROFICIENCY TEST* (Studi Kasus : Universitas Potensi Utama)

Alfa Saleh

Teknik Informatika Universitas Potensi Utama
Jl K.L. Yos Sudarso KM 6.5 No.3-A, Tanjung Mulia, Medan
Email : alfasoleh1@gmail.com

Abstrak

Universitas Potensi Utama merupakan salah satu Perguruan Tinggi Swasta (PTS) di bawah naungan Yayasan Potensi Utama yang bergerak dalam bidang pendidikan khususnya dalam bidang komputer. Tentu saja tidak hanya kualitas dalam ilmu komputer yang menjadi perhatian tetapi juga kompetensi dalam bahasa asing. Oleh karena itu setiap mahasiswa tingkat akhir harus mengikuti *English Profeciency Test* sebagai tolak ukur kemampuan mahasiswa dalam menguasai bahasa asing. 50 data mahasiswa yang mengikuti *English Proficiency Test* telah diuji dengan metode *Naïve Bayes*, didapatkan hasil persentase sebesar 98% untuk keakuratan klasifikasi. Diketahui dari 50 data yang diuji terdapat 49 data yang berhasil diklasifikasikan dengan benar.

Keywords: Data Mining, Naïve Bayes, English Proficiency Test.

1. Pendahuluan

Latar Belakang

Universitas Potensi Utama merupakan salah satu Perguruan Tinggi Swasta (PTS) di bawah naungan Yayasan Potensi Utama yang bergerak dalam bidang pendidikan. Awalnya Universitas Potensi Utama adalah STMIK yang berdiri pada tahun 2003 berdasarkan izin Dirjen Pendidikan Tinggi (DIKTI) dengan SK Nomor : 103/D/O/2003 dan dengan motto "Kami hadir untuk mencerdaskan kehidupan bangsa". Universitas Potensi Utama membuka 3 Program Studi di bidang komputer untuk Fakultas Teknik dan Ilmu Komputer yaitu Teknik Informatika, Sistem Informasi dan Manajemen Informatika dengan Nilai Akreditasi Peringkat B dari Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT). menimbang hal ini tentu saja kualitas pendidikan menjadi prioritas utama guna menciptakan lulusan - lulusan yang kompeten. Tidak hanya di bidang komputer, namun kompetensi bahasa asing khususnya bahasa inggris juga menjadi perhatian, oleh karena itu setiap mahasiswa tingkat akhir harus mengikuti *English Proficiency Test* untuk menguji kompetensi mahasiswa dalam berbahasa inggris baik lisan maupun tulisan. Adapun untuk memprediksi kelulusan mahasiswa yang mengikuti *English Proficiency Test* maka implementasi metode klasifikasi *Naive Bayes* diharapkan mampu mengklasifikasikan kelulusan mahasiswa.

Pada penelitian sebelumnya metode *Naive Bayes* juga digunakan dalam memprediksi penyakit Dermatologi yang diabaikan tapi bahkan dapat menyebabkan kematian di mana metode *Naive Bayes* digunakan untuk mengenal pola data untuk mengungkap kemungkinan penyakit dermatologi[1]. Metode *Naive Bayes* juga dinilai berpotensi baik dalam mengklasifikasi dokumen dibandingkan metode pengklasifikasian yang lain dalam hal akurasi dan efisiensi komputasi [2].

Data Mining

Data Mining merupakan proses pengekstraksian informasi dari sekumpulan data yang sangat besar melalui penggunaan algoritma dan teknik penarikan dalam bidang statistik, pembelajaran mesin dan sistem manajemen basis data[3]. *Data Mining* adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting yang dapat dipakai untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya[4]. Definisi lain mengatakan *Data Mining* adalah kegiatan yang meliputi pengumpulan, pemakaian data hostoris untuk menemukan keteraturan, pola atau hubungan dalam data berukuran besar[5]. Dari beberapa definisi di atas dapat ditarik kesimpulan bahwa *Data Mining* merupakan proses ataupun kegiatan untuk mengumpulkan data yang berukuran besar kemudian mengekstraksi data tersebut menjadi informasi – informasi yang nantinya dapat digunakan.

