PREDIKSI KELANJUTAN STUDI SISWA KE PERGURUAN TINGGI DENGAN NAIVE BAYES

Gentur Wahyu Nyipto Wibowo[1], Zaenal Arifin[2], Muhammad Anwarudin Romli[3], Nurul Ikhsanul Amal[4]
Fakultas Sains dan Teknologi
Universitas Islam Nahdlatul Ulama Jepara
gentur@unisu.ac.id

ABSTRACT

The need for an analysis of the predictions of continuation of student studies to college is a strong reason for this research. Because by knowing the number of students at a school who continue or not continue studies to universities become a reference to improve education services at the school concerned. Naive Bayes is an effective and efficient classification algorithm for data mining and machine learning. So in the research proposed Naive Bayes for predictions of continuation of student studies to college with k-fold validation and confusion matrix. And results from the algorithm Naive Bayes is 86.53%.

Keywords : continuation of student studies to college, Naive Bayes.

ABSTRAK

Kebutuhan akan analisa mengenai prediksi kelanjutan studi siswa ke perguruan tinggi menjadi alasan yang kuat untuk diadakannya penelitian ini. Karena dengan mengetahui jumlah siswa pada suatu sekolah yang melanjutkan atau tidak melanjutkan studi ke perguruan tinggi menjadi acuan untuk meningkatkan pelayanan pendidikan pada sekolah yang bersangkutan. Naive Bayes merupakan algoritma klasifikasi yang efektif dan efisien untuk data mining dan machine learning. Sehingga dalam penelitian ini mengusulkan Naive Bayes untuk memprediksi kelanjutan studi siswa dengan pengujian *k-fold validation*, dan *confusion matrix*. Dimana didapatkan hasil dari algoritma Naive Baye 86.53%.

Kata kunci :kelanjutan studi siswa ke perguruan tinggi, Naive Bayes.

PENDAHULUAN

Latar Belakang

Data mining mencakup berbagai teknik yang telah dikembangkan baik dalam bisnis maupun penelitian. Teknik ini telah digunakan untuk kebutuhan industri, pendidikan, komersial dan ilmiah. Data mining melibatkan banyak algoritma yang berbeda untuk menyelesaikan tugas yang berbeda. Algoritma-algoritma ini mencoba menyesuaikan model dengan data. Dan memeriksa data serta menentukan model yang paling dekat karakteristiknya dengan data yang sedang diperiksa. Tujuan utama data mining dalam praktik cenderung prediksi dan deskripsi [1]. Beberapa teknik, diantaranya klasifikasi dan clustering. Teknik klasifikasi adalah teknik pembelajaran yang digunakan untuk memprediksi nilai dari atribut kategori target [2]. Klasifikasi bertujuan untuk membagi objek yang ditugaskan hanya ke salah satu nomor kategori yang disebut kelas [3]. *Clustering* mengelompokkan objek atau data berdasarkan kemiripan antar data, sehingga anggota dalam satu kelompok memiliki banyak kemiripan dibandingkan dengan kelompok lain [4]. Penggunaan metode klasifikasi data mining yang paling populer digunakan untuk teknik klasifikasi adalah *Decision Trees, Naïve Bayes* (NB), *Statistical analysis*, dan lain lain [4].

Naive Bayes adalah suatu klasifikasi berpeluang yang berdasarkan aplikasi teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen) [5]. Dalam hal ini, diasumsikan bahwa kehadiran atau ketiadaan dari suatu kejadian tertentu dari suatu kelompok tidak berhubungan dengan kehadiran atau ketiadaan dari kejadian lainnya. Naïve Bayes merupakan algoritma klasifikasi yang efektif dalam mendapatkan hasil yang akurat dan efisien dalam proses penalaran memanfaatkan input yang ada dengan cara yang relatif cepat [6]. Algoritma ini bertujuan untuk melakukan klasifikasi data pada klas tertentu. Beberapa penelitian yang menggunakan Naive Bayes dalam berbagai macam keperluan antara lain untuk klasifikasi dokumen [7], deteksi *spam* atau *filtering spam* [8], klasifikasi sms [9], dan masalah klasifikasi lainnya.

