PEMETAAN SEBARAN ASAL SISWA DAN KLASIFIKASI JARAK ASAL SISWA SMA NEGERI DI KABUPATEN PRINGSEWU MENGGUNAKAN METODE *NAÏVE BAYES*

[1]Riska Aprilia, [1]Kurnia Muludi, [1]Aristoteles

[1]Jurusan Ilmu Komputer FMIPA Unila

Abstract

Distance students can be seen from the data stored student's address. Address high school students in the Pringsewu District have different distances. This research was conducted to determine the classification of the distance stored in high school students in the Pringsewu District. Distance students are classified to obtain five categories: "sangat dekat", "dekat", "sedang", "cukup jauh", and "jauh" by using eight attributes are "nomor", "SMA", "kabupaten", "kecamatan", "kelurahan", "jarak", "asal SMP", and "class". The classification performed by using Naive Bayes using Weka tool. Distribution of training data and testing data is different as much as 20 times of testing, resulting in the highest accuracy Naïve Bayes is 89.329% on distribution of 60% training data : 40% testing data. The data of students address and information classification results displayed in the form of digital map that is mapping of student's address in high school in the Pringsewu District.

Keywords: *Classification, Distance Students, Mapping Student, Naïve Bayes.*

1. Pendahuluan

Jarak siswa dapat diketahui dari data asal siswa yang tersimpan. Asal siswa yang berbeda-beda akan menghasilkan jarak yang berbeda pula. Data jarak siswa akan dijadikan salah satu atribut untuk klasifikasi jarak asal siswa SMA negeri di Kabupaten Pringsewu. Jarak asal siswa akan diklasifikasikan dengan menggunakan metode *Naïve Bayes* untuk mendapatkan lima *class* yaitu sangat dekat, dekat, sedang, cukup jauh, dan jauh.

Penelitian sebelumya menjadikan jarak sebagai salah satu atribut dengan menggunakan metode KNN dan *Naïve Bayes* menghasilkan akurasi *Naïve Bayes* sebesar 86% [1]. Pada penelitian ini juga dilakukan visualisasi data dengan peta digital [2]. Visualisasi data bertujuan untuk melihat sebaran asal siswa SMA negeri di Kabupaten Pringsewu. Hasil klasifikasi akan ditambahkan sebagai informasi pada peta sebaran asal siswa.

1.1 Klasifikasi

Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau *class* data, dengan tujuan untuk dapat memperkirakan *class* dari suatu objek yang labelnya tidak diketahui [3].

Target (*class/label*) pada klasifikasi berupa nilai kategorikal/nominal. Proses klasifikasi didasarkan pada empat komponen mendasar [4], yaitu:
1. Kelas (*Class*)
Variabel dependen dari model, merupakan variabel kategorikal yang merepresentasikan "label" pada objek setelah klasifikasinya. Contoh: adanya kelas penyakit jantung, loyalitas pelanggan, dan kelas gempa bumi (badai).

2. Prediktor (*Predictor*)
Variabel independen dari model, direpresentasikan oleh karakteristik (atribut) dari data yang akan diklasifikasikan dan berdasarkan klasifikasi yang telah dibuat. Contoh: tekanan darah, status perkawinan, catatan geologi yang spesifik, kecepatan dan arah angin, musim, dan lokasi terjadinya fenomena.

3. Pelatihan dataset (*Training dataset*)
Kumpulan data yang berisi nilai-nilai dari kedua komponen sebelumnya dan digunakan untuk melatih model dalam mengenali *class* yang cocok/sesuai, berdasarkan prediktor yang tersedia. Contoh: kelompok pasien yang diuji pada serangan jantung, kelompok pelanggan supermarket (diselidiki oleh intern dengan jajak pendapat), database yang berisi gambar untuk monitoring teleskopik dan pelacakan objek astronomi, database badai.

4. Dataset Pengujian (*Testing Dataset*)
Berisi data baru yang akan diklasifikasikan oleh (*classifier*) model yang telah dibangun di atas sehingga akurasi klasifikasi (*model performance*) dapat dievaluasi.

1.2 Naïve Bayes

Naïve Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma ini mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel *class* [5]. Keuntungan penggunan *Naïve Bayes* adalah hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian [1]. Persamaan dari teorema Bayes adalah [6]

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad \text{.....(1)}$$

Di mana:

X:	Data dengan *class* yang belum diketahui.
H:	Hipotesis data merupakan suatu class spesifik.
$P(H\|X)$:	Probabilitas hipotesis H berdasarkan kondisi X (posterior probabilitas).
$P(H)$:	Probabilitas hipotesis H (prior probabilitas) terjadi tanpa memandang bukti apapun.
$P(X\|H)$:	Probabilitas X berdasarkan kondisi pada hipotesis H.
$P(X)$:	Probabilitas kondisi X (prior probabilitas) terjadi tanpa memandang kondisi yang lain.

Proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan *class* apa yang cocok bagi sampel yang dianalisis. Karena itu, metode *Naive Bayes* di atas disesuaikan sebagai berikut:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad \text{..... (2)}$$

Di mana Variabel C merepresentasikan *class*, sementara variabel F1 ... Fn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam *class* C (Posterior) adalah peluang munculnya *class* C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada *class* C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik - karakteristik sampel secara global (disebut juga evidence). Dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad \text{..... (3)}$$

Nilai Evidence selalu tetap untuk setiap *class* pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai-nilai posterior *class* lainnya untuk menentukan ke *class* apa suatu sampel akan diklasifikasikan [7].

1.3 Confusion matrix

Confusion matrix merupakan tabel pencatat hasil kerja klasifikasi. *Confusion matrix* melakukan pengujian untuk memperkirakan objek yang benar dan salah. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya [4]. *Confusion matrix* berisi informasi aktual (*actual)* dan prediksi (*predicted*) pada sistem klasifikasi. Tabel 1 contoh tabel *confusion matrix* yang menunjukkan klasifikasi dua kelas.

