Akurasi Data Mining Untuk Menghasilkan Pola Kelulusan Mahasiswa dengan Metode NAÏVE BAYES

M. Ridwan Effendi

Fakultas Komputer Jurusan Sistem Informasi
Universitas Mohammad Husni Thamrin Jakarta
Email : jundi79@gmail.com

Abstrak

Data mining merupakan cara untuk menemukan informasi dengan mencari pola atau aturan tertentu dari data dalam jumlah besar yang diharapkan dapat mengatasi kondisi tersebut. Dengan memanfaatkan data mahasiswa dan data kelulusan mahasiswa sebagai sumber datanya, diharapkan dapat menghasilkan informasi tentang pola tingkat kelulusan mahasiswa melalui teknik data mining. Kategori tingkat kelulusan di ukur dari nilai IPK (Indek Prestasi Komulatif). Metode Algoritma yang digunakan adalah metode algoritma naïve bayes. Proses pada aplikasi ini ada 2 macam yaitu, proses analisa pola data kelulusan siswa yang telah ada sebelumnya (Learning Phase) berdasarkan atribut – atribut yang di ujikan dan proses dari analisa pola data baru yang diujikan berdasarkan pola yang telah ada (Testing Phase). Informasi yang ditampilkan pada aplikasi tersebut ada 2 macam yaitu, informasi hasil proses Learning Phase dan informasi data berupa nilai probabilitas posterior (kemungkinan kemunculan) dari masing-masing kategori tingkat kelulusan. Pada analisa data yang dilakukan diproses testing, di dapat akurasi datamining pada tingkat keakuratan sekitar 99,33% dan memiliki nilai error 0.27% berdasarkan pengujian 1043 data mahasiswa tahun 2013-2014.

Kata kunci : Data mining, algoritma naïve baye, tingkat kelulusan, data induk siswa, testing phase, learning phase

1.1. Pendahuluan

Dengan kemajuan teknologi informasi, kebutuhan akan informasi yang akurat sangat dibutuhkan menjadi suatu elemen penting dalam perkembangan masyarakat saat ini dan waktu mendatang. Namun kebutuhan informasi yang tinggi kadang tidak di imbangi dengan penyajian informasi yang memadai, seringkali informasi tersebut masih harus digali ulang dari data yang jumlahnya sangat besar. Penggunaan teknik data mining diharapkan dapat memberikan pengetahuan yang sebelumnya tersembunyi di dalam gudang data sehingga menjadi informasi yang berharga. Sekolah Tinggi Manajemen Iformatika & Komputer (STMIK) saat ini dituntut untuk memiliki keunggulan bersaing dan memiliki kualitas yang baik. Untuk mengatasi hal tersebut, pihak perguruan tinggi di tuntut untuk dapat mengambil langkah – langkah yang tepat.

1.2. Latar Belakang

Dalam meningkatkan kualitas nilai kelulusan mahasiswa. Sistem ini akan membantu pihak sekolah mengetahui pola kelulusan dari mahasiswa-mahasiswinya dengan memanfaatkan data mahasiswa dan data kelulusan mahasiswa. Dari pola tersebut, diharapkan bisa menganalisa faktor-faktor yang sangat berpengaruh pada tingkat kelulusan. Hal ini, membantuk pihak perguruan tinggi dalam menyaring mahasiswa-mahasiswi yang lebih kompeten selain berdasarkan rangking dari nilai. Sehingga, membantu pihak kampus untuk menyusun strategi yang tepat dalam meningkatkan kualitas perguruan tinggi dan menjadikan perguruan tinggi memiliki daya saing yang tinggi.

1.3 Perumusan Masalah

1. Bagaimana menerapkan teknik Data Mining dengan Metode Naïve Bayes untuk menampilkan informasi Tingkat Kelulusan dengan Data Induk Maha-

siswa dan Data Kelulusan Mahasiswa sebagai sumber datanya.
2. Bagaimana membuat sistem unttuk menganalisa data, sehingga bisa menjadi informasi yang berguna untuk meningkatkan kualitas mahasiswa
3. Bagaimana menampilkan informasi agar dapat digunakan dalam membantu pengambilan keputusan untuk meningkatkan kualitas perguruan tinggi.

