APLIKASI DATA MINING UNTUK MENGHASILKAN POLA KELULUSAN SISWA DENGAN METODE NAÏVE BAYES

Budanis Dwi Meilani¹⁾, Nofi Susanti²⁾

Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Insitut Teknologi Adhi Tama Surabaya Jl. Arief Rahman Hakim 100 – Surabaya 60117

Email: dwimeilanibudanis@yahoo.com¹, opick_susan2@ymail.com², adifajaryanto@gmail.com³

Abstrak

Pertumbuhan yang pesat dari akumulasi data telah menciptakan kondisi kaya akan data tetapi minim informasi. 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 induk siswa dan data kelulusan siswa sebagai sumber datanya, diharapkan dapat menghasilkan informasi tentang pola tingkat kelulusan siswa melalui teknik data mining. Kategori tingkat kelulusan di ukur dari nilai UNAS. Algoritma yang digunakan adalah 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 tingkat keakuratan sistem sekitar 99,82% dan memiliki nilai error 0.18% berdasarkan pengujian 220 data siswa tahun 2012.

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

Abstract

The rapid growth of data accumulation has created conditions data-rich but information poor. Data mining is a way to find information by looking for patterns or certain rules of large amounts of data which is expected to treat the condition. By leveraging master data and data of students' graduation students as a source of data, is expected to yield information about the pattern of student graduation rates through data mining techniques. Categories graduation rate is measured from the UNAS. The algorithm used is naïve Bayes algorithm. Process in this application there are two kinds, namely, the process of analyzing the data pattern graduation of students who have been there before (Learning Phase) based on attributes - attributes in ujikan and the process of analyzing the new data patterns that are tested based on the pattern that has existed (Testing Phase). Information displayed in the application there are two kinds, namely, information about the results of the Learning Phase and data information in the form of a posterior probability value (probability of occurrences) of each category graduation rates. In the data analysis process testing, at the level of accuracy of the system can be approximately 99.82% and has an error value 0:18 220% based testing of student data in 2012.

Keyword: data mining, naïve Bayes algorithm, graduation rates, student master data, testing phase, learning phase..

1. Pendahuluan

Dengan kemajuan teknologi informasi dewasa ini, kebutuhan akan informasi yang akurat sangat dibutuhkan dalam kehidupan sehari-hari, sehingga informasi akan 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-pengetahuan yang sebelumnya tersembunyi di dalam gudang data sehingga menjadi informasi yang berharga. Sekolah Menengah Atas (SMA) saat ini dituntut untuk memiliki keunggulan bersaing dan memiliki kualitas yang baik. Untuk mengatasi hal tersebut, pihak sekolah di tuntut untuk dapat mengambil langkah yang tepat dalam memajukan kualitas sekolah. Hal ini bisa dilakukan dengan cara meningkatkan kualitas nilai kelulusan siswa.

Sistem ini akan membantu pihak sekolah mengetahui pola kelulusan dari siswa-siswinya dengan memanfaatkan data siswa dan data kelulusanDari pola tersebut, diharapkan bisa menganalisa faktorfaktor yang sangat berpengaruh pada tingkat kelulusan. Hal ini, membantuk pihak sekolah dalam menyaring siswa-siswi yang lebih kompeten selain berdasarkan rangking dari nilai. Sehingga, membantu pihak sekolah untuk menyususn strategi yang tepat dalam meningkatkan kualitas sekolah dan menjadikan sekolah memiliki daya saing yang tinggi.

1.1. Rumusan Masalah

Berdasarkan latar belakang diatas, maka rumusan masalah pada penelitian ini adalah:

- 1. Bagaimana menerapkan teknik *data mining* dengan Metode *Naïve Bayes* untuk menampilkan informasi tingkat kelulusan dengan data induk siswa dan data kelulusan sebagai sumber datanya.
- 2. Bagaimana membuat sistem unttuk menganalisa data, sehingga bisa menjadi informasi yang berguna untuk meningkatkan kualitas siswa.
- 3. Bagaimana menampilkan informasi agar dapat digunakan dalam membantu pengambilan keputusan untuk meningkatkan kualitas sekolah.