Tabel 1 *Confusion Matrix*

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True positives count (TP)	False negatives count (FN)
	Negative	False positive count (FP)	True negative count (TN)

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *false negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negatif. Data uji yang dimasukkan ke dalam *confusion matrix*, akan dihitung nilai-nilai *recall*, *precision* dan *accuracy* [8].

$$p = \frac{TP}{TP+FP} \quad \text{.....(6)}$$

$$r = \frac{TP}{TP+FN} \quad \text{.....(7)}$$

Precision (*p*) = jumlah sampel berkategori positif diklasifikasi benar dibagi dengan total sampel yang diklasifikasi sebagai sample positif.

Recall (*r*) = jumlah sampel diklasifikasi positif dibagi total sampel dalam testing set berkategori positif.

Akurasi dapat diperoleh dengan menggunakan dua persamaan yaitu:

$$accurasy = \frac{TP+TN}{TP+TN+FN+FP} \quad \text{.....(8)}$$

$$akurasi = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ benar}{jumlah\ prediksi\ yang\ dilakukan} \quad \text{.....(9)}$$

1.4 Jarak

Jarak merupakan sesuatu yang harus ditempuh dari suatu lokasi yang lain. Jarak dapat dinyatakan dengan jarak mutlak dan jarak nisbi. Jarak mutlak dinyatakan dalam satuan unit ukuran fisik seperti mil, km, meter, dan sebagainya [9]. Jarak dari tempat tinggal ke setiap prasarana mempunyai standar yang berbeda. Standar jarak dan waktu tempuh untuk sarana fasilitas pendidikan menurut konsep *Neighborhood Unit* dapat dibagi menjadi lima kategori yang ditunjukkan pada Tabel 2 berikut [10].

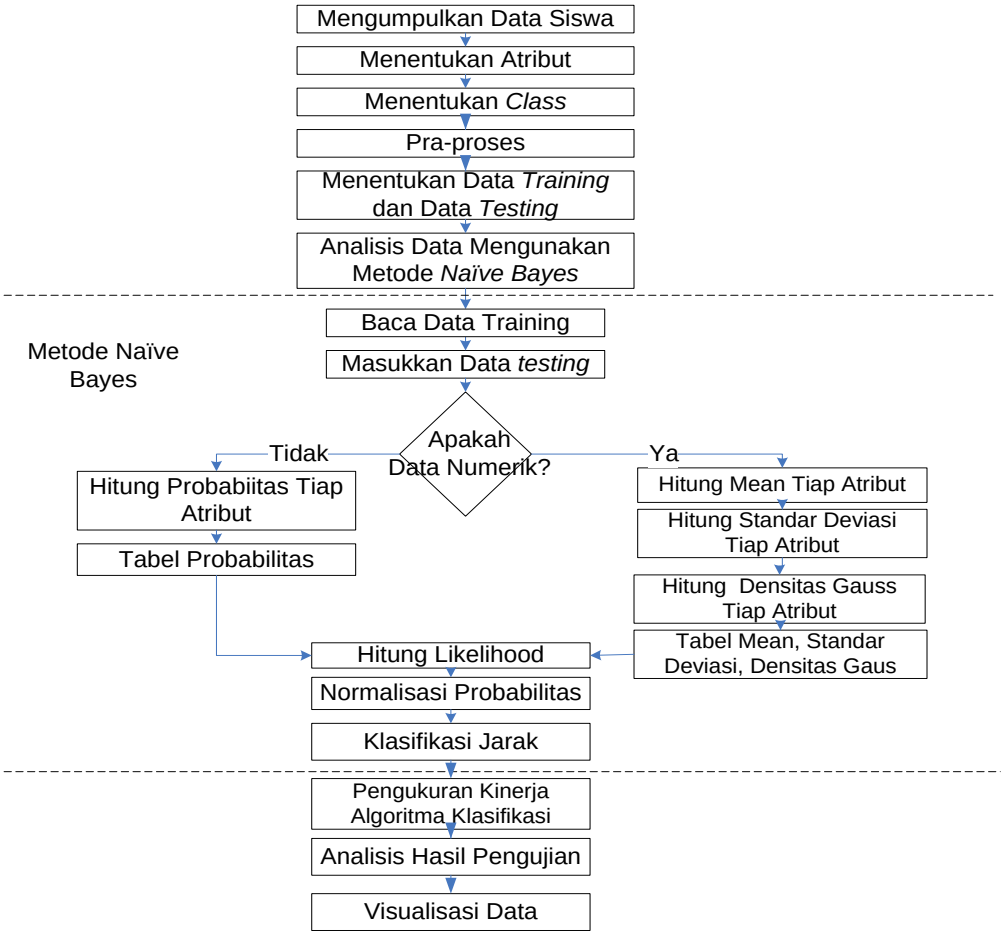
Tabel 2 Kategori Jarak Jangkauan dan Waktu Tempuh

No	Kategori jarak	Jarak tempuh (meter)	Waktu tempuh
1	Sangat Dekat	0-300 meter	0-5 menit
2	Dekat	300-600 meter	5-10 menit
3	Sedang / Cukup	600-1200 meter	10-20 menit
4	Cukup Jauh	1200-3000 meter	20-40 menit
5	Jauh	>3000 meter	>40 menit

Sumber : Udjianto, 1994 dalam Takumangsang 2010

2. Metode Penelitian

Metode pada penelitian ini mengunakan metode *Naïve Bayes.* Diagram alir metode *Naïve Bayes* ditunjukkan pada Gambar 1.



Gambar 1 Diagram Alir Penelitian dengan Metode *Naïve Bayes*

Penelitian ini diawali dengan melakukan pengumpulan data dari 10 SMA negeri di Kabupaten Pringsewu. Data yang didapat berjumlah 5516 data. Kemudian menentuan atribut yaitu NO, SMA, KAB, KEC, KEL, JARAK, ASAL SMP dan *ClASS*. Penjelasan dari masing-masing atribut ditunjukkan pada Tabel 3 berikut.