1.4 Tujuan Penelitian

1. Menerapkan teknik Data Mining dengan Metode Naïve Bayes dan menyajikan informasi kelulusan mahasiswa
2. Mempermudah analisa data kelulusan yang jumlahnya besar agar dapat diketahui faktor-faktor yang sangat berpengaruh pada tingkat kelulusan.
3. Membuat sistem pendukung keputusan untuk membantu meningkatkan kualitas kelulusan mahasiswa

1.5 Batasan Masalah

1. Semua proses perhitungan yang disediakan oleh sistem menggunakan teknik data minning dengan metode Naive Bayes.
2. Informasi yang ditampilkan berupa laporan anlisa pola data mining tingkat kelulusan dan nilai kalkulasi probabilitas posterior pada hubungan antara tingkat kelulusan dengan data induk mahasiswa. Data, formatnya pun disesuaikan dengan kebutuhan data mining.
3. Data Induk Siswa dan data Kelulusan yang diambil sebagai sampel dalam aplikasi ini adalah data tahun 2013 dan 2014.
4. Sistem ini hanya sebagai pendukung keputusan, bukan sebagai faktor utama dalam mengambil keputusan (faktor utama, bisa berdasarkan rangking nilai siswa yang mendaftar).
5. Perancangan dan pembuatan sistem ini dengan menggunakan program aplikasi Rapid Miner 7.0 dan perancangan database dengan menggunakan datasheet campus.

2. Tinjauan Pustaka

2.1 Data Mining

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu-ilmu lain, seperti database system, data warehousing, statistik, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing (Han, 2006). Data mining didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar (Witten, 2005).

2.2 Tahap – Tahap Data Mining

Tahap-tahap data mining ada 6 yaitu :

1. Pembersihan data (data cleaning) Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Integrasi data (data integration)
2. Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.
3. 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.
4. Transformasi data (Data Transformation) Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining.
5. Proses mining, Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.
6. Evaluasi pola (pattern evaluation), Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan.
7. Presentasi pengetahuan (knowledge presentation), Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.3 Metode Naïve Bayes Classifier

Simple Naive Bayesian classifier merupakan salah satu metode pengklasifikasi berpeluang sederhana yang berdasarkan pada penerapan Teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen). Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Dua kelompok peneliti, satu oleh Pantel dan Lin, dan yang lain oleh Microsoft Research memperkenalkan metode statistik Bayesian ini pada teknologi anti spam filter. Tetapi yang membuat algoritma Bayesian filtering ini popular adalah pendekatan yang dilakukan oleh Paul Graham. Dasar dari teorema naive digunakan dalam pemrograman adalah rumus Bayes berikut ini:

P (A|B) = (P(B|A) * P(A))/P(B)

Artinya Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Contoh penggunaan Algoritma Naive Bayes antara lain:

- Untuk klasifikasi dokumen.
- Untuk deteksi SPAM atau fitering SPAM.
- Dan masalah klasifikasi lainnya Teorema Bayes: P(C|X) = P(X|C)·P(C) / P(X).
  Dimana :
  P(X) bernilai konstan utk semua klas
  P(C) merupakan frek relatif sample klas C
  Dicari P(C|X) bernilai maksimum, sama halnya dengan P(X|C)·P(C) juga bernilai maksimum

3. Analisa Dan Perancangansistem

3.1 Langkah Penyelesaian

Berikut ini adalah langkah-langkah mining data gabungan dari data induk mahasiswa dan data kelulusan mahasiswa dengan metode classification naïve bayes agar menghasilkan suatu pola tingkat kelulusan yang diperoleh dari data induk mahasiswa dan data kelulusan. Proses classification dibagi menjadi dua phase yaitu learning dan test.

