METODE *NAIVE* BAYES DENGAN PENDEKATAN DISTRIBUSI GAUSS UNTUK KLASIFIKASI PEMINATAN PESERTA DIDIK

Nur Azizah[1*], Rito Goejantoro[1], Sifriyani[1]

[1]Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Indonesia

Corresponding author: nurazizahb01@gmail.com

Abstrak. Klasifikasi adalah suatu proses menilai objek data untuk memasukkan ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Salah satu teknik klasifikasi adalah *naive* Bayes. *Naive* Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang didasari oleh teorema Bayes dengan mengasumsikan kondisi antar atribut saling bebas. *Naive* Bayes dapat diterapkan pada data atribut kategorik maupun numerik, *naive* bayes mengasumsikan data kontinu ke dalam distribusi tertentu dan memperkirakan parameter distribusi dengan data latih. Penelitian ini menggunakan distribusi Gauss dalam memperkirakan parameter. Penelitian ini bertujuan untuk mendapatkan tingkat akurasi metode *naive* Bayes pada hasil peminatan peserta didik. Data yang digunakan adalah data peserta didik baru di MAN 2 Samarinda Jalan Harmonika Tahun Ajaran 2018/2019 dengan hasil minat yaitu IPA, IPS dan Bahasa. Digunakan 4 variabel bebas yaitu nilai IPA SMP, nilai IPS SMP, nilai Bahasa SMP dan rata-rata UN SMP. Hasil pengukuran akurasi dari metode *naive* Bayes memiliki akurasi yang baik pada klasifikasi hasil peminatan peserta didik yaitu 84% dan 71,05%.

Kata Kunci: Distribusi Gauss, klasifikasi, *naive* Bayes, peminatan peserta didik.

1 PENDAHULUAN

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas dari suatu objek (Mulyanto, 2009). Ada beberapa macam pengklasifikasian dalam *data mining* yaitu *decision tree, naive* Bayes, *Support Vector Machine (SVM)* dan lain-lain [3].

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya dengan asumsi independensi yang kuat (naif). Independensi yang kuat berarti kondisi antar atribut saling bebas atau tidak saling berkaitan [5]. Pada penelitian ini peneliti akan menghitung keakuratan dari metode *naive* Bayes.

Naive Bayes dapat diterapkan pada data atribut kategorik maupun kontinu. Pada data atribut yang bersifat kontinu, *naive* Bayes mengasumsikan data kontinu ke dalam distribusi tertentu dan memperkirakan parameter distribusi dengan data latih. Biasanya digunakan distribusi Gauss untuk menghitung probabilitas bersyarat dari atribut kontinu pada sebuah kelas. Parameter distribusi Gauss adalah rata-rata dan standar deviasi [1]. Distribusi Gauss digunakan dalam penelitian ini karena merupakan distribusi peluang yang paling umum digunakan dalam penelitian untuk metode *naive* Bayes pada data numerik dengan mengasumsikan data berdistribusi normal (Gauss). Salah satu penerapan klasifikasi *naive* Bayes pada data dengan atribut numerik yaitu pada bidang pendidikan.

Kementerian pendidikan dan kebudayaan terhitung sejak semester gasal 2013 memberlakukan Kurikulum 2013 pada jenjang pendidikan dasar hingga menengah. Pemberlakukan ini bertujuan untuk memperbaiki kualitas pendidikan nasional. Salah satu program utama dalam pelaksanaan program peserta didik adalah program peminatan peserta didik. Peminatan peserta didik merupakan suatu proses pengambilan pilihan dan keputusan oleh peserta didik dalam bidang keahlian yang didasarkan atas pemahaman potensi diri dan peluang yang ada [2].

Berdasarkan latar belakang yang diuraikan, penulis tertarik melakukan penelitian tentang *naive* Bayes pada data dengan atribut numerik yaitu pada bidang pendidikan untuk klasifikasi data peminatan siswa menengah atas dalam penelitian ilmiah yang berjudul "Metode *Naive* Bayes dengan Pendekatan Distribusi Gauss untuk Klasifikasi Peminatan Peserta Didik".

2 TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah proses menemukan pola dan hubungan dalam data. *Data mining* terdiri dari pengembangan model, yang biasanya merupakan representasi kompak dari pola yang ditemukan menggunakan data historis dan menerapkan model itu ke data baru [1].

2.2 Klasifikasi

Menurut [5], klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu

membedakan data ke dalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu.

