PREDIKSI HERREGISTRASI CALON MAHASISWA BARU MENGGUNAKAN ALGORITMA NAÏVE BAYES

Selvy Megira[1)], Kusrini[2)], Emha Taufiq Luthfi[3)]

[1), 2), 3)] Teknik Informatika Universitas AMIKOM Yogyakarta
Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281
Email : selvy.megira@students.amikom.ac.id[1)], kusrini@amikom.ac.id[2)]

Abstrak

Data mining merupakan suatu cara dalam menguraikan penemuan pengetahuan, selain untuk memudahkan dalam melakukan pengambilan keputusan saat penerimaan calon mahasiswa baru dengan melakukan prediksi herregistrasi. Salah satu algoritma yang dapat digunakan untuk melakukan prediksi herregistrasi yaitu naive bayes, naive bayes merupakan teknik prediksi yang berbasis probabilistik sederhana dengan penerapan teorema Bayes. Peneliti melakukan penelitian dengan membuat suatu sistem aplikasi untuk melakukan prediksi herregistrasi bagi calon mahasiswa baru dengan menggunakan algoritma naive bayes. Hasil dari penelitian menunjukkan bahwa aplikasi dapat digunakan untuk melakukan prediksi herregistrasi calon mahasiswa baru.

Kata kunci: *Herregistrasi, Data Mining, Naive Bayes*

1. Pendahuluan

Universitas AMIKOM Yogyakarta sebagai salah satu perguruan tinggi swasta di Provinsi Daerah Istimewa Yogyakarta menjadi bagian dari persaingan untuk mendapatkan calon mahasiswa baru setiap tahunnya, setiap penerimaan calon mahasiswa baru sering terjadi ada calon mahasiswa yang tidak melakukan herherregistrasi. Jika kemungkinan calon mahasiswa baru yang tidak melakukan herherregistrasi dapat di ketahui lebih awal maka pihak pengelola dapat melakukan tindakan-tindakan untuk mempertahankan calon mahasiswa baru yang mungkin potensial.

Salah satu cara untuk mengetahui calon mahasiswa baru yang tidak herregistrasi dengan melakukan prediksi apakah seorang calon mahasiswa baru akan cenderung herherregistrasi atau tidak menggunakan *data mining* dengan membandingkan metode klasifikasi. Klasifikasi merupakan suatu pekerjaan yang melakukan pelatihan atau memasukannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia yang menghasilkan suatu model kemudian disimpan sebagai memori [1].

Metode klasifikasi yang umum digunakan antara lain *decision tree, k-nearest* neighbor, *naïve bayes, neural network dan support vector machines* [2]. Beberapa penelitian sebelumnya mengenai klasifikasi data menggunakan metode *naïve bayes* yang berjudul "**Drop Out Estimation Students Based On The Study Period: Comparison Between Naïve Bayes And Support Vector Machines Algorithm Methods**" yang bertujuan untuk melakukan estimasi mahasiswa yang drop out karena tidak bisa menyelesaikan studi tepat waktu. Hasil penelitian ini menunjukkan bahwa *naïve bayes* memiliki akurasi terbaik dalam prediksi mahasiswa yang melakukan pengunduran diri dengan persentase 80,67% sedangkan *error* sebanyak 19,33% [3]. Penelitian selanjutnya yang berjudul "**Perbandingan Kinerja Algoritma C4.5 dan Naïve Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa**" yang bertujuan dalam melakukan klasifikasi data mahasiswa untuk penentuan pemilihan konsentrasi. Hasil penelitian menunjukkan bahwa tingkat akurasi *naïve bayes* sebesar 78,47% dalam ketepatan pemilihan konsentrasi mahasiswa [4].

Penelitian berikutnya yang berjudul "**Implementasi Data Mining dengan Naïve Bayes Classifier untuk Mendukung Strategi Pemasaran di Bagian HUMAS STMIK AMIKOM YOGYAKARTA**" yang bertujuan melakukan prediksi minat studi calon mahasiswa dengan aplikasi *Smart Marketing*. Hasil penelitian dari uji coba 1000 *data testing* yang diambil secara acak menggunakan aplikasi *Smart Marketing* untuk prediksi minat studi didapatkan akurasi sebesar 92,7% dan error sebesar 7,3% [5]. Berdasarkan paparan diatas, peneliti bermaksud untuk melakukan prediksi herregistrasi calon mahasiswa baru menggunakan algoritma *naïve bayes*.