Data pada tabel gabungan di atas ada 2 tipe :

1. Data Statis

Data statis adalah data yang sifatnya tetap, tidak mengalami perubahan nilai. Berikut ini merupakan rumus yang digunakan untuk mencari data yang sifatnya statis :

$$\frac{\text{number of training samples belonging to } C_j}{\text{total number of training samples}}$$

Salah satu contoh atribut yang bersifat statis adalah atribut jenis kelamin, hanya terdapat 2 nilai yaitu laki-laki, perempuan.

2. Data Kontinue

Data kontinue adalah data yang nilainya berubah ubah. Biasanya data setnya berupa data numerik. Berikut ini adalah rumus yang di gunakan untuk mencari nilai probabilitas kemunculan pada data yang sifatnya kontinue.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

$$f(w) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(w-\mu)^2}{\sigma^2}}$$

*Ket : f(w) merupakan nilai kemunculan probabilitas untuk data yang mempunyai nilai (w). Salah satu contoh atribut yang bersifat kontinue adalah atribut danem, setiap siswa memiliki nilai danem yang berbeda – beda.

3.2 Proses Transformasi Data

Tabel ini di jadikan acuan atribut dataset akademik

Tabel 1. Atribut data training dataset akademik

No	Atribut	Keterangan	
1	Nama Mahasiswa	Nama mahasiswa sesuai tahun ajaran	
2	Umur	Umur mahasiswa sesuai tahun ajaran	
3	Jenis Kelamin	Kode	Keterangan
		1	Laki-laki
		2	Perempuan
4	Status Mahasiswa	Aktivitas	
		Kode	Keterangan
		1	Bekerja
		2	Mahasiswa
5	IPS & IPK	Index Prestasi Sementara	
		Index Prestasi Komulatif	
6	Status Lulus	2 = Tepat Waktu	
		1 = Telat Lulus	

tabel predikat kelulusan berdasarkan nilai IPK dapat dikategorikan menjadi tiga yaitu :

1. IPK kelulusan A dengan nilai 3.51– 4.00
2. IPK kelulusan B dengan nilai 3.01 – 3.50
3. IPK kelulusan C dengan nilai 2.50 – 3.00

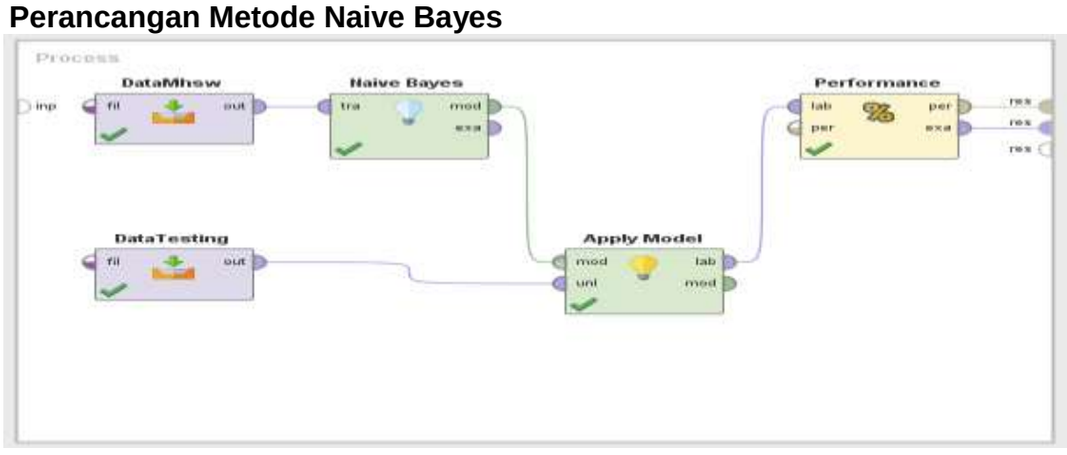
3.3 Perhitungan Learning Phase

Atribut yang digunakan untuk proses testing phase adalah niilai ujian. Berikut ini adalah salah satu contoh perhitungan:

Tabel 2. Perhitungan Learnig Phase

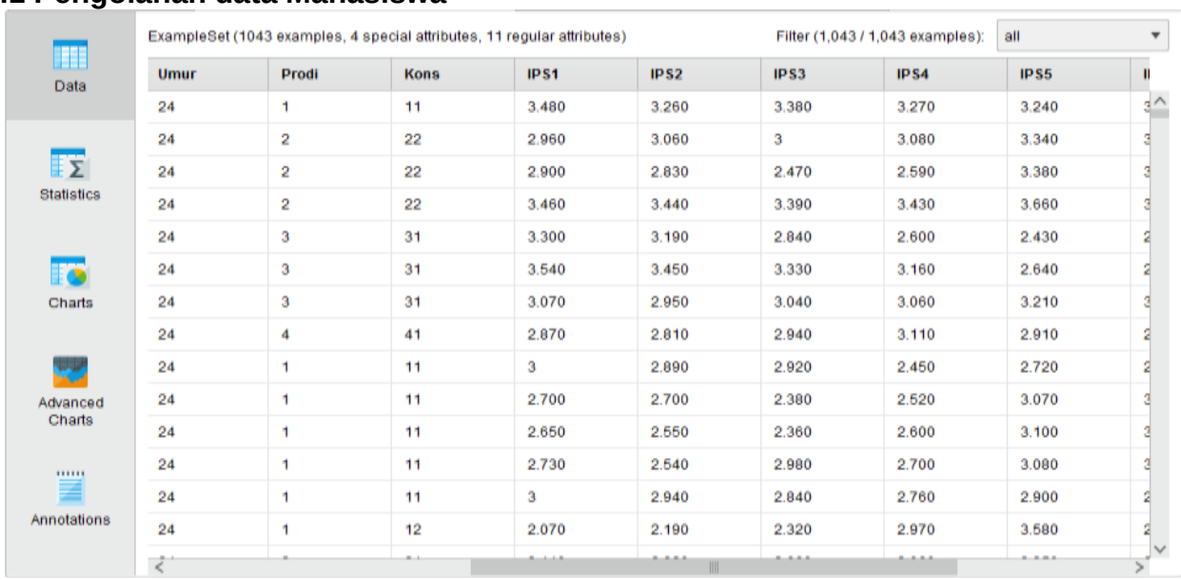
Semerter	IPK Kelulusan A	IPK Kelulusan B	IPK Kelulusan C
1	3.57	3.25	2.75
2	3.68	3.35.	2.90
3	3.76	3.42	2.58
4	3.86	3.30	2.37
5	3.56	3.10	2.82
6	3.55	3.07	3.00
7	3.59	3.00	2.76
8	3.70	3.37	2.98
Mean (μ)			