2.3 Data *Training* dan Data *Testing*

Menurut [6], data untuk pengujian klasifikasi dibagi menjadi data *training* dan data *testing*. Data atau vektor yang sudah diketahui sebelumnya untuk label kelas dan digunakan untuk membangun model *classifier* disebut dengan data *training*. Data atau vektor yang belum diketahui (dianggap belum diketahui) label kelasnya menggunakan model *classifier* yang sudah dibangun disebut data *testing*. Jumlah data *training* dan data *testing* dapat dihitung menggunakan persamaan (1) dan persamaan (2).

$$\sum training = \text{Proporsi data } training \times \text{N} \quad (1)$$

$$\sum testing = \text{N - Jumlah data } training \quad (2)$$

2.4 *Naive* Bayes

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang berdasar pada teorema Bayes yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya dimana diasumsikan kondisi antar atribut saling bebas [5]. Persamaan dari Teorema Bayes adalah :

$$\mathrm{P(C \mid F)} = \frac{\mathrm{P(C)} \times \mathrm{P(F \mid C)}}{\mathrm{P(F)}} \quad (3)$$

di mana :

F : Data dengan kelas yang belum diketahui
C : Hipotesis data merupakan suatu kelas spesifik
$\mathrm{P(C \mid F)}$: Probabilitas hipotesis C dengan syarat F (probabilitas *posterior*)
$\mathrm{P(C)}$: Probabilitas hipotesis C (probabilitas *prior*)
$\mathrm{P(F \mid C)}$: Probabilitas hipotesis F dengan syarat C
$\mathrm{P(F)}$: Probabilitas hipotesis F

Untuk menjelaskan teorema Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, teorema Bayes di atas disesuaikan sebagai berikut :

$$P(C \mid F_1 \ldots F_n) = \frac{P(C).P(F_1 \ldots F_n \mid C)}{P(F_1 \ldots F_n)} \quad (4)$$

di mana variabel C mempresentasikan kelas, sementara variabel $\mathrm{F_1 ... F_n}$ mempresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel dengan karakteristik tertentu dalam kelas C (*posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikalikan dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara umum (disebut juga *evidence*). Karena itu, rumus dapat pula ditulis secara sederhana sebagai berikut :

$$Posterior = \frac{Prior \times likelihood}{evidence} \quad (5)$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel [6].

Untuk fitur bertipe numerik (kontinu), distribusi Gauss biasanya dipilih untuk merepresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(F_i \mid C)$, sedangkan distribusi Gauss dikarakteristikkan dengan dua parameter: *mean*, μ, dan variansi, σ^2. Untuk setiap kelas c_j, probabilitas bersyarat kelas c_j untuk fitur F_i adalah

$$P(F = f_i \mid C = c_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\left(\frac{(f_i-\mu_j)^2}{2\sigma_j^2}\right)} \quad (6)$$

Keterangan :

P : Peluang

f_i : Nilai atribut ke - *i*

c_j : kelompok ke - *j*

μ_j : rata-rata kelompok ke - *j*

σ_j^2 : variansi kelompok ke - *j*

Parameter μ_j bisa didapat dari *mean* sampel F_i $(\bar{f})$ dari semua data latih yang menjadi milik kelas c_j, sedangkan σ_j^2 dapat diperkirakan dari variansi sampel (s^2) dari data latih.[5].

2.5 Tingkat Akurasi Metode

Sebuah sistem klasifikasi harus diukur kinerjanya. Umumnya, cara mengukur kinerja klasifikasi menggunakan *confusion matrix*. *Confusion matrix* merupakan tabel yang mencatat hasil kerja klasifikasi. Setiap sel f_{ij} dalam *matrix* menyatakan jumlah *record*/data dari kelas *i* yang hasil prediksinya masuk ke kelas *j*. Misalnya sel f_{11} adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1 dan f_{10} adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0.

Tabel 1: *Confusion Matrix* untuk Klasifikasi Dua Kelas

f_{ij}		Kelas hasil prediksi (*j*)	
		Kelas = 1	Kelas = 0
Kelas Asli (*i*)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Berdasarkan isi *confusion matrix*, maka dapat diketahui jumlah data dari masing-masing kelas yang diprediksi secara benar yaitu ($f_{11}+f_{00}$). Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi dan prediksi yang dilakukan. Dua kuantitas ini digunakan sebagai *matrix* kinerja klasifikasi. Untuk menghitung akurasi digunakan persamaan sebagai berikut [6]:

$$akurasi = \frac{\sum data\ yang\ diprediksi\ secara\ benar}{\sum\ jumlah\ prediksi\ yang\ dilakukan} \times 100\%$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \times 100\% \quad (7)$$

2.6 Hakekat Peminatan

Peminatan adalah proses yang berkesinambungan untuk memfasilitasi peserta didik mencapai tujuan pendidikan nasional, dan oleh karena itu peminatan harus berpijak pada kaidah-kaidah dasar yang secara eksplisit dan implisit, terkandung dalam kurikulum.

Minat merupakan gejala psikologis, berkaitan dengan pikiran dan perasaan terhadap suatu objek. Dalam kaitannya dengan peminatan peserta didik dalam satuan pendidikan SMA, objek yang dimaksudkan adalah bidang peminatan matematika dan sains, sosial dan bahasa [2].

3 DATA

Data yang digunakan dalam penelitian ini merupakan data peserta didik baru di MAN 2 Samarinda Tahun Ajaran 2018/2019. Dari data tersebut dilakukan analisis menggunakan metode *naive* Bayes. Adapun variabel bebas yang digunakan adalah Nilai IPA SMP, Nilai IPS SMP, Nilai Bahasa SMP pada semester 5 dan rata-rata nilai UN SMP. Serta yang menjadi variabel terikat adalah hasil minat.

