

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

Budanis Dwi Meilani, Nofi Susanti

*Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Insitut Teknologi Adhi Tama
Surabaya*

Jl. Arief Rahman Hakim 100 – Surabaya 60117

E-mail : dwimeilanibudanis@yahoo.com, opick_susan2@ymail.com

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*

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 diimbangi dengan penyajian informasi yang memadai, seringkali informasi

tersebut masih harus digali ulang dari data yang jumlahnya sangat besar.

1.1. Latar Belakang

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 – 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 kelulusan. Dari pola tersebut, diharapkan bisa menganalisa faktor – faktor 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 menyusun strategi yang tepat dalam meningkatkan kualitas sekolah dan menjadikan sekolah memiliki daya saing yang tinggi.

1.2. Perumusan Masalah

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 untuk 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.

1.3 Tujuan Penelitian

1. Menerapkan teknik Data Mining dengan Metode *Naïve Bayes* dan menyajikan informasi kelulusan siswa
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 siswa

1.4 Batasan Masalah

1. Semua proses perhitungan yang disediakan oleh sistem menggunakan teknik data minning dengan metode Naive Bayes.

2. Informasi yang ditampilkan berupa laporan analisa 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 data Kelulusan yang diambil sebagai sampel dalam aplikasi ini 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 dengan menggunakan program aplikasi Visual Basic 6.0 dan perancangan database dengan menggunakan Microsoft Access 2003 pada sistem operasi Windows.

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 populer 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 filtering SPAM
 - Dan masalah klasifikasi lainnya
- Teorema Bayes:
 $P(C|X) = P(X|C) \cdot P(C) / P(X)$

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 siswa dan data kelulusan siswa dengan metode *classification naive 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*.

3.1.1 Learning Phase

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 untuk menentukan pola tingkat kelulusan siswa.

KATEGORI	KETERANGAN
A	danem 60.00 – 48.00
B	danem 47.99 – 36.00
C	danem 35.99 – 00.00

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

1. kelulusan A dengan nilai 60.00 – 48.00
2. kelulusan B dengan nilai 47.99 – 36.00
3. kelulusan C dengan nilai 35.99 – 00.00

3.3 Contoh Perhitungna Learning

3.3.1 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.

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 berjenis 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

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

(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.

Danem SMP	Kelulusa n = A	Kelulusa n = B	Kelulusa n = 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)	31.33	31.03	31.5
Std dev(σ)	7.14325	12.062	12.363

3.4 Contoh Proses Perhitungan Testing

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

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

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 :

- $X' = (\text{jenis kelamin}=p, \text{kota asal}=\text{gresik}, \text{asal sekolah}=\text{smp trisila}, \text{status sekolah}=\text{terakreditasi B}, \text{danem SMP}='b', \text{program studi}=\text{ipa})$

Look up tables

- $P(\text{Jenis kelamin}=p | \text{kelulusan}=A) = 2/5$
- $P(\text{Kota asal}=\text{gresik} | \text{kelulusan}=A) = 2/5$
- $P(\text{asal sekolah}=\text{smp trisila} | \text{kelulusan}=A) = 1/5$
- $P(\text{Danem SMP}='b' | \text{kelulusan}=A) = 0.456$
- $P(\text{Program Studi}=\text{ipa} | \text{kelulusan}=A) = 2/5$
- $P(\text{Kelulusan} = A) = 5/14$
- $P(\text{Jenis kelamin}=p | \text{kelulusan}=B) = 2/5$
- $P(\text{Kota asal}=\text{gresik} | \text{kelulusan}=B) = 1/5$
- $P(\text{asal sekolah}=\text{smp trisila} | \text{kelulusan}=B) = 1/5$
- $P(\text{Danem SMP}='b' | \text{kelulusan}=B) = 0.467$
- $P(\text{Program Studi}=\text{ipa} | \text{kelulusan}=B) = 3/5$
- $P(\text{Kelulusan} = B) = 5/14$

- $P(\text{Jenis kelamin}=p | \text{kelulusan}=C) = 2/4$
- $P(\text{Kota asal}=\text{gresik} | \text{kelulusan}=C) = 3/4$
- $P(\text{asal sekolah}=\text{smp trisila} | \text{kelulusan}=C) = 1/4$
- $P(\text{Danem SMP}='b' | \text{kelulusan}=C) = 0.657$
- $P(\text{Program Studi}=\text{ipa} | \text{kelulusan}=C) = 2/4$
- $P(\text{Kelulusan} = C) = 4/14$

Map rule

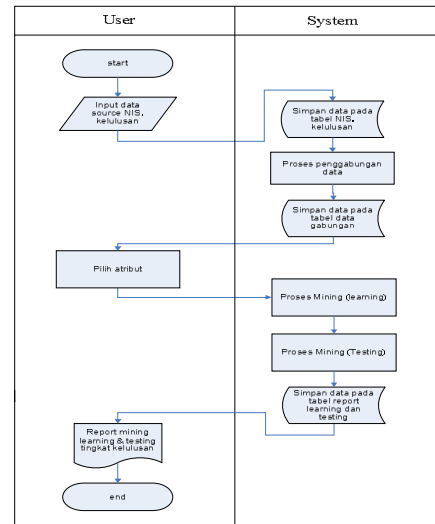
- Likelihood of Kelulusan = A
 $2/5 \times 2/5 \times 1/5 \times 0.456 \times 2/5 \times 5/14 = 0.002$
- Likelihood of Kelulusan = B
 $2/5 \times 1/5 \times 1/5 \times 0.467 \times 3/5 \times 5/14 = 0.0016$
- Likelihood of Kelulusan = C
 $2/4 \times 3/4 \times 1/4 \times 0.657 \times 2/4 \times 4/14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

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

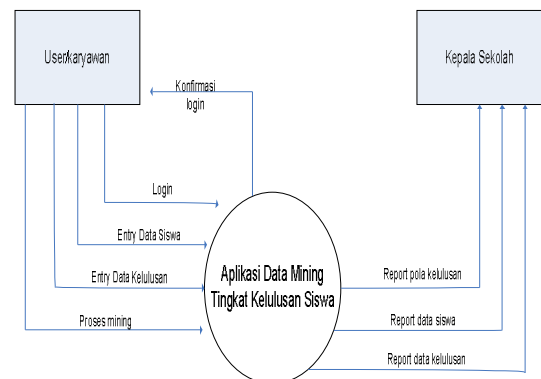
Given the fact $P(A|x') < P(B|x') < P(C|x')$.

3.5 Perancangan Sistem

1. Flowchart Sistem Mining Tingkat Kelulusan Siswa



1. Diagram Konteks Penerimaan Inklusi



4. IMPLEMENTASI

4.1 Pengolahan Data Induk Siswa

Gambar 4.1 Pengolahan Data Induk siswa

4.2 Data Kelulusan Siswa

Gambar 4.2. Pengolahan Data Kelulusan Siswa

4.3 Proses Mining

Gambar 4.3. Proses Mining Learning dan Mining Testing

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 data baru siswa tentang tingkat kelulusan yang akan diperoleh. Pada proses ini atribut yang digunakan adalah tempat lahir, dan 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, dan SMP= 23,828, dan jurusan studi = 31,464

1. 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.

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 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. DAFTAR PUSTAKA

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-](http://www.ilmukomputer.org/wp-content/uploads/2006/08/iko-datamining.zip)

[content/uploads/2006/08/iko-datamining.zip](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 Classifier*1. Bandung: Universitas Pendidikan Indonesia.