1.2. Batasan Masalah

Batasan masalah dalam objek penelitian ini adalah:

- 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 siswa. Data, formatnya pun disesuaikan dengan kebutuhan data mining.
- 3. Data induk siswa dan Kelulusan yang diambil sebagai sampel adalah data tahun 2010 dan 2011.
- 4. Sistem ini hanya sebagai pendukung keputusan, bukan sebagai faktor utama dalam mengambil keputusan (faktot utama, bisa berdasarkan rangking nilai siswa yang mendaftar).
- 5. Perancangan dan pembuatan sistem ini menggunakan program aplikasi Visual Basic 6.0 dan perancangan database dengan menggunakan Microsoft Access 2003 pada sistem operasi Windows.

1.3. Tujuan Penelitian dan Manfaat Penelitian

Tujuan penelitian ini adalah:

- 1. Menerapkan teknik Data Mining dengan Metode Naïve Bayes dan menyajikan informasi kelulusan
- 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 siswat

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. Tahapan Data Mining

Tahap-tahap data mining ada 6 yaitu:

- 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.
- Transformasi data (Data Transformation)
 Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining.
- 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 berikut ini:

$$P(A|B) = (P(B|A) * P(A))/P(B)$$
 (1)

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:

$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$
(2)

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 Perancangan Sistem

3.1. Langkah Penyelesaian dengan learning Phase

Berikut ini adalah langkah – langkah mining data gabungan dari data induk siswa dan data kelulusan siswa dengan metode classification naïve bayes agar menghasilkan suatu pola tingkat kelulusan yang diperoleh dari data induk siswa 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 :

$$\frac{\text{number of training samples belonging to } C_j}{\text{total number of training samples}}$$
(3)

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.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4}$$

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

$$f(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(w-\mu)^2}{\sigma^2}}$$
(5)

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 ditunjukkan oleh tabel 1.

Proses Transformasi Data

Pada tabel 1 dijadikan acuan untuk menentukan pola tingkat kelulusan siswa karena didalamnya terdapat range nilai pada masing-masing kategori. Sehingga tingkat kelulusan siswa dapat diketahui dengan mudah melalui tiga kategori yang ada.

Tabel 1. Tabel Range Danem

KATEGORI	KETERANGAN
A	danem $60.00 - 48.00$
В	danem 47.99 – 36.00
C	danem 35.99 – 00.00

3.2. Contoh Perhitungan Learning

3.2.1. Contoh Perhitungan Learning Phase (Data Statis)

Proses *learning phase* pada data gabungan (data induk siswa dan data kelulusan) antara lain adalah pada atribut jenis kelamin, asal sekolah, tempat lahir, program studi, tahun masuk, tahun lulus, dan tingkat kelulusan (tabel 2).

Tabel 2. Probabilitas Kelulusan Berdasarkan Jenis Kelamin

JENIS KELAMIN	KELULUSAN = A	KELULUSAN = B	KELULUSAN = C
P	2/5	2/5	2/4
L	3/5	3/5	2/4

Di bawah ini adalah salah satu contoh perhitungan mining learning phase pada jenis kelamin

- 1. Probabilitas tingkat kelulusan A
 - a. 2/5 merupakan siswa yang berjeni kelamin perempuan
 - b. 3/5 merupakan siswa yang berjenis kelamin laki laki
- 2. Probabilitas tingkat kelulusan B
 - a. 2/5 merupakan siswa yang berjenis kelamin perempuan
 - b. 3/5 merupakan siswa yang berjenis kelamin laki laki
- 3. Probabilitas tingkat kelulusan C
 - a. 2/4 merupakan siswa yang berjenis kelamin perempuan
 - b. 2/4 merupakan siswa yang berjenis kelamin laki laki

3.3.2 Contoh Perhitungan Learning Phase (Data Kontinue)

Atribut yang digunakan untuk proses testing phase antara lain adalah danem smp, nilai ujian (bahasa Indonesia, bahasa inggris, matematika, fisika, kimia, biologi, sosiologi, geografi, ekonomi). Berikut ini adalah salah satu contoh perhitungan danem SMP (tabel 3).