Data Mining

Data mining dapat disebut menguraikan penemuan pengetahuan di dalam database [6]. Selain itu data mining digunakan untuk mengekstraksi dan mengidentifikasi informasi yang penting dalam *database* besar dengan cara proses semi otomatik menggunakan teknik statistik, mate-matika, kecerdasan buatan dan *machine learning*[7]. Data mining juga biasa dikenal dengan *Knowledge Discovery (mining) in Database* (KDD) yang terdiri dari beberapa tahapan seperti pemilihan data, pra pengolahan, transformasi, data mining dan evaluasi hasil. dan bertujuan untuk menghasilkan informasi baru yang berguna dengan cara memanfaatkan dan mengolah data dalam suatu database. Dalam memudahkan aktifitas *recording* atau mengelola data yang besar dalam suatu transaksi untuk menghasilkan informasi yang berguna dan akurat bagi penggunanya maka dibutuhkan data mining. Secara

sistematis ada tiga langkah utama dalam data mining [2].

a. Eksplorasi (Pemrosesan awal data)
   Eksplorasi terdiri dari pembersihan data, normalisasi data, transformasi data yang salah, reduksi dimensi dan sebagainya.
b. Membangun model dan melakukan validasi terhadapnya
   Melakukan analisis terhadap berbagai model kemudian memilih model yang mempunyai kinerja prediksi yang terbaik. Pada langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis cluster, deteksi anomali, analisis asosiasi, analisis pola sekuensial dan sebagainya.
c. Penerapan
   Penerapan yang berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan atau prediksi masalah yang di investigasi.

Klasifikasi

Klasifikasi merupakan suatu pekerjaan yang melakukan pelatihan atau memasukannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia yang menghasilkan suatu model kemudian disimpan sebagai memori[1]. Dalam klasifikasi terdapat target variabel kategori sebagai contoh penggolongan jurusan SMA dipisahkan menjadi dua kategori, yaitu IPA dan IPS . Klasifikasi data terdiri dari 2 langkah proses, pertama adalah learning (*fase training)* yang digunakan untuk menganalisa data training kemudian merepresentasikan dalam bentuk *rule* klasifikasi. Proses kedua kelasifikasi data tes yang digunakan untuk memperkirakan akurasi dari *rule* klasifikasi. Berdasarkan cara pelatihan algoritma klasifikasi dibagi menjadi dua macam yaitu *eager learner* dan *lazy learner*. Algoritma *eager learner* merupakan proses yang dilakukan menggunakan model yang tersimpan sehingga tidak melibatkan data latih sama sekali kelebihan pada algoritma ini proses yang dilakukan dapat berjalan dengan cepat namun proses pelatihan lama. [8].

Naïve Bayes

Naïve bayes adalah teknik prediksi yang berbasis probabilistik sederhana dengan penerapan teorema Bayes (aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat [1]. Adapun performa naïve bayes yang kompetitif dalam proses klasifikasi walaupun menggunakan asumsi keindependenan atribut (tidak ada kaitannya antar atribut). Walaupun asumsi keindependenan atribut ini pada data jarang terjadi akan tetapi asumsi keindependenan atribut tersebut dilanggar performa pengklasifikasian Naïve Bayes cukup tinggi, hal ini sudah dibuktikan pada berbagai penelitian empiris Selain itu *naïve bayes* dapat dikatakan sebagai asumsi penyederhanaan bahwa nilai atribut secara konditonal saling bebas jika diberikan nilai output. Klasifikasi *naïve bayes* dapat dibuat lebih efisien sebagai bentuk pembelajaran. Adapun parameter yang dapat digunakan untuk perhitungan *naïve bayes* dapat menggunakan metode *maximum likehood* atau kemiripan tertinggi [1]. Untuk menghitung prediksi dengan *naïve bayes* digunakan Persamaan 1:

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

Keterangan:

X : data dengan *class* yang belum diketahui
H : hipotesis data X merupakan suatu *class* spesifik
P(H|X) : probabilitas hipotesis H berdasar kondisi (*posteriori probability*)
P(H) : probabilitas hipotesis H (*prior probability*)
P(X|H) : probabilitas X berdasar kondisi pada hipotesis H
P(X) : probabilitas dari X

Adapun kelebihan Naive Bayes adalah sebagai berikut:

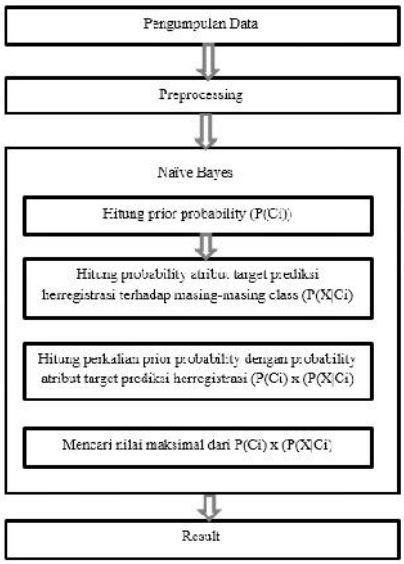
a. Menangani kuantitatif dan data diskrit
b. Kokoh untuk titik noise yang diisolasi, misalkan titik yang dirata – ratakan ketika mengestimasi peluang bersyarat data.
c. Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata – rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi.
d. Menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang
e. Cepat dan efisiensi ruang

Sedangkan kekurangan Naive Bayes sebagai berikut :

a. Tidak berlaku jika probabilitas kondisionalnya adalah nol, apabila nol maka probabilitas prediksi akan bernilai nol juga
b. Mengasumsikan variabel bebas

2. Pembahasan

Diagram Alir Sistem



Gambar 1. *Diagram alir sistem*

Diagram alir data menggambarkan proses kerja perhitungan model algoritma naïve bayes pada sistem seperti pada Gambar 1

Pengumpulan data

Dalam pembahasan ini terdapat 2 dataset yang digunakan yaitu Pertama : Data Penerimaan Mahasiswa Baru yang memiliki atribut dari 6 atribut, dimana 5 atribut predictor dan 1 atribut hasil. Atribut-atribut yang menjadi parameter terlihat pada Tabel 1 yaitu :

Tabel 1. *Atribut dan Nilai Kategori*

Atribut	Nilai	Tipe
Pendapatan ayah	1=0-1000000, 2=1000001-3600000, 3=3600001-7600000, 4=7600001-15000000, 5=15000001-25000000, 6=25000001- 1000000000	Diskrit
Nilai UN	7= 5-6 8= 7-8 9= 9-10	Diskrit
Minat Studi	S1 Teknik Informatika, S1 Sistem Informasi, D3 Teknik Informatika, D3 Manajemen Informatika	Diskrit
Gelombang Pendaftaran	Khusus, I, II, III, III-P	Diskrit
Jurusan	IPA, IPS, TKJ, Akuntansi	Diskrit
Status	True, False	

Preprocessing

Proses pembersihan data mencakup antara lain memeriksa data yang tidak konsisten, data dengan *missing value* dan *redundant* data. Seluruh atribut pada dua kelompok data (tabel) dibersihkan karena hal tersebut merupakan syarat awal untuk proses data mining yang akan menghasilkan dataset yang bersih dan siap digunakan pada tahap mining data. Dikatakan *missing value* jika pada salah satu atribut nilai *record* tersebut hilang maka *record* yang dimaksud akan dihapus, karena *record* tersebut dinilai kehilangan data atau *missing value*. Apabila dalam dataset yang sama terdapat lebih dari satu *record* yang berisi nilai yang sama, maka *record* yang dimaksud juga harus dihapus karena tidak akan memberi informasi yang berarti jika dipertahankan. Tahap ini tidak hanya membersihkan data yang mengandung *missing value* saja akan tetapi terhadap data yang tidak konsisten juga dilakukan. Data pada penelitian ini merupakan data yang sudah konsisten. Karena dua kelompok data (tabel) diambil seluruhnya tidak ada data yang *dicleaning*, maka jumlah

atribut dan record pada kelompok data (tabel) adalah tetap. Pada tahap ini data sudah bersih dan siap untuk digunakan pada tahap selanjutnya yaitu implementasi perhitungan dengan algoritma naïve bayes.

Implementasi Perhitungan Algoritma *Naive Bayes*

Setelah data hasil cleaning rampung maka selanjutnya, data tersebut akan di implementasikan dengan persamaan 1 *naïve bayes*. Adapun cara kerja dari proses perhitungan *naïve bayes* yaitu sebagai berikut: Tahapan diawali dengan mengambil data sample atau contoh data seperti pada Tabel 1 data tes mahasiswa.

Tabel 2. *Sample Data Testing*

No	Pendapatan Ayah	Nilai rata-rata UN	Minat Studi	Jurusan Sekolah	Gelombang Pendaftaran
1	2	2	S1 Teknik Informatika	TKJ	I
2	1	3	S1 Sistem Informasi	IPA	III
3	3	2	S1 Sistem Informasi	Akuntansi	II
4	4	1	S1 Sistem Informasi	IPS	I
5	2	2	D3 Manajemen Informatika	IPS	III-P
6	1	2	D3 Manajemen Informatika	IPS	III
7	1	2	S1 Sistem Informasi	TKJ	II
8	3	3	S1 Teknik Informatika	IPA	III-P
9	2	2	D3 Manajemen Informatika	Akuntansi	I
10	3	2	S1 Sistem Informasi	TKJ	II