3.4 Perancangan Metode Naive Bayes



4. Implementasi

4.1 Pengolahan data Mahasiswa

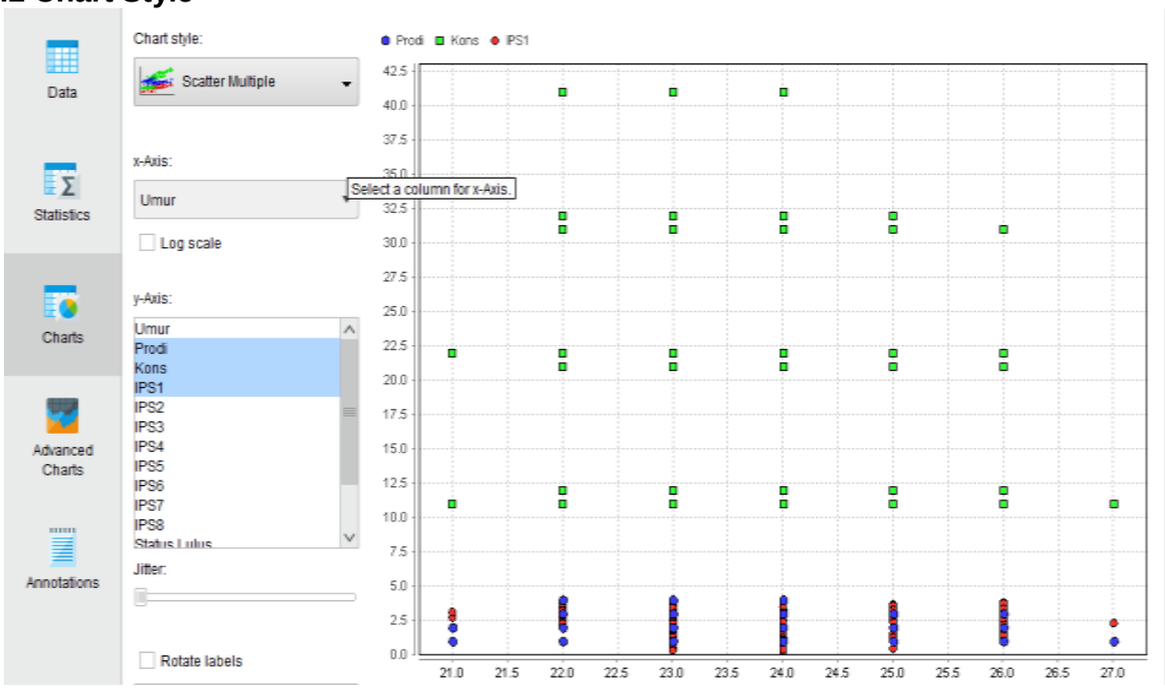


Data
Statistics
Charts
Advanced Charts
Annotations

ExampleSet (1043 examples, 4 special attributes, 11 regular attributes) Filter (1,043 / 1,043 examples): all

Umur	Prodi	Kons	IPS1	IPS2	IPS3	IPS4	IPS5
24	1	11	3.480	3.260	3.380	3.270	3.240
24	2	22	2.960	3.060	3	3.080	3.340
24	2	22	2.900	2.830	2.470	2.590	3.380
24	2	22	3.460	3.440	3.390	3.430	3.660
24	3	31	3.300	3.190	2.840	2.600	2.430
24	3	31	3.540	3.450	3.330	3.160	2.640
24	3	31	3.070	2.950	3.040	3.060	3.210
24	4	41	2.870	2.810	2.940	3.110	2.910
24	1	11	3	2.890	2.920	2.450	2.720
24	1	11	2.700	2.700	2.380	2.520	3.070
24	1	11	2.650	2.550	2.360	2.600	3.100
24	1	11	2.730	2.540	2.980	2.700	3.080
24	1	11	3	2.940	2.840	2.760	2.900
24	1	12	2.070	2.190	2.320	2.970	3.580

4.2 Chart Style



Data
Statistics
Charts
Advanced Charts
Annotations

Chart style: Scatter Multiple

x-Axis: Umur

Log scale

y-Axis: Umur, Prodi, Kons, IPS1, IPS2, IPS3, IPS4, IPS5, IPS6, IPS7, IPS8, Status Lulus

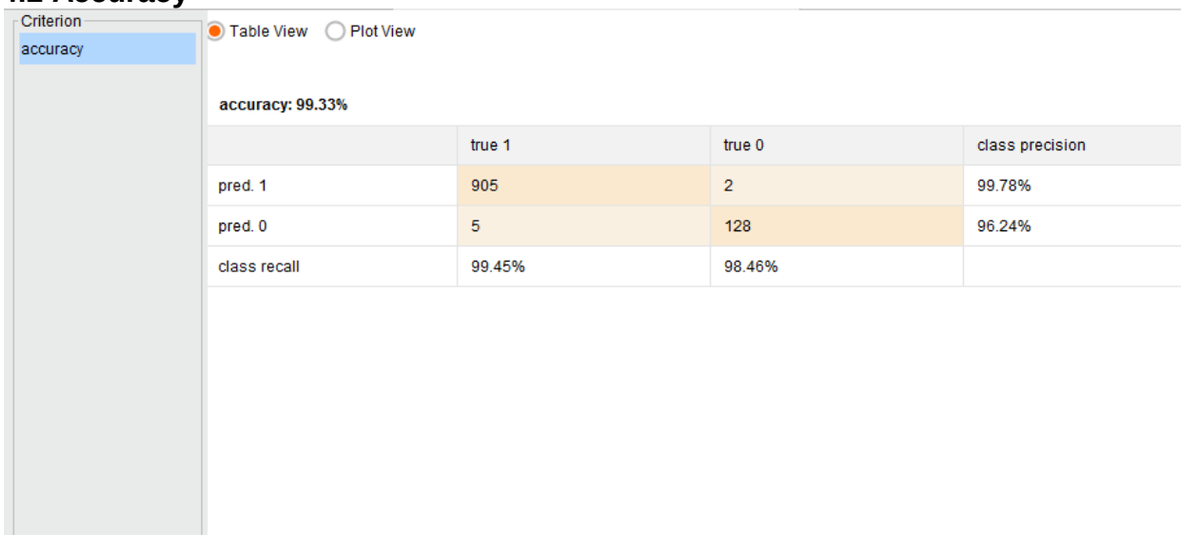
Jitter:

Rotate labels

Prodi Kons IPS1

Select a column for x-Axis.

4.2 Accuracy



Criterion: accuracy

Table View Plot View

accuracy: 99.33%

	true 1	true 0	class precision
pred. 1	905	2	99.78%
pred. 0	5	128	96.24%
class recall	99.45%	98.46%	

5. Kesimpulan Dan Saran

5.1 Kesimpulan

Dari perancangan dan implementasi Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa ini, dapat ditarik kesimpulan bahwa : .

1. Proses testing digunakan untuk memprediksi akurasi data mahasiswa tentang tingkat kelulusan yang akan diperoleh. Pada proses ini atribut yang digunakan adalah IPK, Program studi. Atribut tersebut dipilih karena memiliki nilai perkalian support dan confidence yang tinggi dibandingkan atribut yang lain.
2. Pada analisa data yang dilakukan diproses testing, di dapat tingkat keakuratan sistem sekitar 99,33% dan memiliki nilai error 0.27%, berdasarkan pengujian 1043 data siswa tahun 2013 - 2014.

5.2 Saran

Dari hasil evaluasi aplikasi yang telah dibuat, penulis menyadari bahwa aplikasi yang dibuat masih terdapat kekurangan. Report yang di hasilkan berupa nilai kelulusan menggunakan metode naive bayes, karena perhitungan yang dilakukan masih mengacu pada perhitungan data mahasiswa dengan aplikasi Rapid miner. Untuk pengembangan lebih lanjut, bisa di inputkan data-data yang lebih bervariasi dalam proses analisa. Sehingga hasil analisa yang di dapatkan lebih mendekati tingkat keberhasilan. Sehingga bisa membantu pihak manajemen perguruan tinggi dalam menindak lanjuti perbaikan kualitasnya.

Daftar Pustaka

Davies, and Paul Beynon, 2004, “Database Systems Third Edition”, Palgrave Macmillan, New York.

Elmasri, Ramez and Shamkant B. Navathe, 2000, “Fundamentals of Database Systems. Third Edition”, Addison Wesley Publishing Company, New York.

Kadir, Abdul, 1999, “Konsep dan Tuntunan Praktis Basis Data”, Penerbit Andi, Yogyakarta.

Kusrini, dan Emha Taufik Luthfi, 2009, “Algoritma Data Mining”, Penerbit Andi, Yogyakarta.

Pramudiono, I. 2007. Pengantar Data Mining : Menambang Permata Pengetahuan di Gunung Data.http://www.ilmukomputer.org/wpcontent/uploads/2006/08/iko-datamining.zip Diakses pada tanggal 15 Maret 2009 jam 08.54.

Nurul Pratiwi, Oktariani. 2009. Klasifikasi Posting Blog Berbahasa Indonesia dengan Menggunakan Algoritma Naïve Bayes. Bandung : Universitas Pendidikan Indonesia.

Wibisono, Yudi. 2005. Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier1. Bandung: Universitas Pendidikan Indonesia