# PENERAPAN DATA MINING DALAM MENENTUKAN JURUSAN SISWA

**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

Penentuan jurusan bagi siswa SMA/MA sederajat merupakan proses untuk memfokuskan siswa dalam bidang konsentrasi tertentu, hal ini dilakukan agar setiap siswa dapat mempelajari lebih dalam mata pelajaran – mata pelajaran yang sesuai dengan jurusan yang telah ditentukan untuk setiap siswa. untuk menentukan jurusan siswa, maka diterapkanlah metode Naive Bayes dalam mengklasifikasikan jurusan siswa berdasarkan data yang dilatih, data tersebut akan ditentukan probabilitasnya baik yang menggunakan data dengan nilai string maupun numerik dan dari probabilitas tersebut dapat diprediksi jurusan yang sesuai untuk siswa. dalam penelitian ini ada 100 data siswa yang digunakan sebagai data untuk melihat keakuratan metode Naive Bayes dalam mengklasifikasikan jurusan siswa. hasilnya, dari 100 data siswa yang diuji, terdapat 90 data siswa yang berhasil diklasifikasikan dengan presentasi keberhasilan 90 % sedangkan 10 data siswa tidak berhasil diklasifikasikan.

**Kata Kunci**: Data Mining, Naive Bayes, Jurusan Siswa

## 1. Pendahuluan

### Latar Belakang

Perkembangan ilmu pengetahuan dan teknologi telah membawa perubahan di hampir semua aspek kehidupan manusia di mana berbagai permasalahan hanya dapat dipecahkan kecuali dengan upaya penguasaan dan peningkatan ilmu pengetahuan dan teknologi. Penentuan konsentrasi bagi siswa SMA/MA sederajat merupakan proses untuk memfokuskan siswa dalam bidang konsentrasi tertentu, hal ini dilakukan agar setiap siswa dapat mempelajari lebih dalam mata pelajaran – mata pelajaran yang sesuai dengan konsentrasi yang telah ditentukan untuk setiap siswa. Yang menjadi masalah ialah penulis ingin mendapatkan informasi dari histori nilai akademik siswa kelas 11 dan 12 Aliyah Swasta PAB 2 Helvetia sehingga setiap siswa kelas 10 dapat diklasifikasikan dalam kategori konsentrasi yang sesuai berdasarkan nilai yang mereka peroleh.

Hal ini juga kiranya telah menjadi bahan penelitian untuk kategori Sistem pendukung keputusan dalam menentukan jurusan di SMA yang sesuai dengan kemampuan siswa dengan dasar yang digunakan dalam penentuan jurusan adalah nilai semester, nilai potensi dan nilai pilihan siswa[1]. Penelitian lainnya seputar pemilihan jurusan juga penulis temukan, di mana dalam proses pemilihan jurusan ini digunakanlah metode 360 derajat[2]. Untuk melakukan perhitungan yang ada dalam penelitian ini maka digunakan teknik pengklasifikasian dengan metode *Naive Bayes.* 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[3]. Metode *Naive Bayes* juga dinilai berpotensi baik dalam mengklasifikasi dokumen dibandingkan metode pengklasifikasian yang lain dalam hal akurasi dan efisiensi komputasi [4].

### 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[5]. *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 [6]. Definisi lain mengatakan *Data Mining* adalah kegiatan yang meliputi pengumpulan, pemakaian data hostoris untuk menemukan keteraturan, pola atau hubungan dalam data berukuran besar [7]. 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[8]:

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[9]. 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[10].
*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[11].*

**Persamaan Metode Naive Bayes**

Persamaan dari teorema *Bayes* adalah (Bustami,2013) :

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 :

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 :

Untuk klasifikasi dengan data kontinyu digunakan rumus *Densitas Gauss* :

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.

Berikut langkah – langkah penyelesaian 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 :

atau

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 :