1. Menghitung prior probabilitas untuk melakukan prediksi herregistrasi maka akan menghitung prior probabilitas terlebih dulu

   P(Registrasi) = $\frac{176}{196}$ = 0,90

   P(Tidak Registrasi) = $\frac{20}{196}$ = 0,102

2. Hitung probability atribut target prediksi herregistrasi terhadap masing-masing *class* (P(X|Ci). Sampel yang digunakan yaitu no 1 dengan atribut pendapatan ayah 2, nilai rerata UN 2, minat studi S1 teknik informatika, jurusan sekolah tkj, dan gelombang pendaftaran I. Untuk nilai probabilitasnya dapat dilihat pada Tabel 3.

Tabel 3. *Probability atribut target prediksi*

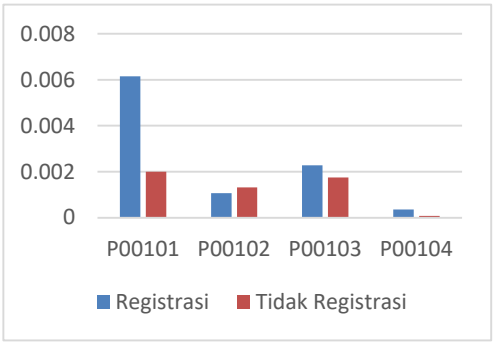
Atribut	Nilai	Probabilitas	
		Registrasi	Tidak Registrasi
Pendapatan ayah	2	0,2444	0,35
Nilai rerata UN	2	0,625	0,65

Minat Studi	S1 Teknik Informatika	0,341	0,35
Jurusan sekolah	TKJ	0,426	0,25
Gelombang pendaftaran	I	0,278	0,1

3. Tahapan berikutnya yaitu hitung perkalian prior probability dengan probability atribut target prediksi herregistrasi(P(Ci) x (P(X|Ci). Adapun hasil perkalian probabilitas terhadap sampel data uji dapat dilihat pada Tabel 4, sedangkan untuk grafik hasil perkalian probabilitas dapat dilihat pada Gambar 2

Tabel 4. *Hasil perkalian probabilitas*

No	Id Pendaftar	Probabilitas	
		Registrasi	Tidak Registrasi
1	P00101	0,00616	0,00199
2	P00102	0,00106	0,00132
3	P00103	0,00227	0,00175
4	P00104	0,00035	0,00008
5	P00105	0,00042	0,00153
6	P00106	0,00064	0,00156
7	P00107	0,00884	0,00097
8	P00108	0,00087	0,00334
9	P00109	0,00075	0,00068
10	P00110	0,00583	0,00219



Gambar 2. Grafik hasil perkalian probabilitas

Pada Gambar 2. Dapat dilihat bahwa sampel P00104 memiliki probabilitas tidak registrasi paling rendah untuk P00101 memiliki probabilitas registrasi paling tinggi dan pada P00102 memiliki probabilitas tidak registrasi dibandingkan dengan prbabilitas registrasi.

4. Selanjutnya mencari nilai maksimal dari P(Ci) x (P(X|Ci) untuk menentukan hasil prediksi herregistrasi. berikut Tabel 5 hasil dari prediksi herregistrasi

Tabel 5. Hasil prediksi Herregistrasi

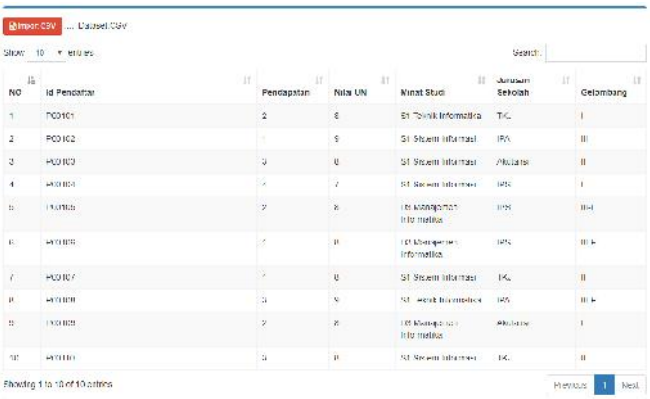
No	Id pendaftar	Status Prediksi
1	P00101	Registrasi
2	P00102	Tidak Registrasi
3	P00103	Registrasi
4	P00104	Registrasi
5	P00105	Tidak Registrasi
6	P00106	Tidak Registrasi
7	P00107	Registrasi
8	P00108	Tidak Registrasi
9	P00109	Registrasi
10	P00110	Registrasi

Pada Tabel 5 dapat dilihat bahwa id pendaftar P00102, P00105, P00106, dan P00108 menghasilkan status prediksi tidak registrasi sedangkan id pendaftar P00101, P001010, P00103, P00104, P00107, P00109 di prediksi akan melakukan registrasi.