Dikarenakan Naïve Bayes dalam mengklasifikasi lebih efektif dan hasil yang

didapatkan akurat serta efisien, maka Naïve Bayes akan digunakan dalam melakukan prediksi seberapa banyak siswa lulusan SMA yang melanjutkan ke perguruan tinggi, baik perguruan tinggi negeri maupun swasta. Karena kelanjutan pendidikan ke jenjang yang lebih tinggi ikut berperan dalam meningkatkan kemajuan suatu bangsa. Oleh karena itu, pendidikan harus terus menerus diperbaiki baik dari segi kualitas maupun kuantitasnya. Seiring dengan berjalannya waktu dan pembangunan di bidang pendidikan. Kompetisi antar sekolahpun juga semakin ketat, sekolah-sekolah yang berhasil meningkatkan jumlah peserta didiknya ke jenjang perguruan tinggi akan menjadi tujuan pilihan utama sebagai sekolah kelanjutan dari sekolah tingkat dasar. Sehingga prediksi mengenai jumlah peserta didik yang melanjutkan ke pergurutan tinggi dianggap sebagai sesuatu yang penting dikarenakan dengan mengetahui prosentase siswa yang melanjutkan ke perguruan tinggi dapat dijadikan acuan untuk perbaikan mutu pelayanan pendidikan di SMA. Karena salah satu tujuan penyelenggaraan pendidikan menengah di Sekolah Menengah Atas (SMA) adalah meningkatkan pengetahuan siswa untuk melanjutkan pendidikan pada jenjang yang lebih tinggi. Berdasarkan kurikulum tahun 1994 program pengajaran di SMA terbagi menjadi tiga program pengajaran khusus yang dapat dipilih oleh siswa sesuai bakat dan kemampuannya yaitu program IPA, IPS dan Bahasa [10].

Batasan Masalah

Data yang digunakan dalam penelitian ini adalah data *private* hasil rekap siswa SMA Negeri 1 Pecangaan dengan jumlah 313 record dengan 12 atribut. Penentuan atribut mengacu pada penelitian sebelumnya yaitu faktor-faktor yang mempengaruhi siswa melanjutkan studi ke perguruan tinggi [10]. Sehingga atribut yang terdapat pada hasil rekap siswa meliputi: jenis kelamin, pendapatan orang tua, pekerjaan orang tua, PIP, pendidikan ayah, pendidikan ibu, nilai bahasa Indonesia, nilai bahasa Inggris, nilai Matematika, nilai jurusan, nilai ijazah, dan lanjut yang akan dijadikan label. Dalam melakukan eksperimen menggunakan bantuan *tools RapidMiner 5.3*

Tujuan Penelitian

Tujuan penelitian ini adalah untuk melakukan prediksi kelanjutan studi siswa ke perguruan tinggi dengan menggunakan Naive Bayes, sehingga dapat mengetahui efektifitas dan efisiensi dari algoritma Naive Bayes.

Metode penelitian

Langkah-langkah yang diterapkan dalam penelitian ini sebagai berikut:

1. Analisa Masalah dan Studi Literatur

Langkah awal ini digunakan untuk penentuan rumusan masalah dalam penelitian. Pada langkah ini dilakukan pengamatan dari faktor-faktor yang mempengaruhi siswa melanjutkan studi ke perguruan tinggi. Sehingga selanjutnya dapat dianalisa untuk mengetahui cara penyelesaian dari masalah tersebut. Kemudian mempelajari dari berbagai literatur tentang penerapan metode Naive Bayes melalui jurnal-jurnal untuk meningkatkan dasar pengetahuan untuk melakukan penelitian selanjutnya.

2. Mengumpulkan Data

Data yang digunakan dalam penelitian ini adalah data *private* hasil rekap siswa SMA Negeri 1 Pecangaan dengan jumlah 313 record dengan 12 atribut.

3. Pengolahan Data Awal

Pada tahapan ini dataset yang mempunyai missing value harus diperlakukan secara khusus. Agar dikenali dan dapat digunakan sebagai model algoritma yang diusulkan. Data akan diproses dengan dibantu tools data mining. Jika masih terdapat data yang missing value maka akan dilakukan filter data. Dan data yang masih inkonsistensi dilakukan pembersihan data atau cleaning missing value.