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 | Jenis Kelamin | Nilai IPA | Nilai IPS | Angket | Rekomendasi | Jurusan |
|---|---|---|---|---|---|---|
| 1 | L | 75 | 73 | IPA | IPA | IPA |
| 2 | P | 85 | 75 | IPA | IPA | IPA |
| 3 | P | 73,3 | 75 | IPS | IPS | IPS |
| 4 | L | 81,6 | 80,1 | IPA | IPA | IPA |
| 5 | P | 81 | 74,5 | IPA | IPA | IPA |
| 6 | P | 85 | 70 | IPA | IPA | IPA |
| 7 | P | 85,3 | 84 | IPA | IPA | IPA |
| 8 | P | 83,5 | 77,4 | IPA | IPA | IPA |
| 9 | P | 83,75 | 82,6 | IPS | IPS | IPS |
| 10 | L | 80 | 80 | IPS | IPS | IPS |
| 11 | P | 80,1 | 78,8 | IPS | IPS | IPS |
| 12 | L | 79,75 | 79,1 | IPS | IPS | IPS |
| 13 | P | 85 | 81,5 | IPA | IPA | IPA |
| 14 | P | 73,3 | 80 | IPS | IPS | IPS |
| 15 | L | 75,8 | 73,2 | IPA | IPA | IPA |
| 16 | L | 80,5 | 74,4 | IPS | IPA | IPS |
| 17 | L | 75 | 85 | IPS | IPA | IPS |
| 18 | P | 78 | 80 | IPA | IPS | IPS |
| 99 | P | 81,6 | 80,8 | IPS | IPS | IPS |
| 100 | P | 78 | 70,5 | IPS | IPA | IPA |

b. Kriteria dan Probabilitas

Adapun nilai probabilitas setiap kriteria didapatkan dari data latih pada tabel 1. Adapun nilai probabilitas setiap kriteria dapat dilihat pada pengujian dengan tools Weka di bawah ini.

**Gambar 3. Nilai Probabilitas Setiap Kriteria**

Dari nilai probabilitas di atas akan diuji data sebanyak 120 data siswa. Hasil uji coba dengan tools Weka untuk melihat seberapa akurat klasifikasi metode *Naive Bayes* dalam menentukan konsentrasi siswa dapat dilihat pada gambar 4 sebagai berikut.

**Gambar 4. Hasil Klasifikasi Naive Bayes**

Berdasarkan gambar 4 di atas, dapat diketahui dari 100 data siswa yang diuji dengan 6 buah kriteria sebagai pendukung pengklasifikasian di mana setiap kriteria memiliki nilai probabilitas tersendiri untuk setiap *class*-nya terdapat 90 data siswa yang berhasil diklasifikasikan dengan benar sementara sebanyak 10 data siswa tidak berhasil diklasifikasikan dengan benar. Dengan begitu keakuratan metode Naive Bayes dalam mengklasifikasikan 100 data siswa adalah sebesar 90%. Persentasi ini dapat dilihat pada gambar 5 berikut :

**Gambar 5. Presentase Keakuratan Metode Naive Bayes**

dilihat persentase untuk *Correctly Classified Instance* adalah sebesar 90 % sementara persentase untuk *Incorrectly Classified Instance* adalah sebesar 10%. Dengan *Confusion Matrix* untuk *class* IPA sebanyak 46 data siswa yang berhasil diklasifikasikan dan sebanyak 5 data siswa yang tidak berhasil diklasifikasikan. Sedangkan untuk *class* IPS sebanyak 44 data siswa yang diklasifikasikan dengan benar dan 5 data siswa yang tidak berhasil diklasifikasikan.

**3. Kesimpulan**

Berdasarkan penelitian tentang menentukan jurusan siswa dapat ditarik beberapa kesimpulan sebagai berikut :

1. Berdasarkan data akademik siswa yang diperoleh, proses *Data Mining* membantu dalam penerapan metode *Naive Bayes* dalam mendapatkan informasi dari hasil klasifikasi jurusan siswa.
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 jurusan siswa berdasarkan proses klasifikasi yang dilakukan oleh metode *Naive Bayes* itu sendiri.
3. Berdasarkan data akademik siswa yang dijadikan data *training*, metode *Naive Bayes* berhasil mengklasifikasikan 90 data siswa dari 100 data yang diuji. Sehingga dengan demikian metode N*aive Bayes* ini berhasil memprediksi jurusan siswa dengan persentase keakuratan sebesar 90 %.

**Daftar Pustaka**


[1] Tresna yudha Prawira, Dimara kusuma Hakim, (2011). *Sistem Pendukung Keputusan berbasis Web untuk Menentukan Penjurusan (IPA/IPS/Bahasa) pada SMA Islam Bumiayu,* JUITA ISSN : 2086-9398 Vol. 1 Nomor 4, November 2011.

[2] Stefanie G.N.L. Worang, Natalia K. Toeera, (2013), *Penerapan Metode 360 Derajat dalam Sistem Pendukung Keputusan penentuan Jurusan SMA Berbasis, Seminar nasional Aplikasi Teknologi Informasi (SNATI) 2013,15 Juni 2013,* Yogyakarta.

[3] 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.*

[4] 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.*

[5] 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.*

[6] 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.

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

[8] 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.

[9] 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.*

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

[11] 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.*