Tabel 3 Atribut Klasifikasi

Atribut Klasifikasi	Keterangan
NO	Berisi nomor sebagai pembeda pada setiap data.
SMA	Berisi nama SMA negeri di Kabupaten Pringsewu.
Kabupaten (KAB)	Berisi kabupaten asal siswa di SMA negeri di Kabupaten Pringsewu.
Kecamatan (KEC)	Berisi kecamatan asal siswa di SMA negeri di Kabupaten Pringsewu.
Kelurahan (KEL)	Berisi kelurahan asal siswa di SMA negeri di Kabupaten Pringsewu.
Jarak	Berisi jarak asal siswa dari rumah ke masing-msing sekolah dengan berjalan kaki dalam satuan kilometer (km). Jarak yang digunakan adalah jarak mutlak.
Asal SMP	Berisi asal SMP siswa di SMA negeri di Kabupaten Pringsewu.
Class	Berisi *class*/ label yang merepresentasikan jarak asal siswa.

Setelah atribut dibentuk, dilakukan pelabelan sesuai dengan *class* yang telah ditentukan yaitu sangat dekat, dekat, sedang, cukup jauh, dan jauh yang masing-masing nilainnya ditunjukkan pada Tabel 2 tentang kategori jarak tempuh. Kemudian dilakukan pra-proses klasifikasi dengan perapihan data sebagai berikut.

1. Penghapusan kata yang tidak dipakai dalam klasifikasi.
   Contoh : Jl. Satria, Pringsewu Barat → Pringsewu Barat
2. Konversi menjadi huruf kapital.
   Contoh : Pringsewu Barat → PRINGSEWU BARAT
3. Melakukan perbaikan data apabila ada kata yang diperlukan tetapi data tidak sesuai.
   Contoh: ada kata “gemuk mas” kemudian diubah menjadi “GUMUKMAS”.
4. Melakukan penghapusan spasi
   Contoh : PRINGSEWU BARAT → PRINGSEWUBARAT

Setelah itu, dilakukan pembersihan data pada data yang mengandung *missing value*. Setelah dilakukan pra-proses, data siswa berjumlah 5438 data. Setelah itu, dilakukan pembagian data menjadi dua bagian yaitu *training* dan *testing* dengan format yang diubah dari file bentuk .csv menjadi .arff pada Weka. Setelah itu, dilakukan implementasi algoritma *Naïve Bayes*. Tahap terakhir adalah visualisasi dengan peta digital yaitu peta sebaran asal siswa dengan informasi berupa *class* hasil klasifikasi.

3. Hasil dan pembahasan

3.1 Penerapan metode Naive Bayes

1. Baca data *training*
   Contoh:

Tabel 4 Data *Training*

No	SMA	Kab	Kec	Kel	Jarak (km)	Asal SMP	*Class*
1	SMAN 1 Adiluwih	Pesawaran	Negeri katon	Bangun sari	4.8	SMPN 1 Adiluwih	Jauh
2	SMAN 1 Adiluwih	Pesawaran	Negeri katon	Trirahayu	7.8	MTs Guppi Trirahayu	Jauh
3	SMAN 1 Adiluwih	Pringsewu	Adiluwih	Adiluwih	3.5	SMPN 1 Adiluwih	Cukup Jauh
470	SMAN 1 Ambarawa	Pringsewu	Ambarawa	Ambarawa	1.5	SMPN 2 Ambarawa	Cukup Jauh
471	SMAN 1 Ambarawa	Pringsewu	Ambarawa	Sumber Agung	7.2	SMPN 1 Ambarawa	Jauh
1093	SMAN 1 Banyumas	Pringsewu	Sukoharjo	Sukoharjo III	8.8	SMP PGRI 1 Sukoharjo	Jauh
1178	SMAN 1 Banyumas	Pringsewu	Banyumas	Sriwungu	1	SMPN 1 Sukoharjo	Sedang
1374	SMAN 1 Pagelaran	Tanggamus	Air Naningan	Air Naningan	35.4	SMPN 1 Talang Padang	Jauh
252	SMAN 1 Pardasuka	Pringsewu	Pardasuka	Kedaung	9.5	SMPN 1 Pardasuka	Jauh
253	SMAN 1 Pardasuka	Pringsewu	Pardasuka	Tanjung Rusia	4	SMPN 1 Pardasuka	Jauh
2027	SMAN 1 Sukoharjo	Pringsewu	Sukoharjo	Siliwangi	9.2	SMP PGRI 1 Sukoharjo	Jauh
2028	SMAN 1 Sukoharjo	Pringsewu	Adiluwih	Adiluwih	4.2	SMPN 1 Adiluwih	Jauh
2031	SMAN 1 Sukoharjo	Pringsewu	Adiluwih	Waringin sari Timur	4.1	SMPN 2 Adiluwih	Jauh
4772	SMAN 2 Pringsewu	Pringsewu	Pringsewu	Pringsewu Utara	1.8	SMP Muhamma diyah 1 Pringsewu	Cukup Jauh
4700	SMAN 2 Pringsewu	Pringsewu	Pringsewu	Podosari	0.3	MTsN 1 Pringsewu	Sangat Dekat
2637	SMAN 1 Pringsewu	Pringsewu	Pringsewu	Pringsewu Utara	0.6	SMPN 1 Pringsewu	Dekat
2638	SMAN 1 Pringsewu	Pringsewu	Pringsewu	Pringsewu Utara	0.5	SMPN 3 Pringsewu	Dekat
3754	SMAN 2 Gadingrejo	Pringsewu	Gading rejo	Tegal Sari	3	MTs Pelita Gedong Tataan	Cukup Jauh
3884	SMAN 1 Gadingrejo	Pringsewu	Gading rejo	Tegal Sari	1.1	SMPN 1 Gading rejo	Sedang
3879	SMAN 1 Gadingrejo	Lampung Tengah	Bangun rejo	Sidorejo	26.7	SMPN 1 Bangun Rejo	Jauh

2. Masukkan Data *Testing*

Contoh, dicari *class* dari data *testing* berikut:

No 3892, SMAN1 Gadingrejo, Kabupaten Pringsewu, Kecamatan Gadingrejo, Kelurahan Tegalsari, jarak 1.1 km, dan asal SMP N 1 Gadingrejo

3. Perhatikan jenis data pada data *testing*

- Data non-numerik/diskrit

Cari nilai probabilistik pada data *testing* yang bersifat non-numerik/diskrit dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.