Tabel 3. Perhitungan Danem

DANEM SMP	KELULUSAN = A	KELULUSAN = B	KELULUSAN = C
1	30.65	36.8	34.45
2	31.75	28.25	26.75
3	27.15	28.4	33.85
4	33	31.2	30.95
5	34.1	30.55	-
Mean(μ) Std dev(σ)	31.33 7.14325	31.03 12.062	31.5 12.363

3.3. Contoh Proses Perhitungan Testing

Berikut ini adalah rumus yang digunakan untuk menghitung MAP rule pada testing phase :

$$[P(x_1 \mid c^*) \cdots P(x_n \mid c^*)]P(c^*) > [P(x_1 \mid c) \cdots P(x_n \mid c)]P(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$
(7)

Contoh:

Jika ada seorang siswa yang di sekolah barunawati dengan data-data sebagai berikut

- 1. Jenis kelamin = perempuan
- 2. Kota asal sekolah = gresik
- 3. Asal sekolah = SMP Trisila
- 4. Danem SMP = 29.50
- 5. Program studi = ipa

Bagaimana pola tingkat kelulusan pada siswa tersebut?

Jawab:

➤ X' = (jenis kelamin=p, kota asal=gresik, asal sekolah=smp trisila, status sekolah=terakreditasi B, danem SMP='b', program studi=ipa)

Look up tables

P(Jenis kelamin=p|kelulusan=A)= 2/5

P(Kota asal=gresik|kelulusan=A)=2/5

P(asal sekolah=smp trisila|kelulusan=A)= 1/5

P(Danem SMP=b'|kelulusan=A)= 0.456

P(Program Studi=ipa|kelulusan=A)= 2/5

P(Kelulusan = A) = 5/14

P(Jenis kelamin=p|kelulusan=B)= 2/5

P(Kota asal=gresik|kelulusan=B)= 1/5

P(asal sekolah=smp trisila|kelulusan=B)= 1/5

P(Danem SMP=b'|kelulusan=B)= 0.467

P(Program Studi=ipa|kelulusan=B)= 3/5

P(Kelulusan = B) = 5/14

P(Jenis kelamin=p|kelulusan=C)= 2/4

P(Kota asal=gresik|kelulusan=C)= 3/4

P(asal sekolah=smp trisila|kelulusan=C)= 1/4

P(Danem SMP=b'|kelulusan=C)= 0.657

P(Program Studi=ipa|kelulusan=C)= 2/4

P(Kelulusan = C) = 4/14

Map rule

- Likelihood of Kelulusan = A 2/5 x 2/5 x 1/5 x 0.456 x 2/5 x 5/14 = 0.002
- Likelihood of Kelulusan = B 2/5 x 1/5 x 1/5 x 0.467 x 3/5 x 5/14 = 0.0016
- Likelihood of Kelulusan = C 2/4 x 3/4 x 1/4 x 0.657 x 2/4 x 4/14 = 0.0087

Nilai probabilitas dapat dihitung dengan melakukan normalisasi terhadap *likelihood* tersebut sehingga jumlah nilai yang diperoleh = 1.

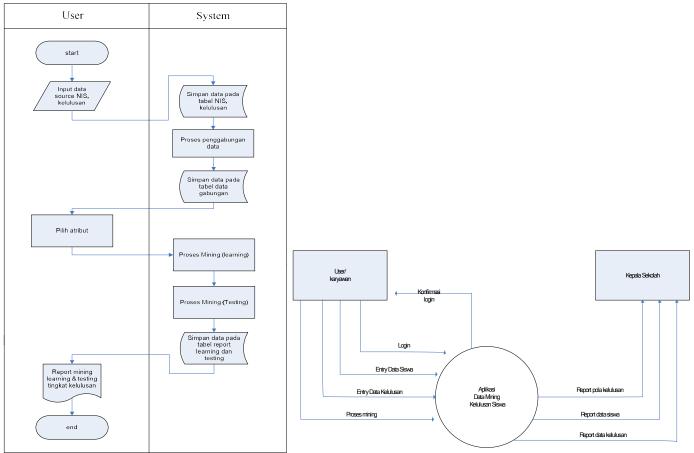
Given the fact P(A|x') < P(B|x') < P(C|x').