4. Metode yang diusulkan

No	Langkah Metode
1.	Pengumpulan dataset
2.	Pengolahan data awal (preprocessing) yang bertujuan untuk meningkatkan efisiensi klasifikasi yaitu dengan penghapusan data yang missing value dengan menggunakan tool data mining.
3.	Pelatihan data dan pengetesan data menggunakan algoritma Naive Bayes.
4.	Nilai hasil dari pelatihan data dan pengetesan data menggunakan algoritma Naive Bayes

HASIL PENELITIAN dan pembahasan

Pada penelitian ini bertujuan mengetahui nilai akurasi algoritma Naïve Bayes Dalam penelitian ini digunakan dataset *private* rekap siswa SMA. Dimana dataset tersebut sebelum pengujian, terlebih dahulu diberikan tipe nilai pada tiap-tiap atribut datasetnya.

1. Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini adalah *dataset private* rekap siswa SMA. Dataset awal terdiri dari 313 record dan 12 variabel yaitu jenis kelamin, pendapatan orang tua, pekerjaan orang tua, PIP, pendidikan ayah, pendidikan ibu, nilai bahasa Indonesia, nilai bahasa Inggris, nilai Matematika, nilai jurusan, nilai ijazah dan lanjut.

2. Pengolahan Data Awal

Pada tahap ini dilakukan penghapusan data yang kosong atau missing value sebanyak 38 data, sehingga data yang terkumpul menjadi 275 record. Setelah itu ditentukan variabel-variabel yang akan digunakan sebagai variabel prediktor dan variabel target. Variabel jenis kelamin, pendapatan orang tua, pekerjaan orang tua, PIP, pendidikan ayah, pendidikan ibu, nilai bahasa Indonesia, nilai bahasa Inggris, nilai Matematika, nilai jurusan, nilai ijazah sebagai variabel prediktor, sedangkan variabel lanjut sebagai variabel target.

3. Eksperimen

3.1 Model Algoritma Naive Bayes

Pada penelitian dan pengujian metode ini akan dilakukan eksperimen algoritma Naive Bayes dengan percobaan 2-20 *number of validation*. Percobaan menggunakan 275 data dan 11 variabel prediktor yang sudah dinormalisasi, yaitu jenis kelamin, pendapatan orang tua, pekerjaan orang tua, PIP, pendidikan ayah, pendidikan ibu, nilai bahasa Indonesia, nilai bahasa Inggris, nilai Matematika, nilai jurusan, dan 1 variabel target lanjut. Sebelum dilakukan pengujian, terlebih dahulu dilihat probabilitas lanjut atau tidak lanjut yang dihasilkan dari tiap-tiap atribut yaitu jenis kelamin, pendapatan orang tua, pekerjaan orang tua, PIP, pendidikan ayah, pendidikan ibu, nilai bahasa Indonesia, nilai bahasa Inggris, nilai Matematika, nilai jurusan, nilai ijazah. Pada tabel 1.1 menunjukkan probabilitas lanjut atau tidak lanjut.

Tabel 1.1 Probabilitas masing-masing atribut

Variabel		Probabiltiy	
		Yes	No
Jenis Kelamin	Laki-laki	0,424	0,318
	Perempuan	0,576	0,682
Pendapatan Orang Tua	Rp. 1000.000-Rp.1.999.999	0,628	0,500
	Rp. 2000.000-Rp.4.999.999	0,247	0,273
	Rp.500.000-Rp.999.999	0,126	0,182
Pekerjaan Orang Tua	Wiraswasta	0,476	0,477
	PNS/TNI/POLRI	0,177	0,182
	Buruh	0,091	0,068
	Karyawan Swasta	0,082	0,114
	Pedagang Kecil	0,043	0,023
	Pedagang Besar	0,043	0,068
	Sudah Meninggal	0,022	0,068
	Petani	0,022	0,000
	Peternak	0,004	0,000
	Pensiunan	0,013	0,000
	Nelayan	0,022	0,000
	Wirausaha	0,004	0,000
PIP	Tidak Mendapat	0,779	0,682
	Mendapat	0,221	0,318
Pendidikan Ayah	SMA/Sederajat	0,369	0,432
	S1	0,121	0,136
	SD/Sederajat	0,242	0,227
	SMP/Sederajat	0,186	0,159
	D3	0,043	0,023
	D2	0,004	0,000
	D1	0,004	0,023
	S2	0,017	0,000
	Tidak sekolah	0,004	0,023
	Putus SD	0,017	0,000

