KLASIFIKASI JURUSAN MENGGUNAKAN METODE *NAÏVE BAYES* PADA SEKOLAH MENENGAH ATAS NEGERI (SMAN) 1 FATULEU TENGAH

Arddy H. Hailitik [1], Bertha S. Djahi [2], Yelly Y. Nabuasa [3]
[1,2,3] Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana

INTISARI

***Naïve bayes* merupakan metode pengklasifikasian yang memanfaatkan probabilitas dan statistik untuk memprediksi peluang di masa depan dengan memanfaatkan pengalaman di masa sebelumnya. Sistem penjurusan di Sekolah Menengah Atas (SMA) merupakan upaya untuk lebih mengarahkan siswa berdasarkan minat dan kemampuan akademiknya. Penjurusan pada SMA Negeri 1 Fatuleu Tengah terdiri dari jurusan IPA dan IPS. Penelitian ini menggunakan metode *naïve bayes* untuk mengklasifikasikan jurusan siswa. Data siswa yang digunakan merupakan data siswa kelas XI semester 2 tahun 2011-2015 dengan jumlah 470 data. Dalam proses pengujian digunakan 420 data (89%) sebagai data latih dan 50 data (11%) sebagai data uji. Hasil penelitian ini menunjukan akurasi sebesar 99.31% dalam proses pengklasifikasian jurusan.**

***Kata kunci*: *Naïve bayes*, Data *mining*, Pengklasifikasian jurusan**

ABSTRACT

Naïve bayes is the classification method which utilizes the both probabilities and statistics to predict the future opportunity by using the last experiance. The system of major in the senior high school is the means of students directing to be more based on their interest and academic competence. The major in East SMAN 1 Fatuleu consists of the Science and Social majors. This research is using the Method of Naïve bayesto classify the student major. The data of student that is used here is the grade XI for second semester in the years of 2011 to 2015 with the 470 for the total data. For the testing proces is used 420 data (89%) as trains data and 50 data (11%) as tests data. The result of this research shows the amount of 99.31% accuracy in the process of major classification.

Keyword: Naïve bayes, Data mining, the major classification.

I. PENDAHULUAN

Pemilihan jurusan bagi siswa SMA adalah upaya untuk mengenalkan siswa terhadap minat serta kemampuan akademik siswa. Banyak siswa yang bingung dalam memilih jurusan yang sesuai dengan minat atau kemampuan mereka. Hal ini juga dialami oleh para siswa di SMAN 1 Fatuleu Tengah, dimana mereka biasanya hanya berkonsultasi langsung dengan wali kelas atau dengan orang tua masing-masing. Setelah itu, pihak sekolah dalam hal ini wali kelas akan melakukan perhitungan data siswa berupa nilai rapor, nilai minat, dan nilai bakat untuk memutuskan jurusan yang tepat untuk siswa. Kegiatan ini membutuhkan waktu yang cukup lama karena perhitungan penentuan jurusan masih dilakukan secara manual. Untuk mengatasi masalah ini, terdapat beberapa metode yang dapat digunakan untuk menentukan jurusan pada SMAN 1 Fatuleu Tengah. Salah satunya dengan menggunakan metode klasifikasi menggunakan data siswa terdahulu sebagai acuan. Terdapat beberapa metode dalam pengklasifikasian antara lain algoritma ID3, C4.5, *K-Nearest Neighbor* (KNN), dan *naïve bayes*. Dalam penelitian ini, metode yang digunakan oleh peneliti adalah *naïve bayes* yang merupakan metode statistik sederhana dan memiliki akurasi yang baik dalam proses pengklasifikasian [6].

ISSN 2337-7631

ISSN 2337-7631

II. MATERI DAN METODE

2.1 *Dataset* siswa

Data siswa yang digunakan dalam penelitian ini adalah data siswa SMAN 1 Fatuleu Tengah kelas XI semester 2 yaitu data nilai rapor (matematika, fisika, kimia, biologi, geografi, ekonomi, sosiologi, sejarah), data nilai bakat (nilai IQ, verbal, numeral, spasial, persepsional, teknik) dan data nilai minat (minat orangtua dan minat siswa). Data tersebut akan digunakan sebagai parameter yang akan diolah dalam sistem klasifikasi jurusan menggunakan algoritma *Naive Bayes*. Data yang digunakan terdiri dari 470 data, dengan perbandingan 420 data (89%) sebagai data latih dan 50 data (11%) sebagai data uji.