Tahap-tahap *Data Mining*

Sebagai suatu rangkaian proses, *Data Mining* dapat dibagi menjadi beberapa tahap proses. Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*.
Tahap-tahap *Data Mining* adalah sebagai berikut[6]:
a. Pembersihan data *(Data Cleaning)*
Pembersihan data merupakan proses menghilang-kan *noise* dan data yang tidak konsisten atau data tidak relevan.
b. Integrasi data *(Data Integration)*
Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru.
c. Seleksi data *(Data Selection)*
Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.
d. Transformasi data *(Data Transformation)*
Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *Data Mining.*
e. Proses *Mining*
Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan *Data Mining.*
f. Evaluasi pola *(Pattern Evaluation)*
Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.
g. Presentasi pengetahuan *(Knowledge Presentation)*
Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

Metode Naive Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas[7]. Definisi lain mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya [8].
Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu[8]. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan[9].

Persamaan Metode Naive Bayes

Persamaan dari teorema *Bayes* adalah[8] :

$$P\ H\ X\ = \frac{P\ X\ H\ .P(H)}{P(X)} \qquad (1)$$

Di mana :
X :Data dengan *class* yang belum diketahui
H : Hipotesis data merupakan suatu *class* spesifik
P(H|X) :Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori probabilitas)
P(H) : Probabilitas hipotesis *H* (prior probabilitas)
P(X|H) :Probabilitas *X* berdasarkan kondisi pada hipotesis *H*
P(X) : Probabilitas *X*

Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naive Bayes* di atas disesuaikan sebagai berikut :

$$P\ C\ F1 \dots Fn\ = \frac{P\ C\ P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \qquad (2)$$

Di mana Variabel *C* merepresentasikan kelas, sementara variabel *F1 ... Fn* merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas *C* (*Posterior*) adalah peluang munculnya kelas *C* (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik karakteristik sampel pada kelas *C* (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik karakteristik sampel secara global (disebut juga *evidence*). Karena itu, rumus diatas dapat pula ditulis secara sederhana sebagai berikut :

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \qquad (3)$$

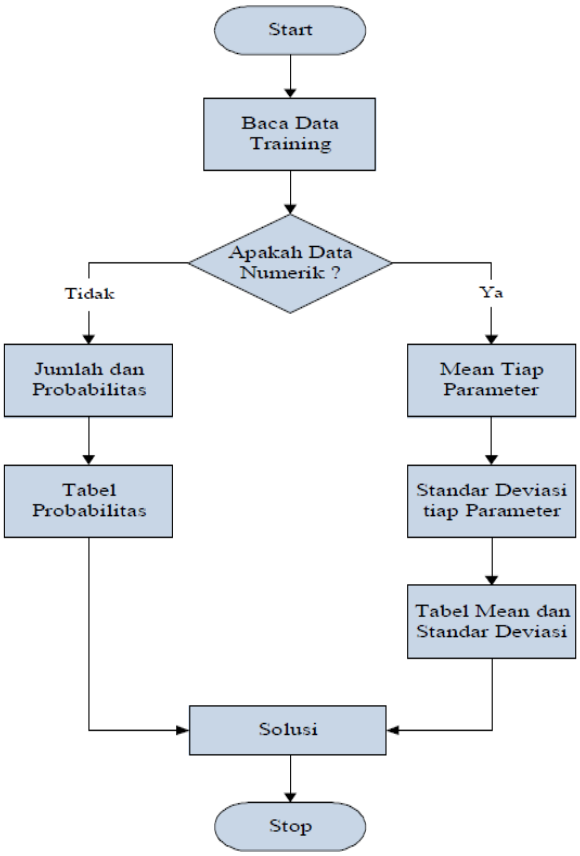
Nilai *Evidence* selalu tetap untuk setiap kelas pada satu sampel.Untuk klasifikasi dengan data kontinyu digunakan rumus *Densitas Gauss* :

$$P\ X_i = x_i\ Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2 ij}} \qquad (4)$$

Di mana :
P : Peluang
Xi : Atribut ke *i*
xi : Nilai atribut ke *i*
Y : Kelas yang dicari

yi : Sub kelas *Y* yang dicari
μ : *mean*, menyatakan rata – rata dari seluruh atribut
σ :Deviasi standar, menyatakan varian dari seluruh atribut.