Contoh :

Probabilitas dari setiap kelas.

- P(class= "Sangat Dekat") $= \frac{1}{20} = 0.05$
- P(class= "Dekat") $= \frac{2}{20} = 0.1$
- P(class= "Sedang") $= \frac{2}{20} = 0.10$
- P(class= "Cukup Jauh") $= \frac{4}{20} = 0.2$
- P(class= "Jauh") $= \frac{11}{20} = 0.55$

Probabilitas dari setiap *class* untuk atribut data diskrit.

- P(SMA = "SMAN 1 Gadingrejo" | class= "Sangat Dekat") $= \frac{0}{1} = 0$
- P(SMA = "SMAN 1 Gadingrejo" | class= "Dekat") $= \frac{0}{2} = 0$
- P (SMA = "SMAN 1 Gadingrejo" | class= "Sedang") $= \frac{1}{2} = 0.5$
- P(SMA = "SMAN 1 Gadingrejo" | class= "Cukup Jauh") $= \frac{0}{4} = 0$
- P(SMA = "SMAN 1 Gadingrejo" | class= "Jauh") $= \frac{1}{11} = 0.090$
- P(Kab = "Pringsewu" | class= "Sangat Dekat") $= \frac{1}{1} = 1$
- P(Kab = "Pringsewu" | class= "Dekat") $= \frac{2}{2} = 1$
- P(Kab = "Pringsewu" | class= "Sedang") $= \frac{1}{2} = 0.5$
- P(Kab = "Pringsewu" | class= "Cukup Jauh") $= \frac{4}{4} = 1$
- P(Kab = "Pringsewu" | class= "Jauh") $= \frac{7}{11} = 0.63$
- P(Kec = "Gadingrejo" | class= "Sangat Dekat") $= \frac{0}{1} = 0$
- P(Kec = "Gadingrejo" | class= "Dekat") $= \frac{0}{2} = 0$
- P(Kec = "Gadingrejo" | class= "Sedang") $= \frac{1}{2} = 0.5$
- P(Kec = "Gadingrejo" | class= "Cukup Jauh") $= \frac{1}{4} = 0.25$
- P(Kec = "Gadingrejo" | class= "Jauh") $= \frac{1}{11} = 0.090$
- P(Kel = "Tegalsari" | class= "Sangat Dekat") $= \frac{0}{1} = 0$
- P(Kel = "Tegalsari" | class= "Dekat") $= \frac{0}{2} = 0$
- P(Kel = "Tegalsari" | class= "Sedang") $= \frac{1}{2} = 0.5$
- P(Kel = "Tegalsari" | class= "Cukup Jauh") $= \frac{1}{4} = 0.25$

- P(Kel = "Tegalsari" | class= "Jauh") $= \frac{0}{11} = 0$
- P(AsalSMP= "SMPN 1 Gadingrejo" | class= "Sangat Dekat") $= \frac{0}{1} = 0$
- P(AsalSMP= "SMPN 1 Gadingrejo" | class= "Dekat") $= \frac{0}{2} = 0$
- P(AsalSMP= "SMPN 1 Gadingrejo" | class= "Sedang") $= \frac{1}{2} = 0.5$
- P(AsalSMP= "SMPN 1 Gadingrejo" | class= "Cukup Jauh") $= \frac{0}{4} = 0$
- P(AsalSMP= "SMPN 1 Gadingrejo" | class= "Jauh") $= \frac{0}{11} = 0$

• Data numerik/kontinu

Cari nilai *mean*, standar deviasi, dan densitas gauss dari masing-masing parameter yang merupakan data numerik/kontinu.

- Persamaan menghitung *mean*:

$$\mu = \frac{\sum_{i=1}^{n} xi}{n} \quad \text{..... (1)}$$

atau

$$\mu = \frac{x1+x2+x3+\cdots+xn}{n} \text{..... (2)}$$

Contoh : Jarak Sedang → $\mu = \frac{1.1+1}{2} = 1.05$

- Persamaan untuk menghitung nilai simpangan baku (standar deviasi):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi-\mu)^2}{n-1}} \quad \text{..... (3)}$$

Contoh : Jarak→ $\sigma = \sqrt{\frac{(1.1-1.05)^2+(1-1.05)^2}{2-1}} = 0.071$

• Persamaan fungsi Densitas Gauss:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{..... (4)}$$

Contoh: $f(x) = \frac{1}{\sqrt{2\pi}\times 0.070710678} e^{\frac{-(5-1.05)^2}{2\times(0.070710678)^2}} = 4.29E+48$

Nilai Densitas Gauss yang didapat digunakan dalam perhitungan likelihood.

4. Mendapatkan nilai dalam tabel *mean*, standar deviasi, densitas gauss dan probabilitas.

Hasil *mean*, standar deviasi, densitas gauss menggunakan persamaan (2), (3), dan (4) dapat dilihat pada Tabel 5 dan Tabel 6. sedangkan hasil probabilitas ditunjukkan pada Tabel 7 berikut.

Tabel 5 Hasil Perhitungan *Mean* dan Standar deviasi, dan Densitas Gauss jarak siswa

Jarak	Sangat Dekat	Dekat	Sedang	Cukup Jauh	Jauh
Mean	0.300	0.550	1.050	2.450	11.063

Standar Deviasi	0	0.071	0.071	0.954	10.286
Densitas Gauss	0	7.74E+13	4.29E+48	11.320	0.069

Tabel 6 Hasil Perhitungan *Mean* dan Standar deviasi, dan Densitas Gauss no siswa.