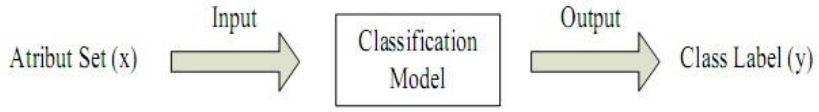
2.2 Sistem penjurusan di Sekolah Menengah Atas (SMA)

Pada pembelajaran tingkat SMA kita mengenal adanya sistem penjurusan. Penjurusan diperkenalkan sebagai upaya untuk lebih mengarahkan siswa berdasarkan minat dan kemampuan akademiknya. Hal ini diberlakukan karena siswa SMA berada pada jenjang yang strategis dan kritis bagi perkembangan dan masa depannya. Pada masa ini siswa berada di pintu gerbang untuk memasuki dunia perguruan tinggi yang merupakan wahana untuk membentuk integritas cita-cita yang diinginkan di masa mendatang [7].

Dalam penentuan jurusan di SMA ada tiga sistem penentuan yang digunakan dalam memilih jurusan, yaitu penilaian jurusan berdasarkan prestasi akademik berupa nilai rapor, penilaian jurusan berdasarkan minat dan penilaian jurusan berdasarkan nilai tes bakat.

2.3 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Proses klasifikasi bertujuan untuk membentuk suatu model yang mampu membedakan data kedalam kelas-kelas yang berbeda berdasarkan aturan fungsi [2].



Gambar 1. Blok diagram model klasifikasi

Gambar 1 menjelaskan bahwa *input*-an akan di klasifikasi dan menghasilkan *output* berupa label kelas. Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data uji digunakan untuk memperkirakan akurasi dari *rule* klasifikasi.

2.4 Algoritma *Naïve Bayes*

Algoritma *naive bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang ditemukan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai *Teorema Bayes.* Klasifikasi *naive bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas, tidak ada hubungannya dengan ciri dari kelas lainnya [2].

$$P(C|F_1 \ldots F_n) = \frac{P(C)P(F_1 \ldots . F_n|C)}{P(F_1 . \ldots F_n)} \quad (1)$$

Dimana variabel C merepresentasikan kelas, sementara variabel $F_{1 \ldots} F_n$ 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*). Dari penjelasan tersebut dapat dirumuskan sebagai berikut :

$$Posterior = \frac{Prior \times likelihood}{evidence} \quad (2)$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan.

2.5 Perhitungan probabilitas untuk klasifikasi jurusan

Dalam perhitungan nilai probabilitas terdapat 2 tipe data dari masing-masing parameter yaitu data yang bersifat *numeric* dan data yang bersifat *text*. Untuk menghitung nilai dari data-data tersebut, dapat digunakan persamaan sebagai berikut:

1. Menghitung nilai probabilitas untuk data *numeric*. Untuk menghitung jumlah dan proabilitas dari data yang bersifat *numeric* harus dicari terlebih dahulu nilai rata-rata hitung (*mean*) dan standar deviasi dari setiap parameter dari data yang memiliki data *numeric.*

$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \quad (3)$$

dimana:
μ : rata-rata hitung (*mean*)
x_i : nilai sampel ke –*i*
n : jumlah sampel
dan persamaan untuk menghitung nilai standar deviasi adalah sebagai berikut:

$$\sigma = \sqrt{\frac{(\sum_{i=1}^{n}(x_i - \mu)^2) + 1}{n-1}} \quad (4)$$

dimana:
σ : standar deviasi
x_i : nilai sampel ke –*i*
μ : rata-rata hitung (*mean*)
n : jumlah sampel

2. Menghitung nilai probabilitas untuk data *text.* Untuk menghitung jumlah dan probabilitas data yang bersifat *text* akan digunakan rumus *laplacian smoothing* dengan nilai K=1 untuk menghindari *P(x)* = 0.