		0,004 0,009	0,000 0,000
Pendidikan Ibu	SMA/Sederajat S1 SD/Sederajat SMP/Sederajat D3 D2 D1 S2 Tidak sekolah Putus SD	0,320 0,117 0,225 0,208 0,048 0,004 0,004 0,004 0,009	0,363 0,114 0,182 0,295 0,023 0,000 0,023 0,000 0,000 0,000
Nilai Bahasa Indonesia	Mean Standard deviation	75,066 11,547	76,163 11,972
Nilai Bahasa Inggris	Mean Standard deviation	54,642 15,298	57,745 14,581
Nilai Matematika	Mean Standard deviation	45,590 13,597	44,568 12,952
Nilai Jurusan	Mean Standard deviation	60,284 12,551	61,109 12,194
Nilai Ijazah	Mean Standard deviation	81,951 2,228	81,595 2,122

3.2 Pengujian Model Algoritma Naive Bayes

Dalam pengujian ini akan digunakan *K-fold cross validation* untuk validasi dan *confussion matrix* untuk mengetahui tingkat akurasi. Dalam pengujian *K-Fold Cross Validation* Algoritma Naïve Bayes, peneliti menggunakan 2-20 *number of validation* dengan *sampling type Stratified* (bertingkat-tingkat). Dan nilai hasil dari pengujian Naïve Bayes ini ditunjukkan pada tabel 1.2 hasil percobaan Naïve Bayes.

Tabel 1.2 Hasil Percobaan Naïve Bayes

K-fold	Akurasi
2	83.26%
3	85.09%
4	84.01%
5	84.73%
6	85.81%
7	85.47%
8	83.97%
9	86.18%
10	85.82%
11	84.36%
12	85.08%
13	84.40%
14	84.34%
15	86.24%
16	85.80%
17	85.83%
18	86.53%
19	85.11%
20	86.21%

Berdasarkan tabel 1.2 hasil percobaan pada alagoritma Naïve Bayes dengan menggunakan 2-20 *number of validation* yang mendapatkan nilai akurasi tertinggi adalah pada *number of validation* 18 dengan nilai akurasinya 86.53%.

Pada pengujian ini digunakan data sebanyak 275 record dengan 12 atribut di antaranya adalah jenis kelamin, pendapatan orang tua, pekerjaan orang tua, PIP, pendidikan ayah, pendidikan ibu, nilai bahasa Indonesia, nilai bahasa Inggris, nilai Matematika, nilai jurusan, nilai ijazah dan lanjut yang dijadikan sebagai label. Dengan menggunakan metode Naive Bayes dan pengujian model *k-fold cross validation* diperoleh hasil yang dapat dilihat pada Tabel 1.3 sebagai berikut :

Tabel 1.3 Confusion Matrix Naïve Bayes

Accuracy: 86.53%			
	True Lanjut Studi	True Tidak Lanjut Studi	Class Precision
Pred. Lanjut Studi	228	34	87.02 %
Pred. Tidak Lanjut Studi	3	10	76.92 %
Class Recall	98.70 %	22.73 %	

Berdasarkan tabel 1.3 dapat diketahui: *True Positive* (TP) 228, nilai *True Negative* (TN) 10, nilai *False Positive* (FP) 34, nilai *False Negative* (FN) 3. Sehingga jika dihitung untuk mencari nilai *accuracy, sensitivity, spesificity, ppv*, dan *npv* pada persamaan di bawah ini:

TP=228 TN=10
FP=34 FN=3

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{228 + 10}{228 + 10 + 34 + 3}$$

$$Sensitivity = \frac{tp}{tp + fn} = \frac{228}{228 + 3}$$

$$specitivity = \frac{tn}{tn + fp} = \frac{10}{10 + 34}$$

$$PPV = \frac{tp}{tp + fp} = \frac{228}{228 + 34}$$

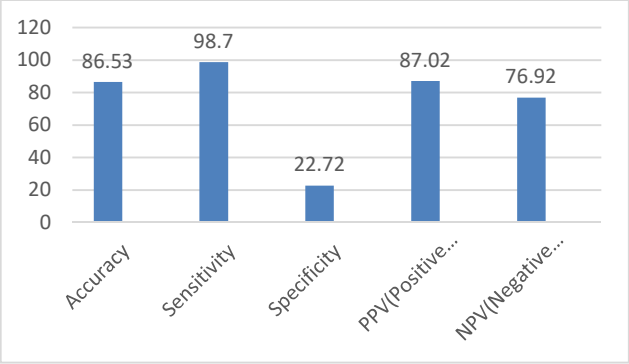
$$NPV = \frac{tn}{tn + fn} = \frac{10}{10 + 3}$$

Hasil perhitungan persamaan di atas dapat dilihat pada Tabel 1.4

Tabel 1.4 Nilai *accuracy, sensitivity, specificity, ppv*, dan *npv*

Accuracy	86.53%
Sensitivity	98.70%
Specificity	22.72%
PPV(Positive Predictive Value)	87.02%
NPV(Negative Predictive Value)	76.92%

Hasil perhitungan persamaan nilai *accuracy, sensitivity, specificity, ppv*, dan *npv* algoritma Naïve Bayes dapat dilihat gambar grafiknya pada gambar 1.1



Gambar 1.1 Hasil perhitungan persamaan nilai *accuracy, sensitivity, specificity, ppv*, dan *npv* algoritma Naïve Bayes

Dari perhitungan di atas maka diketahui tingkat akurasi hasil prediksi kelanjutan studi siswa ke perguruan tinggi menggunakan algoritma Naïve Bayes sebesar 86.53%.

PENUTUP

KESIMPULAN

Berdasarkan percobaan dalam penelitian yang sudah dilakukan, dapat ditarik kesimpulan bahwa :

1. Metode Naive Bayes dapat menghasilkan probabilitas di setiap kriteria dengan class yang berbeda, sehingga nilai-nilai probabilitasnya dapat digunakan untuk mengoptimalkan prediksi kelanjutan studi siswa ke perguruan tinggi mengacu pada pengklasifikasian yang dilakukan Naive Bayes sendiri.
2. Pemilihan pembobotan pada atribut dapat meningkatkan nilai akurasi pengklasifikasian dari data yang diujikan sehingga menghasilkan nilai prediksi 86.53%.

Saran

Saran yang dapat digunakan untuk penelitian berikutnya untuk mencapai hasil yang lebih baik adalah diharapkan pada penelitian berikutnya dapat dikembangkan dengan menggunakan metode optimasi lainnya, seperti Particle Swarm Optimization, Genetic Algorithm, Ant Colony, Bee Colony dan lain-lain.

DAFTAR PUSTAKA

[1] D. Hand, D. Hand, H. Mannila, H. Mannila, P. Smyth, and P. Smyth, *Principles of data mining*, vol. 30. 2001.

[2] C. Vercellis, “Business Intelligence: Data Mining and Optimization for Decision Making,” p. 436, 2009.

[3] M. Bramer, *Principles of Data Mining*. 2013.

[4] Florin Gorunescu, *Intelegent System Reference Library*. .

[5] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier under Zero-One Los,” *Mach. Learn.*, vol. 29, no. 1, pp. 103–130, 1997.

[6] A. J. B. Bumiputera and B. Branch, “Prediksi Hasil PEMILU Legislatif Kabupaten Lombok Timur

Menggunakan Algoritma Naive Bayes Berbasis PSO," vol. 5, no. 2, pp. 1359–1368, 2017.

[7] M. Gogoi and S. K. Sarma, "Document Classification of Assamese Text Using Naïve Bayes Approach," *Int. J. Comput. Trends Technol.*, vol. 30, no. 4, pp. 182–186, 2015.

[8] T. Sun, "Spam Filtering based on Naive Bayes Classification," 2009.

[9] I. Ahmed, D. Guan, and T. C. Chung, "SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 2, pp. 183–187, 2014.

[10] Aden Ginanjar Andanawari, "Faktor-faktor yang mempengaruhi minat siswa melanjutkan pendidikan ke perguruan tinggi dengan menggunakan regresi logistik," 2010.