No	Sangat Dekat	Dekat	Sedang	Cukup Jauh	Jauh
Mean	4700.000	2637.500	2531.000	2249.750	1219.182
Standar Deviasi	0	0.500	1913.077	2369.241	942.673
Densitas Gauss	0	0	0.000269	0.000214	0.02357

Tabel 7 Probabilitas dari setiap *class* untuk atribut data diskrit.

Atribut	Sangat Dekat	Dekat	Sedang	Cukup Jauh	Jauh
P(X)	$^1/_{20}$	$^2/_{20}$	$^2/_{20}$	$^4/_{20}$	$^{11}/_{20}$
P(SMAN1 Gadingrejo|X)	$^0/_1$	$^0/_2$	$^1/_2$	$^0/_4$	$^1/_{11}$
P(Pringsewu| X)	$^1/_1$	$^2/_2$	$^1/_2$	$^4/_4$	$^7/_{11}$
P(Gadingrejo|X)	$^0/_1$	$^0/_2$	$^1/_2$	$^1/_4$	$^1/_{11}$
P(Tegalsari|X)	$^0/_1$	$^0/_2$	$^1/_2$	$^1/_4$	$^0/_{11}$
P(SMPN3 Pringsewu|X)	$^0/_1$	$^0/_2$	$^1/_2$	$^0/_4$	$^0/_{11}$

5. Perhitungan likelihood

Perhitungan likelihood dari data *testing* yang diberikan adalah sebagai berikut.

P(X|Sangat dekat) = $0 \times 0 \times \left(\frac{1}{20}\right) \times \left(\frac{0}{1}\right) \times \frac{1}{1} \times \frac{0}{1} \times \frac{0}{1} \times \frac{0}{1} = 0$

P(X|Dekat) = $(7.74\text{E} + 13) \times 0 \times \left(\frac{2}{20}\right) \times \left(\frac{0}{2}\right) \times \frac{2}{2} \times \frac{0}{2} \times \frac{0}{2} \times \frac{0}{2} = 0$

P(X|Sedang) = $(4.29\text{E} + 48) \times 0.000269 \times \left(\frac{1}{20}\right) \times \left(\frac{1}{2}\right) \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = 3.6\text{E} + 43$

P(X|Cukup jauh) = $11.31973 \times 0000214 \times \frac{4}{20} \times \frac{0}{4} \times \frac{4}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{0}{4} = 0$

P(X|Jauh) = $0.069176 \times 0.02357 \times \frac{11}{20} \times \frac{1}{11} \times \frac{7}{11} \times \frac{1}{11} \times \frac{0}{11} \times \frac{0}{11} = 0$

6. Normalisasi Probabilitas

Mengambil keputusan sebuah data *testing* masuk ke dalam *class* apa, perlu dilakukan normalisasi probabilitas.

Probabilitas Sangat Dekat = $\frac{0}{0+0+(3.6\text{E}+43)+0+0} = 0$

Probabilitas Dekat = $\frac{0}{0+0+(3.6\text{E}+43)+0+0} = 0$

Probabilitas Sedang = $\frac{3.6\text{E}+43}{0+0+(3.6\text{E}+43)+0+0} = 1$

Probabilitas Cukup Jauh = $\frac{0}{0+0+(3.6E+43)+0+0} = 0$

Probabilitas Jauh = $\frac{0}{0+0+(3.6E+43)+0+0} = 0$

Kesimpulan *class* dari data *testing* yang diberikan adalah Sedang.

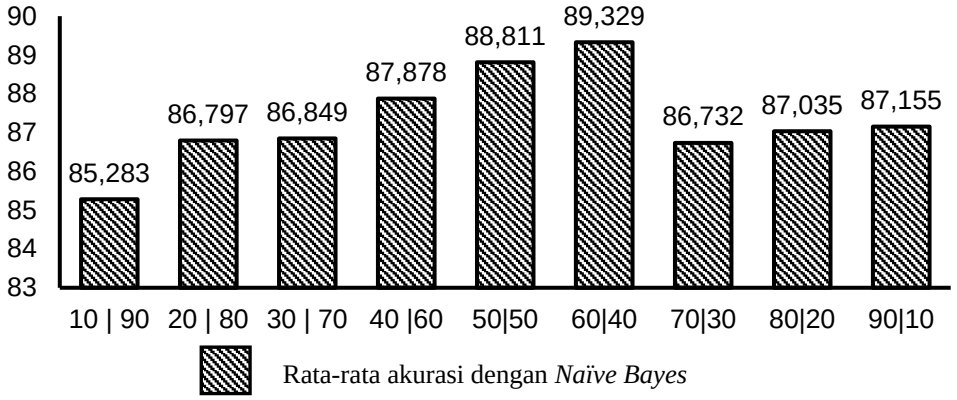
3.2 Pengujian Metode Naive Bayes

Proses awal klasifikasi jarak asal siswa SMA negeri di Kabupaten Pringsewu untuk memprediksi jarak asal siswa masuk ke dalam *class* sangat dekat, dekat, sedang, cukup jauh, dan jauh adalah pembagian data. Pembagian data dilakukan dengan menggunakan teknik *HoldOut* yaitu membagi data menjadi dua bagian yaitu *training* dan *testing* untuk mendapatkan akurasi yang tinggi. Pembagian data *training* dan data *testing* dilakukan dengan perbandingan *training*:*testing* sebesar 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, dan 10:90 [11]. Setiap perbandingan/percobaan dilakukan pengujian sebanyak 20 kali dengan data yang berbeda di setiap pengujian. Akurasi hasil pengujian dengan metode *Naive Bayes* ditunjukkan pada Tabel 8 berikut.