$$P(x) = \frac{Count(x) + K}{N + K|x|} \quad (5)$$

dimana:
P : probabilitas dari variable *x*
Count (*x*) : jumlah kemunculan dari sampel *x*
K : parameter *smoothing*
N : jumlah total kejadian dari sampel *x*
$|x|$: jumlah kelas pada sampel

3. Menghitung probabilitas menggunakan rumus *dentitas gauss*. Rumus dentitas gaus yang digunakan adalah sebagai berikut:

$$\mathrm{P}(X|C) = \frac{1}{\sqrt{2\sigma\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (6)$$

dimana:
P : probabilitas
X : variabel
C : kelas
σ : standar deviasi

ISSN 2337-7631

μ : *mean*

4. Menghitung nilai *likelihood*. Menghitung nilai likelihood dilakukan untuk mendapatkan hasil akhir.

$$\mathrm{P}(X|\mathrm{C}) = P(F_1|C) \times P(F_2|C) \times \ldots \times P(F_n|C) \quad (7)$$

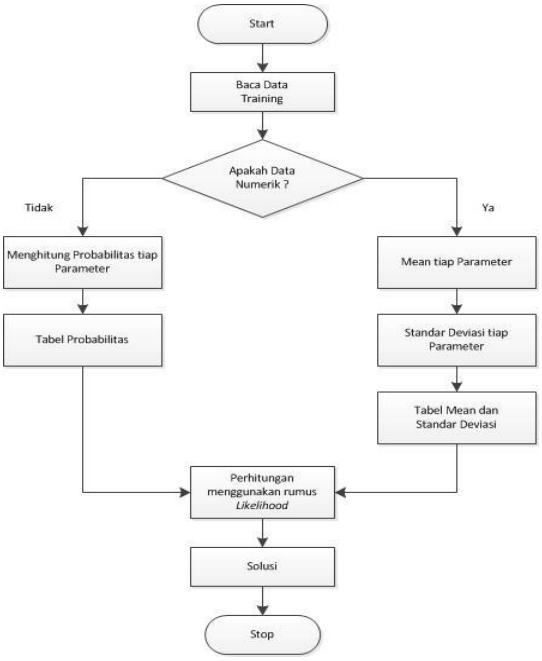
dimana:
P : probabilitas
X : variabel
C : kelas
F : atribut

5. Normalisasi. Pada proses normalisasi dalam klasifikasi data, data ditransformasi ke dalam interval yang ternormalisasi rentang nilai [-1..1] atau [0..1].
Untuk menghasilkan nilai probabilitas maka dilakukan normalisasi terhadap *likelihood* kelas IPA dan kelas IPS.

$$P(x) = \frac{Likelihood\ prior}{Likelihood\ prior + likelihoodposterior} \quad (8)$$

dimana:
P : probabilitas
X : variabel
$likelihood\ prior$: kelas
$likelihood\ posterior$: atribut

Secara garis besar, alur dari metode *naïve bayes* dapat dilihat pada gambar 2.



Gambar 2. *Flowchart naïve bayes*

Berdasarkan gambar 2, dapat dilihat bahwa proses pengklasifikasian *naïve bayes* dimulai dengan membaca data *training*, jika data tersebut merupakan data numerik maka akan dilanjutkan ke tahap perhitungan nilai *mean* dan standar deviasi dari tiap parameter, sedangkan apabila data

training tersebut bukan data numerik maka akan masuk ke tahap perhitungan probabilitas dari tiap parameter. Setelah itu akan dilanjutkan ke tahap perhitungan nilai *likelihood* dimana dari hasil perhitungan tersebut akan diperoleh solusi.