3.4. Perancangan Sistem

Berikut ini terdapat perancangan sistem dalam bentuk diagram alir tentang sistem mining tingkat kelulusan siswa pada gambar 1. dan diagram konteks penerimaan inklusi pada gambar 2.

4. Implementasi

Pada tahap implementasi menunjukkan hasil dari aplikasi data mining agar menghasilkan pola kelulusan siswa yang mana pada gambar 3 menunjukkan pengolahan data induk siswa sesuai dengan atributnya. Pada gambar 4 menunjukkan pengolahan data kelulusan dengan menunjukkan nilai dari masing-masing siswa. Sedangkan gambar 5 menunjukkan proses pembelajaran naïve bayes mulai dari proses *learning* dan *testing*.



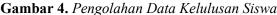
Gambar 1. Sistem Mining Tingkat Kelulusan Siswa

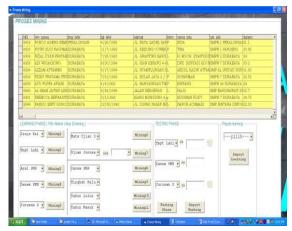
Gambar 2. Diagram Penerimaan Inklusi



Gambar 3. Pengolahan Data Induk siswa







Gambar 5. Proses Learning dan Testing

5. Kesimpulan dan Saran

Kesimpulan

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

- Proses testing digunakan untuk memprediksi data baru siswa tentang tingkat kelulusan yang akan diperoleh. Pada proses ini atribut yang digunakan adalah tempat lahir, danem SMP, jurusan studi. Atribut tersebut dipilih karena memiliki nilai perkalian support dan confidence yang tinggi dibandingkan atribut yang lain. Nilai masing masing atribut tersebut berdasarkan perhitungan antara lain adalah tempat lahir = 39,083, danem SMP= 23,828, dan jrurusan studi = 31,464
- 2. Pada analisa data yang dilakukan diproses testing, di dapat tingkat keakuratan sistem sekitar 99,82% dan memiliki nilai error 0.18% berdasarkan pengujian 220 data siswa tahun 2012.

Saran

Dari hasil evaluasi aplikasi yang telah dibuat, penulis menyadari bahwa aplikasi yang dibuat masih terdapat kekurangan. Report yang di hasilkan berupa nilai probabilitas, karena perhitungan yang dilakukan masih mengacu pada perhitungan data lama. 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 sekolah dalam menindak lanjuti perbaikan kualitas sekolahnya

6. DaftarPustaka

- [1] Davies, and Paul Beynon, 2004, "Database Systems Third Edition", Palgrave Macmillan, New York.
- [2] Elmasri, Ramez and Shamkant B. Navathe, 2000, "Fundamentals of Database Systems. Third Edition", Addison Wesley Publishing Company, New York.
- [3] Kadir, Abdul, 1999, "Konsep dan Tuntunan Praktis Basis Data", Penerbit Andi, Yogyakarta.
- [4] Kusrini, dan Emha Taufik Luthfi, 2009, "Algoritma Data Mining", Penerbit Andi, Yogyakarta.
- [5] Pramudiono, I. 2007. Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data. http://www.ilmukomputer.org/wp-content/uploads/2006/08/iko-datamining.zip Diakses pada tanggal 15 Maret 2009 jam 08.54.
- [6] Nurul Pratiwi, Oktariani. 2009. Klasifikasi Posting Blog Berbahasa Indonesia dengan Menggunakan Algoritma Naïve Bayes. Bandung : Universitas Pendidikan Indonesia.
- [7] Wibisono, Yudi. 2005. Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier1. Bandung: Universitas Pendidikan Indonesia.