Tabel 8 Akurasi Hasil Pengujian setiap percobaan dengan *Naive Bayes.*

Pengujian	Data *Training* : Data *Testing*								
	10:90	20:80	30:70	40:60	50:50	60:40	70:30	80:20	90:10
1	93.011	91.864	92.985	83.936	86.061	87.454	87.439	89.154	88.419
2	90.376	77.109	81.371	82.894	83.965	86.765	81.618	86.213	86.581
3	89.027	78.419	88.650	94.175	85.362	79.734	84.130	86.489	88.971
4	76.262	84.165	84.104	81.484	84.995	94.761	87.868	87.500	85.294
5	75.235	92.483	80.894	89.761	87.164	79.825	85.294	88.695	87.132
6	76.931	92.529	85.737	83.911	87.164	95.770	90.619	87.765	93.370
7	78.407	77.614	94.169	84.861	84.222	83.172	85.408	88.041	88.398
8	91.152	84.552	79.695	84.156	87.422	87.678	88.106	85.741	84.530
9	91.050	84.029	82.817	87.374	88.488	82.207	86.695	86.937	86.556
10	93.524	93.218	94.115	82.317	85.693	89.241	86.573	87.673	79.742
11	88.221	91.467	91.507	88.804	90.427	92.516	86.300	88.991	89.599
12	89.489	93.444	81.104	90.365	92.268	91.873	87.140	85.073	89.416
13	93.171	80.682	83.885	86.851	90.611	92.854	86.169	87.431	85.192
14	91.965	81.436	91.827	94.847	88.660	89.302	86.957	86.618	88.665
15	75.726	93.191	83.623	90.798	91.053	91.322	86.108	88.451	85.949
16	75.388	91.456	83.447	88.736	91.844	92.354	88.315	89.391	85.240
17	76.889	93.246	93.912	89.685	91.183	91.847	87.400	85.161	85.000
18	78.277	80.381	92.629	87.027	92.135	92.400	86.408	82.673	88.148
19	91.464	80.466	81.212	94.735	96.141	89.590	82.852	84.055	86.507
20	90.096	94.189	89.294	90.848	91.360	95.908	93.239	88.653	90.388
Rata-rata akurasi	85.283	86.795	86.849	87.878	88.811	89.329	86.732	87.035	87.155

Tabel 8 berisi hasil 20 kali pengujian dalam satuan persen (%) dengan Metode *Naïve Bayes*. Nilai yang didapat berupa nilai akurasi *class* yang benar diklasifikasikan oleh Weka. Hasil akurasi pada Metode *Naïve Bayes* di setiap pengujian menghasilkan akurasi yang berbeda-beda pada setiap pembagian data *training* dan *testing* yang berbeda. Rata-rata akurasi dengan *Naïve Bayes* ditunjukkan pada Gambar 2.



Gambar 2 Rata-rata akurasi dengan *Naïve Bayes*

Akurasi dari Metode *Naïve Bayes* pada pembagian data *training* dan data *testing* yang berbeda sebanyak 20 kali pengujian menghasilkan rata-rata akurasi diatas 85%. Akurasi *Naïve Bayes* tertinggi adalah 89.329% pada pembagian data *training* 60% dan data *testing* 40%.

3.2 Pengukuran Kinerja Algoritma

Pengukuran kinerja algoritma digunakan *precision* dan *recall* yang dapat dihitung dari Tabel *Confusion matrix*.

Tabel 9 *Confusion Matrix* Klasifikasi Jarak untuk Menghitung *Precision, Recall,* dan Akurasi pada pembagian data *training*:*testing* 60:40 pada pengujian ke-20 dengan Metode *Naïve Bayes*

		predicted class					
	Class	Jauh	cukupjauh	sedang	dekat	sangatdekat	total *actual*
actual class	Jauh	1388	61	0	0	0	1449
	cukupjauh	4	605	0	8	0	617
	sedang	0	6	41	6	0	53
	dekat	0	0	4	38	0	42
	sangat dekat	0	0	0	0	**14**	14
	total *predicted*	1392	672	45	52	14	

1. *Precision* =
   a. *Precision* Jauh = $\frac{1388}{1392} = 0.9971$
   b. *Precision* Cukup Jauh = $\frac{605}{672} = 0.9003$
   c. *Precision* Sedang = $\frac{41}{45} = 0.9111$

d. *Precision* Dekat = $\frac{38}{52} = 0.7308$
e. *Precision* Sangat Dekat = $\frac{14}{14} = 1$

Precision menjelaskan nilai ketepatan sebuah *class* sebenarnya di antara *class* prediksi/hasil klasifikasi yang diberikan oleh sistem. Sebagai contoh pada *class* dekat, terdapat 38 data aktual *class* dekat yang diprediksi ke dalam *class* dekat, akan tetapi ada 6 data aktual *class* sedang, dan 8 data aktual cukup jauh yang diprediksikan juga ke dalam *class* dekat. Sehingga nilai ketepatan sebuah *class* sebenarnya di antara *class* prediksi adalah 38 data aktual dari 52 data.

2. Recall =

a. *Recall* Jauh = $\frac{1388}{1449} = 0.9579$
b. *Recall* Cukup Jauh = $\frac{605}{617} = 0.9806$
c. *Recall* Sedang = $\frac{41}{53} = 0.7736$
d. *Recall* Dekat = $\frac{38}{42} = 0.9048$
e. *Recall* Sangat Dekat= $\frac{14}{14} = 1$

Recall menjelaskan tingkat keberhasilan sistem dalam menemukan/menghasilkan prediksi yang sesuai dengan *class* sebenarnya. Sebagai contoh pada *class* sedang, terdapat 53 data aktual *class* sedang, akan tetapi hanya 41 data yang diprediksi ke dalam *class* sedang dan sisanya yaitu 6 data diprediksi ke dalam *class* cukup jauh dan 6 data diprediksi ke dalam *class* dekat. Sehingga keberhasilan sistem menghasilkan prediksi yang sesuai dengan *class* sebenarnya adalah 41 data dari 53 data aktual.