2.6 Kriteria Evaluasi

Untuk permasalahan dalam *binary classification*, kriteria evaluasi yang biasa digunakan adalah sebagai berikut:

1. *Precision*

Dalam *binary classification*, *precision* dapat disamakan dengan *positive predictive value* atau nilai kelas yang diklasifikasi secara benar. Rumus *precision* adalah:

$$Precision = \left(\frac{True\ Positif}{True\ positif + False\ positif}\right) \times\ 100\% \quad (9)$$

2. *Recall*

Recall adalah pengambilan data yang berhasil dilakukan terhadap bagian data yang relevan dengan *query*. Rumus *recall* adalah:

$$Recall = \left(\frac{True\ negatif}{False\ negatif + True\ negatif}\right) \times\ 100\% \quad (10)$$

3. *Accuracy*

Accuracy adalah persentase dari total jurusan yang benar diidentifikasi. Rumus *Accuracy* adalah:

$$Accuracy = \left(\frac{True\ positif + True\ negatif}{totaldata}\right) \times\ 100\% \quad (11)$$

4. *F1-Measure*

F1-measure merupakan nilai rata-rata dari *precision* dan *recall*. Skor *F1-measure* mencapai nilai terbaik pada 1 dan skor terburuk pada 0. Rumus menghitung *F1-measure* adalah:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

III. HASIL DAN PEMBAHASAN

3.1 Hasil pengujian menggunakan weka

Pengujian menggunakan weka dilakukan menggunakan 470 dataset yang dibagi dalam 5 kali pengujian dengan variasi data yang diambil secara manual.

Tabel 1. Hasil pengujian menggunakan weka

Data Uji ke-	***Total Data Testing***	***Precision***	***Recall***	***F1-Measure***	***Accuracy (%)***	***True Negative* (TN)**	***True Positive* (TP)**	***False Positive* (FP)**	***False Negative* (FN)**
1	50	1.000	1.000	1.000	100	27	23	0	0
2	100	0.99	0.99	0.99	99	50	49	1	0
3	150	0.993	0.993	0.993	99.33	77	72	1	0
4	200	0.995	0.995	0.995	99.5	104	95	1	0
5	250	0.996	0.996	0.996	99.6	128	121	1	0

Pada Tabel 1 menunjukan bahwa pengujian menggunakan weka dengan presentase tertinggi sebesar 100% diperoleh pada percobaan pertama dengan jumlah data 50 dengan

ISSN 2337-7631

parameter *precision* sebesar 1.000, *recall* sebesar 1.000, *F1-measure* sebesar 1.000, *true negative* sebesar 27, *true positif* sebesar 23, *false positif* sebesar 0, *false negative* sebesar 0. Pada pengujian selanjutnya dengan menggunakan jumlah data lebih besar dari 50, diperoleh akurasi terbaik sebesar 99.6% pada percobaan ke 5 dengan jumlah data 250 data, dengan nilai parameter *precision* sebesar 0.996, *recall* sebesar 0.996, *F1-measure* sebesar 0.996, *true negative* sebesar 128, *true positif* sebesar 121, *false positif* sebesar 1, *false negative* sebesar 0.

3.2 Hasil pengujian menggunakan model program

Pengujian menggunakan model program dilakukan menggunakan 470 dataset yang dibagi dalam 5 kali pengujian dengan variasi data yang diambil secara acak.

Tabel 2. Hasil pengujian menggunakan model program

Data Uji ke-	*Total Data*	*Precision*	*Recall*	*F1-Measure*	*Accuracy (%)*	*True Negative* (TN)	*True Positive* (TP)	*False Positive* (FP)	*False Negative* (FN)
1	50	1.000	1.000	1.000	100	27	23	0	0
2	100	0.98	1.000	0.99	99	49	50	1	0
3	150	0.987	1.000	0.993	99.33	74	75	1	0
4	200	0.989	0.989	0.989	99	110	88	1	1
5	250	0.992	0.992	0.992	99.2	121	127	1	1

Pada Tabel 2 menunjukan bahwa pengujian menggunakan model program memiliki persentase tertinggi pada pengujian ke 1 dengan nilai akurasi sebesar 100%. Pada pengujian ke 1 dengan jumlah data 50 diperoleh parameter *precision* sebesar 1.000, *recall* sebesar 1.000, *f1-measure* sebesar 1.000, *true negative* sebesar 27, *true positif* sebesar 23, *false positif* sebesar 0, *false negative* sebesar 0. Pada pengujian selanjutnya dengan jumlah data lebih besar dari 50 diperoleh akurasi terbaik sebesar 99.33% pada percobaan ke 3 dengan jumlah data 150 data, dengan nilai parameter *precision* sebesar 0.987, *recall* sebesar 1.000, *f1-measure* sebesar 0.993, *true negative* sebesar 74, *true positif* sebesar 75, *false positif* sebesar 1, *false negative* sebesar 0.