Alur dari metode *Naive Bayes* dapat dilihat pada gambar 1 sebagai berikut :



Gambar 1. Alur Metode Naive Bayes

1. Baca *data training*
2. Hitung Jumlah dan probabilitas, namun apabila data numerik maka :
   a. Cari nilai *mean* dan standar deviasi dari masing masing parameter yang merupakan data numerik. Adapun persamaan yang digunakan untuk menghitung nilai rata – rata hitung (*mean*) dapat dilihat sebagai berikut :

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (5)$$

atau

$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \qquad (6)$$

di mana :
μ : rata – rata hitung (*mean*)
x_i : nilai sample ke -*i*
n : jumlah sampel
dan persamaan untuk menghitung nilai simpangan baku (standar deviasi) dapat dilihat sebagai berikut :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \qquad (7)$$

di mana :
σ : standar deviasi
x_i : nilai x ke -*i*
μ : rata-rata hitung
n : jumlah sampel

   b. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.
3. Mendapatkan nilai dalam tabel *mean*, standart deviasi dan probabilitas.
4. solusi kemudian dihasilkan.

2. Pembahasan

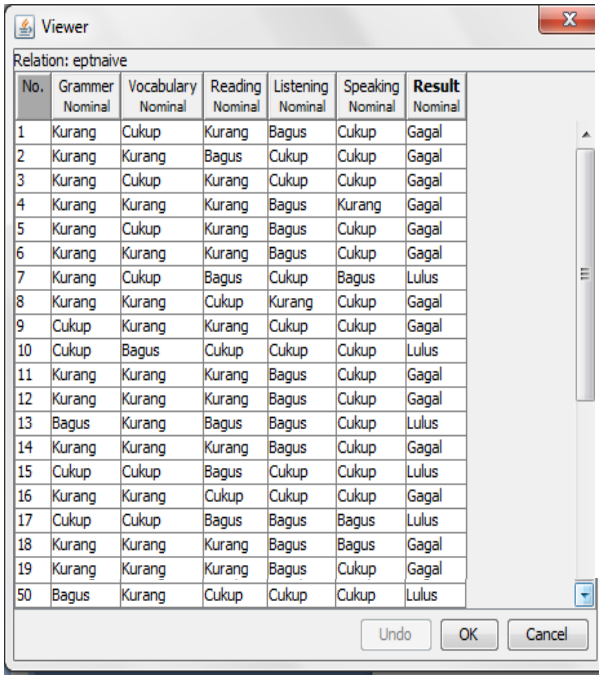
Penerapan Metode *Naive Bayes*

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Dalam metode Naive Bayes data String yang bersifat konstan dibedakan dengan data numerik yang bersifat kontinyu, perbedaan ini akan terlihat pada saat menentukan nilai probabilitas setiap kriteria baik itu kriteria dengan nilai data string maupun kriteria dengan nilai data numerik. Adapun penerapan metode Naive Bayes sebagai berikut.

a. Baca *Data Training*

Untuk menentukan data yang nantinya akan dianalisis dengan metode *Naive Bayes* maka langkah pertama yang dilakukan adalah membaca data latih. Adapun data latih yang digunakan dapat dilihat pada tabel 1 berikut :