3. Akurasi = $\frac{2086}{2175} = 95.908\%$

Akurasi menjelaskan seberapa banyak data aktual yang benar diklasifikasikan oleh sistem dengan ketentuan jumlah data yang benar diklasifikasikan sistem dibagi jumlah data keseluruhan.

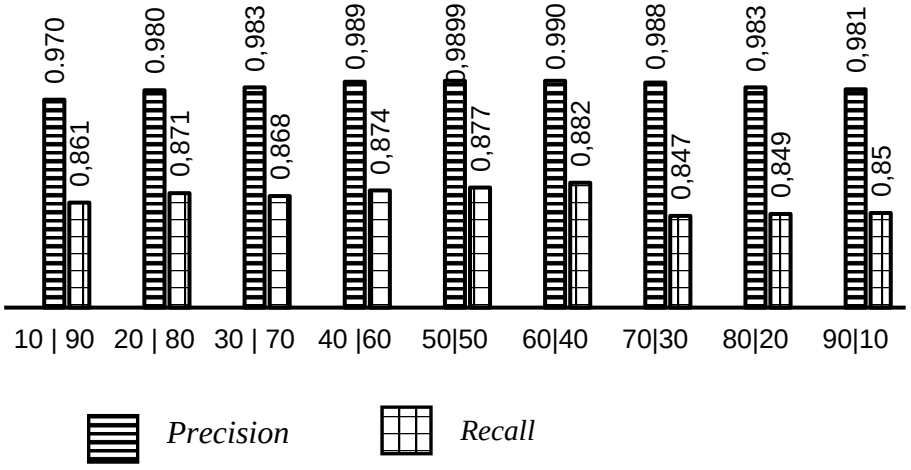
Rata-rata presentase nilai *precision* dan *recall* metode *Naïve Bayes* pada setiap percobaan untuk *class* jauh ditunjukkan pada Tabel 10.

Tabel 10 Rata-rata Presentase Nilai *Precision* dan *Recall* Metode *Naïve Bayes* untuk *Class* Jauh

Rata-rata	*Training*:*Testing*								
	10 | 90	20 | 80	30 | 70	40 |60	50|50	60|40	70|30	80|20	90|10
Precision	0.970	0.980	0.983	0.989	0.989	0.990	0.988	0.983	0.981
Recall	0.861	0.871	0.868	0.874	0.877	0.882	0.847	0.849	0.850

Tabel 10 menunjukkan presentase nilai *precision* tertinggi terjadi pada pembagian data *training* 60% dan data *testing* 40% yaitu sebesar 0.990. Presentase nilai *recall* tertinggi juga

terjadi pada pembagian data *training* 60% dan data *testing* 40% yaitu sebesar 0.882. Perbedaan rata-rata presentase nilai *precision* dan *recall* ditunjukkan pada Gambar 3.



Gambar 3 Rata-Rata Presentase *Precision* dan *Recall* metode *Naïve Bayes*

untuk *Class* Jauh

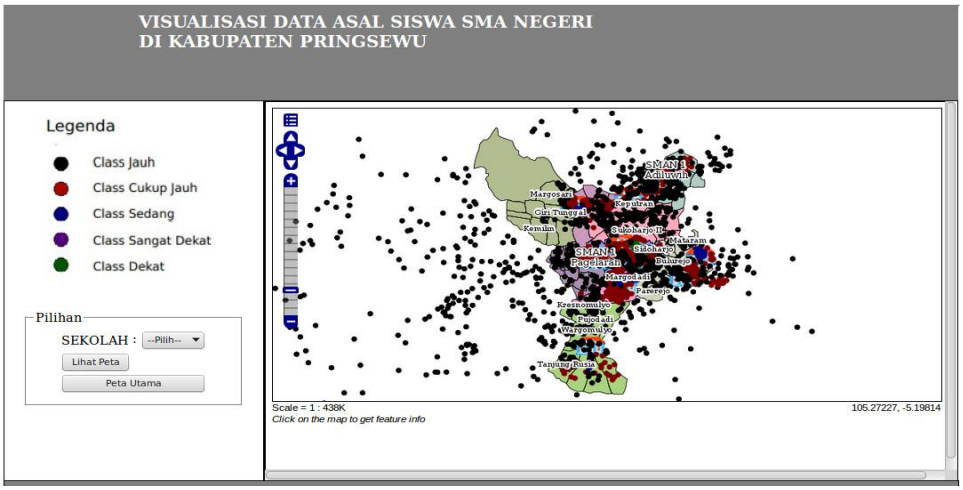
Tabel 10 dan Gambar 3 menunjukkan presentase *precision* dan *recall Naïve Bayes* untuk *class* jauh pada setiap percobaan di atas 0.800. Di mana presentase nilai *precision* tertinggi terjadi pada pembagian data *training* 60% dan data *testing* 40% yaitu sebesar 0.990. Presentase nilai *recall* tertinggi juga terjadi pada pembagian data *training* 60% dan data *testing* 40% yaitu sebesar 0.882.

Hal ini menunjukkan pada *Naïve Bayes*, jumlah data *training* dan data *testing* tidak menjadi patokan menghasilkan akurasi yang tinggi, Selain itu, hasil *precision* dan *recall* yang tinggi akan menghasilkan akurasi yang tinggi pula.