3.3 Analisis Hasil Pengujian

Berdasarkan hasil pengujian (Tabel 1 dan Tabel 2) maka diketahui bahwa penerapan metode *naïve bayes* pada pengujian untuk mengklasifikasi jurusan memiliki akurasi baik. Hal ini dapat ditunjukkan pada pengujian menggunakan weka diperoleh akurasi sebesar 100% pada percobaan ke 1 dan nilai akurasi terendah pada pengujian ke 2 sebesar 99% dimana pada proses klasifikasi terdapat 1 data yang tidak diklasifikasi secara tepat. Rata-rata akurasi pengujian menggunakan weka yaitu 99.49%. Pada pengujian menggunakan model program, diperoleh nilai akurasi tertinggi sebesar 100% pada pengujian ke 1 dan terdapat dua pengujian yang memiliki nilai akurasi terendah sebesar 99%, masing-masing pada pengujian ke 2 diperoleh hasil klasifikasi terdapat 1 data yang tidak diklasifikasi secara tepat dan pada pegujian ke 4 diperoleh hasil klasifikasi terdapat 2 data yang tidak diklasifikasi secara tepat. Dari pengujian ke 1 sampai pengujian ke 5 maka diperoleh rata-rata akurasi dalam pengklasifikasian jurusan menggunakan model program yaitu 99.31%.

IV. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan hasil pengujian maka dapat disimpulkan:

1. *Naïve bayes* dapat digunakan untuk klasifikasi jurusan dengan tingkat akurasi sebesar 99.31%.
2. Hasil pengujian menggunakan weka dan model program memiliki nilai tingkat kecocokan yang baik karena hasil persentase akurasinya tidak berbeda jauh yaitu 0.18% dimana weka memiliki akurasi sebesar 99.49% dan model program memiliki akurasi sebesar 99.31%.

4.2 Saran

Diperlukan adanya penelitian lanjutan menggunakan algoritma *naïve bayes* untuk pengklasifikasian dengan jumlah data dan parameter yang lebih banyak.

DAFTAR PUSTAKA

[1] Anandita, E. R., 2014, *Klasifikasi Tebu dengan Menggunakan Algoritma Naïve Bayes Classification pada Dinas Kehutanan dan Perkebunan Pati*, Universitas Dian Nuswantoro, Semarang.

[2] Bustami, 2014, *Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Data Nasabah Asuransi*, Universitas Malikussaleh, Lhokseumawe.

[3] Kristanto, O., 2014, *Penerapan Algoritma Klasifikasi Data Mining ID3 untuk Menentukan Penjurusan Siswa SMAN 6 Semarang*, Universitas Dian Nuswantoro, Semarang.

[4] Nugroho, Y. S., 2014, *Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro*, Universitas Dian Nuswantoro, Semarang.

[5] Rahayu, E. B., 2015, *Algoritma C4.5 untuk Penentuan Jurusan Siswa SMA Negeri 3 Pati*, Universitas Dian Nuswantoro, Semarang.

[6] Rosdiyansyah, S. F., 2012, *Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung menggunakan Naive Bayesian Classification*, Universitas Pendidikan Indonesia, Bandung.

[7] Rufaidah, A., 2015, *Pengaruh Intelegensi dan Minat Siswa Terhadap Putusan Pemilihan Jurusan*, Universitas Indraprasta PGRI, Jakarta Selatan.

[8] Saleh, A., 2015, *Implementasi Metode Klasifasi Naïve Bayes dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga*. Universitas Potensi Utama, Medan.

[9] Tefnai, M., 2016, *Klasifikasi Jurusan di Sekolah Menengah Atas Negeri (SMAN) 1 Fatuleu Tengah Menggunakan Metode K-Nearest Neighbor*, Universitas Nusa Cendana, Kupang.