4 HASIL DAN PEMBAHASAN

4.1 Membagi Data *Training* dan Data *Testing*

Proporsi pembagian data *training* yang digunakan adalah 90:10 dan 70:30. Didapatkan hasil untuk proporsi 90:10, 227 data pertama hasil pengacakan data sebagai data *training* sedangkan 25 data terakhir sebagai data *testing.* Didapatkan hasil untuk proporsi 70:30, 176 data pertama hasil pengacakan data sebagai data *training* sedangkan 76 data terakhir sebagai data *testing.*

4.2 Membagi Data *Training* dan Data *Testing*

Tabel 2: Nilai *Posterior* Ketiga Kelompok untuk Proporsi 90:10

No	Kelompok 1	Kelompok 2	Kelompok 3	Hasil Klasifikasi
172	0,719	0,697	0,262	IPA
134	0,694	0,0904	0,18	
⋮	⋮	⋮	⋮	
No	Kelompok 1	Kelompok 2	Kelompok 3	Hasil Klasifikasi
164	0,0729	0,00279	0,0388	IPA
130	0,719	0,127	0,199	
115	0,0328	0,0708	0,0194	IPS
51	0,291	0,482	0,11	
⋮	⋮	⋮	⋮	
123	0,124	0,72	0,132	
108	0,0417	0,0304	0,058	

Hasil klasifikasi menggunakan metode *naive* Bayes pada nilai *posterior* dengan 25 data *testing* yang ada diperoleh hasil terdapat 16 peserta didik terpilih kedalam

peminatan IPA, 9 peserta didik terpilih kedalam peminatan IPS dan tidak terdapat peserta didik terpilih kedalam peminatan Bahasa.

$$\text{Akurasi} = \frac{(15+6)}{(15+0+0+0+6+0+1+3+0)} \times 100\%$$
$$= \frac{21}{25} \times 100\%$$
$$= 84\%$$

Dapat diketahui bahwa pada metode *naive* Bayes menunjukkan yang benar klasifikasi dalam memprediksi klasifikasi sebesar 84%. Hal ini menunjukkan bahwa tingkat akurasi pengklasifikasian metode *naive* Bayes baik.

4.3 **Klasifikasi *Naive* Bayes untuk Proporsi 70:30**

Tabel 3: Nilai *Posterior* Ketiga Kelompok untuk Proporsi 70:30

No	Kelompok 1	Kelompok 2	Kelompok 3	Hasil Klasifikasi
104	0,055200	0,004840	0,000654	IPA
31	1,080000	0,050600	0,054000	
⋮	⋮	⋮	⋮	
179	0,375000	0,054600	0,122000	
250	0,030200	0,163000	0,031900	IPS
79	0,024200	0,115000	0,038100	
⋮	⋮	⋮	⋮	
2	0,000757	0,031500	0,011000	

Hasil klasifikasi menggunakan metode *naive* Bayes pada nilai posterior dengan 76 data *testing* yang ada diperoleh hasil terdapat 51 peserta didik terpilih kedalam peminatan IPA, 25 peserta didik terpilih ke dalam peminatan IPS.

$$\text{Akurasi} = \frac{(38+16)}{(38+4+0+6+16+0+7+5+0)} \times 100\%$$
$$= \frac{54}{76} \times 100\%$$
$$= 71{,}05\%$$

Dapat diketahui bahwa pada metode *naive* Bayes menunjukkan yang benar klasifikasi dalam memprediksi klasifikasi sebesar 71,05%.

5 KESIMPULAN

Hasil pengukuran tingkat akurasi klasifikasi pada hasil minat pada data peserta didik baru MAN 2 Samarinda Tahun Ajaran 2018/2019 dengan metode *naive* Bayes sebesar 84% untuk proporsi 90:10 dan 71,05% untuk proporsi 70:30. Hal ini menunjukkan bahwa metode *naive* Bayes memberikan ketepatan prediksi klasifikasi sudah cukup baik dan jika data *training* diperbesar maka hasil klasifikasi yang didapatkan akan semakin baik

DAFTAR PUSTAKA

[1] Han, J., Micheline, K., dan Jian, P. (2012). *Data Mining Concepts and Techniques, Third Edition*. Waltham : Elsevier Inc

[2] Kemendikbud. (2013). *Materi Diklat Peminatan Peserta Didik*. Diakses dari

https://akhmadsudrajat.wordpress.com/2013/07/15/download-materi-diklat-peminatan-peserta-didik/comment-page-1/ diakses tanggal 19 Januari 2019 pukul 20:12

[3] Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey : John Wiley & Sons.

[4] Mulyanto, A. (2009). *Sistem Informasi Konsep dan Aplikasi*. Yogyakarta : Pustaka Pelajar.

[5] Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi menggunakan MATLAB Edisi 1*. Yogyakarta: ANDI.

[6] Prasetyo, E. (2014). *Data Mining, Mengolah Data Menjadi Informasi menggunakan MATLAB Edisi 1*. Yogyakarta: ANDI.