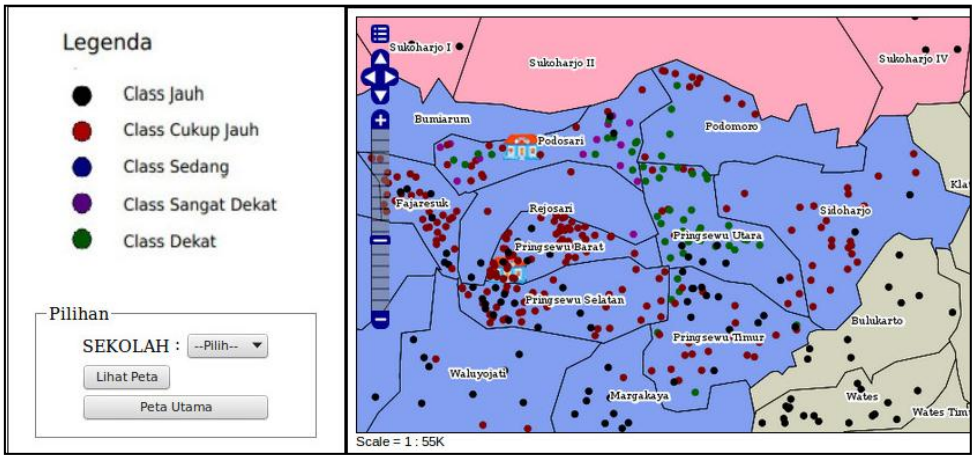
Hasil *precision*, *recall* dan akurasi menunjukkan bahwa metode *Naïve Bayes* cukup sesuai diterapkan pada penelitian ini karena menggunakan pendekatan statistik yaitu peluang. Setiap atribut bersifat independen, dan tidak membutuhkan jumlah data *training* yang besar untuk menghasilkan akurasi yang tinggi.

Hasil klasifikasi pada pembagian data *training*:*testing* 60:40 pada pengujian ke-20 dengan metode *Naïve Bayes* divisualisasikan dalam peta digital yaitu peta sebaran asal siswa SMA negeri di Kabupaten Pringsewu. Peta yang ditampilkan adalah peta sebaran asal siswa dan informasi *class* hasil klasifikasi. Hasil visualisasi ditunjukkan pada Gambar 4, Gambar 5, dan

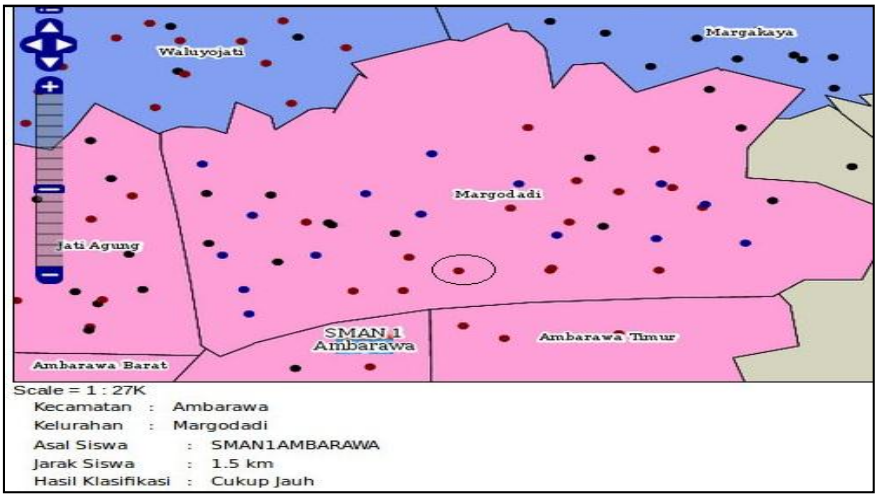
Gambar 6.



Gambar 4 Visualisasi Hasil Klasifikasi dengan *Naïve Bayes*



Gambar 5 Perbedaan Warna Visualisasi Hasil Klasifikasidengan *Naïve Bayes*



Gambar 6 Visualisasi Hasil Klasifikasi dengan *Naïve Bayes*
pada Satu *Point*

4. Kesimpulan

Kesimpulan yang didapat berdasarkan penelitian yang telah dilakukan adalah sebagai berikut.

1. Setelah 20 kali pengujian diperoleh rata-rata akurasi tertinggi pada *Naïve Bayes* sebesar 89.329% pada pembagian data *training* dan data *testing* 60:40.
2. Pada proses klasifikasi dengan *Naïve Bayes* jumlah data *training* tidak mempengaruhi tingginya nilai akurasi.

5. Referensi

[1] Satoto, B. D,. dan Yasid, A. 2015. Aplikasi Sales Report untuk Klasifikasi Area Penjualan Menggunakan K-Nearest Neighbor dan Naive Bayes Berbasis Android. Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015). ISSN: 2089-9815.

[2] Rodiyansyah, S. F. dan Winarko, E. 2013. Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. Indonesian Journal of Computing and Cybernetics Systems (IJCCS). Vol.7, No.1. ISSN: 1978-1520.

[3] Han, J., Kamber, M. dan Pei, J. 2011. Data Mining Concept and Techniques Third Edition. SanFrancisco: Morgan Kauffman.

[4] Gorunescu, F. 2011. Data Mining Concept Model Technique. Romania: Springer.

[5] Ridwan, M., Suyono, H. dan Sarosa, M. 2013. Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Electrical, Electronics, control, communications, and informatics Seminar. Vol.7, No. 1.

[6] Kusrini dan Luthfi, E.T. 2009. Algoritma data mining. Yogyakarta: Andi Offset.

[7] Saleh, Alfa. 2015. Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. Creative Information Technology Journal. Vol. 2, No. 3. ISSN:2354-5771.

[8] Defiyanti, Sofi. 2013. Analisis dan Prediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining. Syntax Vol. 2 Ed. 1.

[9] Daldjoeni. 1997. Geografi Baru-Organisasi Keruangan dalam Teori dan Praktek. Alumni. Bandung.

[10] Takumangsang, Esli D. 2010. Kajian Penempatan Fasilitas Pendidikan Dasar dan Menengah dalam Aspek Sistem Informasi Geografis. TEKNO Vol. 08 No.54.

[11] Nasution, N., Djahara, K. dan Zamsuri, A. 2015. Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes (Studi Kasus: Fasilkom Unilak). Jurnal Teknologi Informasi & Komunikasi Digital Zone. Vol.6, No.2.