Antarmuka Sistem

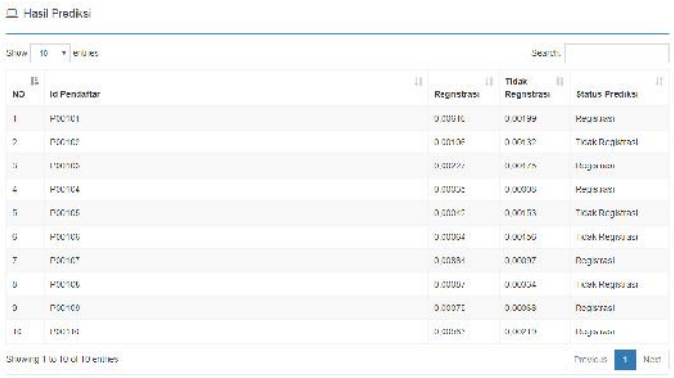
Tampilan antarmuka pada sistem prediksi herregistrasi dengan menggunakan algoritma *naïve bayes* dapat dilihat pada Gambar 3 dan Gambar 4. Sebagai contoh yang akan ditampilkan pada antarmuka sistem, akan menggunakan data pada pembahasan sebelumnya yaitu :

Pada Gambar 3 merupakan hasil tampilan data uji dengan id prediksi P00101, P00102, P00103, P00104, P00105, P00106, P00107, P00108, P00109, P00110. Pada Gambar 4 merupakan hasil tampilan sistem prediksi herregistrasi ini berupa tampilan hasil prediksi herregistrasi calon mahasiswa baru.



Gambar 3. *Antarmuka Data Uji*

Gambar 3 merupakan antarmuka untuk admin dalam mengelola atau memasukan data uji dalam format CSV (*Comma Separated Values*).



Gambar 4 *Antarmuka Hasil Prediksi Herregistrasi*

Gambar 4 merupakan antarmuka untuk melakukan hasil prediksi herregistrasi calon mahasiswa baru.

3. Kesimpulan

Atribut yang digunakan untuk memprediksi herregistrasi calon mahasiswa baru adalah Pendapatan ayah, Nilai rata-rata UN, Minat studi, Jurusan sekolah dan Gelombang pendaftaran. Penerapan algoritma *data mining* menggunakan *naïve bayes* dapat dilakukan untuk memprediksi Herregistrasi calon mahasiswa baru.

Daftar Pustaka

[1] Prasetyo, E., *Data Mining Konsep dan Aplikasi Menggunakan Matlab*, Andi Offset, Yogyakarta, 2012.
[2] Prasetyo, E., *Data Mining: Mengolah Data menjadi Informasi Menggunakan Matlab*, Andi Offset, Yogyakarta, 2014.
[3] Harwati, et al., *Drop out Estimation Students based on the Study Period: Comparison between Naïve Bayes and Support Vector Machines Algorithm Methods*, IOP Conf. Series: Materials Science and Engineering, 2016.
[4] Supriyanti, W et al., *Perbandingan Kinerja Algoritma C4.5 dan Naïve Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa*, Jurnal INFORMA Politeknik Indonusa Surakarta, ISSN: 2442-7942, Vol.1 No 3, 2016.
[5] Hadi, Erik. S., & Burhan, A. M, 2014, *Implementasi Data Mining dengan Naïve Bayes Classifier untuk Mendukung Strategi Pemasaran di Bagian Humas STMIK AMIKOM Yogyakarta*, Semnasteknomedia, 2014.
[6] Kusrini, & Emha T. Luthfi., 2009, Algoritma Data Mining, Andi Offset, Yogyakarta
[7] Turban, E; Aronston; Liang p T, *Sistem Pendukung Keputusan dan Sistem Cerdas*, Jilid I. Andi Offset, Yogyakarta, 2005
[8] Larose D.T, *Discovering Knowledge in Data*. New Jersey : John Willey & Sons, Inc, 2005

Biodata Penulis

Selvy Megira, Saat ini menjadi Mahasiswa di Universitas AMIKOM Yogyakarta.

Kusrini, Saat ini menjadi Dosen di Universitas AMIKOM Yogyakarta.

Emha Taufiq Luthfi, Saat ini menjadi Dosen di Universitas AMIKOM Yogyakarta.

ISSN : 2302-3805