Tabel 1. Data Training



No.	Grammer Nominal	Vocabulary Nominal	Reading Nominal	Listening Nominal	Speaking Nominal	Result Nominal
1	Kurang	Cukup	Kurang	Bagus	Cukup	Gagal
2	Kurang	Kurang	Bagus	Cukup	Cukup	Gagal
3	Kurang	Cukup	Kurang	Cukup	Cukup	Gagal
4	Kurang	Kurang	Kurang	Bagus	Kurang	Gagal
5	Kurang	Cukup	Kurang	Bagus	Cukup	Gagal
6	Kurang	Kurang	Kurang	Bagus	Cukup	Gagal
7	Kurang	Cukup	Bagus	Cukup	Bagus	Lulus
8	Kurang	Kurang	Cukup	Kurang	Cukup	Gagal
9	Cukup	Kurang	Kurang	Cukup	Cukup	Gagal
10	Cukup	Bagus	Cukup	Cukup	Cukup	Lulus
11	Kurang	Kurang	Kurang	Bagus	Cukup	Gagal
12	Kurang	Kurang	Kurang	Bagus	Cukup	Gagal
13	Bagus	Kurang	Bagus	Bagus	Cukup	Lulus
14	Kurang	Kurang	Kurang	Bagus	Cukup	Gagal
15	Cukup	Cukup	Bagus	Cukup	Cukup	Lulus
16	Kurang	Kurang	Cukup	Cukup	Cukup	Gagal
17	Cukup	Cukup	Bagus	Bagus	Bagus	Lulus
18	Kurang	Kurang	Kurang	Bagus	Bagus	Gagal
19	Kurang	Kurang	Kurang	Bagus	Cukup	Gagal
50	Bagus	Kurang	Cukup	Cukup	Cukup	Lulus

b. Kriteria dan Probabilitas

Adapun nilai probabilitas setiap kriteria didapatkan dari data latih pada tabel 1. nilai probabilitas setiap kriteria sebagai berikut :

1. Probabilitas Kriteria *Grammer*

   Berdasarkan data hasil uji english proficiency test pada table 1 diketahui jumlah data latih (*data training*) adalah sebanyak 50 data mahasiswa, di mana dari 50 mahasiswa tersebut terdapat 36 mahasiswa gagal dengan nilai *grammer* kurang, 2 mahasiswa gagal dengan nilai *grammer* cukup dan tidak ada mahasiswa gagal dengan nilai *grammer* bagus, sementara itu terdapat 2 mahasiswa yang lulus dengan nilai *grammer* kurang, 5 mahasiswa yang lulus dengan nilai *grammer* cukup dan 5 mahasiswa lulus dengan nilai *grammer* bagus. Probabilitas kriteria *Grammer* dapat dilihat pada tabel 2.

Tabel 2. Probabilitas Kriteria *Grammer*

Grammer	Jumlah Kejadian		Probabilitas	
	Gagal	Lulus	Gagal	Lulus
Kurang	36	2	0,95	0,17
Cukup	2	5	0,05	0,42
Bagus	0	5	0,00	0,42
Jumlah	38	12	0,76	0,24

2. Probabilitas Kriteria *Vocabulary*.

   Pada kriteria *vocabulary* dapat diketahui dari 50 mahasiswa tersebut terdapat 25 mahasiswa gagal dengan nilai *vocabulary* kurang, 12 mahasiswa gagal dengan nilai *vocabulary* cukup dan 1 mahasiswa gagal dengan nilai *vocabulary* bagus, sementara itu terdapat 3 mahasiswa yang lulus dengan nilai *vocabulary* kurang, 5 mahasiswa yang lulus dengan nilai *vocabulary* cukup dan 4 mahasiswa lulus dengan nilai *vocabulary* bagus. Probabilitas kriteria *vocabulary* dapat dilihat pada tabel 3.

Tabel 3. Probabilitas Kriteria *Vocabulary*

Vocabulary	Jumlah Kejadian		Probabilitas	
	Gagal	Lulus	Gagal	Lulus
Kurang	25	3	0,66	0,25
Cukup	12	5	0,32	0,42
Bagus	1	4	0,03	0,33
Jumlah	38	12	0,76	0,24

3. Probabilitas Kriteria *Reading*.

   Pada kriteria *reading* dapat diketahui dari 50 mahasiswa tersebut terdapat 26 mahasiswa gagal dengan nilai *reading* kurang, 9 mahasiswa gagal dengan nilai *reading* cukup dan 3 mahasiswa gagal dengan nilai *reading* bagus, sementara itu terdapat 1 mahasiswa yang lulus dengan nilai *reading* kurang, 6 mahasiswa yang lulus dengan nilai *reading* cukup dan 5 mahasiswa lulus dengan nilai *reading* bagus. Probabilitas kriteria *reading* dapat dilihat pada tabel 4.

Tabel 4. Probabilitas *Reading*

Reading	Jumlah Kejadian		Probabilitas	
	Gagal	Lulus	Gagal	Lulus
Kurang	26	1	0,68	0,08
Cukup	9	6	0,24	0,50
Bagus	3	5	0,08	0,42
Jumlah	38	12	0,76	0,24

4. Probabilitas *Listening.*

   Pada kriteria *listening* dapat diketahui dari 50 mahasiswa tersebut terdapat 6 mahasiswa gagal dengan nilai *listening* kurang, 17 mahasiswa gagal dengan nilai *listening* cukup dan 15 mahasiswa gagal dengan nilai *listening* bagus, sementara itu tidak ada mahasiswa yang lulus dengan nilai *listening* kurang, 5 mahasiswa yang lulus dengan nilai *listening* cukup dan 7 mahasiswa lulus dengan nilai *listening* bagus. Probabilitas kriteria *listening* dapat dilihat pada tabel 5.

Tabel 5. Probabilitas *Listening*

Listening	Jumlah Kejadian		Probabilitas	
	Gagal	Lulus	Gagal	Lulus
Kurang	6	0	0,16	0,00
Cukup	17	5	0,45	0,42
Bagus	15	7	0,39	0,58
Jumlah	38	12	0,76	0,24

5. Probabilitas *Speaking*.
   Pada kriteria *speaking* dapat diketahui dari 50 mahasiswa tersebut terdapat 1 mahasiswa gagal dengan nilai *speaking* kurang, 32 mahasiswa gagal dengan nilai *speaking* cukup dan 5 mahasiswa gagal dengan nilai *speaking* bagus, sementara itu tidak ada mahasiswa lulus dengan nilai *speaking* kurang, 10 mahasiswa lulus dengan nilai *speaking* cukup dan 2 mahasiswa lulus dengan nilai *speaking* bagus. Probabilitas kriteria *speaking* dapat dilihat pada tabel 6.

Tabel 6. Probabilitas *Speaking*

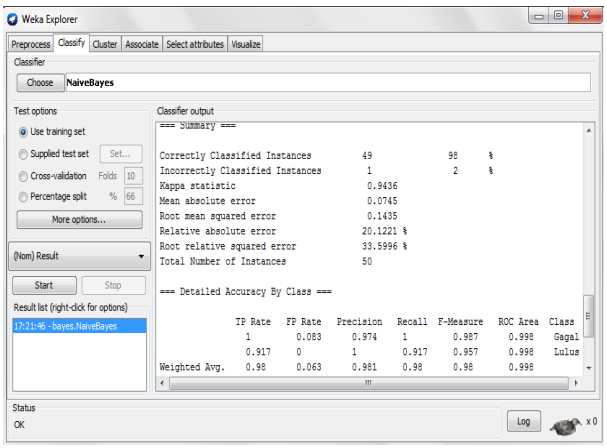
Listening	Jumlah Kejadian		Probabilitas	
	Gagal	Lulus	Gagal	Lulus
Kurang	1	0	0,03	0,00
Cukup	32	10	0,84	0,83
Bagus	5	2	0,13	0,17
Jumlah	38	12	0,76	0,24

6. Probabilitas Kriteria *Result*.
   Berdasarkan tabel 1 diketahui dari 50 mahasiswa yang mengikuti English Proficiency Test terdapat 38 mahasiswa yang gagal, 12 mahasiswa yang lulus. Probabilitas kriteria *result* dapat dilihat pada tabel 7.

Tabel 7. Probabilitas Kriteria *Result*

Result	Jumlah Kejadian		Probabilitas	
	Gagal	Lulus	Gagal	Lulus
Jumlah	38	12	0,76	0,24

Adapun hasil klasifikasi kelulusan mahasiswa yang mengikuti *English Proficiency Test* dengan weka terlihat pada gambar 2 dengan jumlah data yang diuji sebanyak 50 data mahasiswa sebagai berikut :



Gambar 2. Hasil Klasifikasi Metode Naive Bayes

dilihat persentase untuk *Correctly Classified Instance* adalah sebesar 98% sementara persentase untuk *Incorrectly Classified Instance* adalah sebesar 2%. Dimana dari 50 data, sebanyak 49 data berhasil diklasifikasikan dengan benar dan sebanyak 1 tidak berhasil diklasifikasikan dengan benar.

3. Kesimpulan

Berdasarkan penelitian mengenai prediksi kelulusan mahasiswa yang mengikuti *English Proficiency Test* ada beberapa kesimpulan sebagai berikut :

1. Berdasarkan data mahasiswa yang diperoleh, proses *Data Mining* membantu dalam penerapan metode *Naive Bayes* dalam mendapatkan informasi dari hasil klasifikasi kelulusan mahasiswa pada uji kompetensi *English Proficiency Test*.

2. Metode *Naive Bayes* memanfaatkan data *training* untuk menghasilkan probabilitas setiap kriteria untuk *class* yang berbeda, sehingga nilai-nilai probabilitas dari kriteria tersebut dapat dioptimalkan untuk memprediksi kelulusan mahasiswa berdasarkan proses klasifikasi yang dilakukan oleh metode *Naive Bayes* itu sendiri.

3. Berdasarkan data mahasiswa yang mengikuti *English Proficiency Test* yang dijadikan data *training*, metode *Naive Bayes* berhasil mengklasifikasikan 49 data dari 50 data yang diuji. Sehingga dengan demikian metode N*aive Bayes* ini berhasil memprediksi kelulusan mahasiswa dengan persentase keakuratan sebesar 98 %.

Daftar Pustaka

[1] Manjusha K.K, et al, (2014). *Prediction of Different Dermatological Conditions Using Naive Bayesian Classification*, International Journal of Advanced Research in Computer Science and Software Engineering, 2014.

[2] S.L. Ting , et al, (2011). *Is Naive Bayes a Good Classifier for Document Classification ?*, International journal of Software Engineering and Its Applications, Vol. 5, 3, July, 2011.

[3] Shyara taruna R, Saroj Hiranwal, (2013). *Enhanced Naive Bayes Algorithm for Intrusion Detection in Data Mining*, International Journal of Computer Science and information Technologies, Vol. 4, 2013.

[4] Angga Ginanjar Mabrur, Riani Lubis, (2012). *Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit*, Jurnal Komputer dan Informatika (KOMPUTA) Edisi 1, Vol. 1, Maret 2012.

[5] Surbekti Mujiasih, (2011). *Pemanfaatan Data Mining Untuk Prakiraan Cuaca*, Jurnal Meteorologi dan Geofisika, Volume 12, Nomor 2, September 2011.

[6] Mujib Ridwan, dkk, (2013), *Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier*, Jurnal EECCIS Vol. 7, No. 1, Juni 2013.

[7] Tina R. Patil, S.S. Sherekar, (2013). *Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*, International Journal of Computer Science and Applications, Vol. 6, No. 2, April 2013.

[8] Bustami, (2013). *Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi*, TECHSI : Jurnal Penelitian Teknik Informatika.

[9] Shadab Adam Pattekari, Asma Parveen, *(2012), Prediction System for Heart Disease Using Naive Bayes*, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624, Vol. 3, Issue 3, 2012