PENGKLASIFIKASIAN KEMAMPUAN AKADEMIK MAHASISWA MENGGUNAKAN METODE *INFORMATION GAIN* DAN *NAIVE BAYES CLASSIFIER* DALAM PREDIKSI PENYELESAIAN STUDI TEPAT WAKTU

Rd Muhammad Alfajri*, Yulison Herry Chrisnanto, Rezki Yuniarti
Jurusan Informatika, Fakultas MIPA, Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman PO BOX 148 Cimahi
*Email: alfajri607@gmail.com

Abstrak

Data mahasiswa merupakan data yang berlimpah terdiri dari awal masuk sampai dengan kelulusan . dikarenakan data yang sangat banyak faktor data warehouse menjadi peluang besar untuk mencul, data yang masuk secara beruntun pada setiap tahunnya memiliki pola tertentu sehingga diperlukan suatu teknik analisis data yang dapat mengambil informasi yang berharga dari sekian banyak data yang terkumpul pada suatu perangkat komputer atau pelaporan. penelitian ini dilakukan pada Universitas Jenderal Achmad Yani tepatnya pada jurusan informatika diperlukan suatu teknik analisis data untuk menganalisis data tersebut. Pada penelitian kali ini data latih sebanyak 82 dan data uji sebanyak 35 diperlukan beberapa atribut dalam menyelesaikan persolanan tersebut diantaranya NIM, nama mahasiswa matakuliah semester 1 sampai dengan 6 dengan total jumlah matakuliah sebanyak 65 dengan penyelesaian studi tepat waktu <= 4 tahun dan > 4 tahun metode yang digunakan adalah Naïve Bayes classifier penelitian ini memperoleh nilai akurasi sebelum menggunakan seleksi fitur sebesar 62 % dan setelah menggunakan seleksi fitur meningkat tidak signifikan sebesar 6 % atau menjadi 68 % sehingga seleksi fitur menggunakan information gain dapat menghilangkan noise dalam proses pengklasifikasian menggunakan naive bayes classifier.

***Kata kunci**: KHS, mata kuliah semester 1-6, naïve bayes classifier*

1. PENDAHULUAN

Mahasiswa merupakan peserta didik yang sedang dalam proses pembelajaran pada tingkat perguruan tinggi, institut maupun akademi. Mahasiswa tingkat akhir telah melakukan proses pembelajaran yang sangat lama dalam kurun waktu 4 tahun lamanya. Sehingga untuk memprediksi mahasiswa sangatlah mungkin terjadi untuk penelitian ini salah satu penyebab yang digunakan berupa nilai hasil akhir matakuliah karena proses rekapitulasi nilai akademis adalah rutin dilakukan dan selain itu hampir setiap mahasiswa yang lulus tepat waktu atau terlambat tercerminkan melalui nilai matakuliah yang diperoleh sehingga tidak salahnya untuk menentukan kelulusan berdasarkan suatu kemampuan akademis untuk data yang digunakan berupa nilai hasil akhir matakuliah yang berada pada kartu hasil studi atau dikenal dengan KHS untuk studi kasus yang digunakan di jurusan informatika universitas jenderal achmad yani.Untuk mengetahui mahasiswa yang lulus tepat waktu atau terlambat digunakan suatu teknik metode naïve bayes classifier Naïve bayes classifier yaitu suatu teknik pengklasifikasian dengan probabilitas sederhana yang mengaplikasikan teorema bayes dalam teori probabilitas dan statistika, teorema bayes adalah sebuah penafsiran berbeda dalam penafsiran teorema ini menyatakan seberapa jauh derajat kepercayaan subjektif harus berubah secara rasional ketika ada petunjuk baru.

Beberapa penelitian telah dilakukan menggunakan distribusi normal sebagai perhitungan probabilitas dalam mengklasifikasikan sebuah kasus baru seperti klasifikasi status gizi (Arifin, 2015), kemudian terdapat penelitian yang memiliki kasus sama dan tidak menggunakan distribusi normal seperti Evaluasi kinerja akademik mahasiswa(Yunus, 2015), Sistem pendukung keputusan untuk menentukan kelulusan(Kusumadewi, 2009), tidak menggunakan metode naive bayes seperti desain model kelulusan dengan decision tree (Laily dkk, 2014), kemudian seleksi fitur menggunakan information gain dalam klasifikasi dalam prediksi komunikasi (Mujib dkk, 2013), klasifikasi berita hoax (Purwarianti, 2015), klasifikasi data momentum (Wasiati, 2015), dan penelitian yang memiliki kasus berbeda tetapi menggunakan distribusi normal seperti kelayakan calon tenaga kerja (Maharani, 2009).

ISBN 978-602-99334-5-1

2. METODOLOGI PENELITIAN

2.1. Data Cleaning

Data cleaning adalah proses pengecekan data untuk konsistensi data meliputi pemeriksaan data yang *out of range* yaitu tidak konsisten secara logika terdapat nilai ekstrim dan nilai – nilai yang tidak terdefinisi.

2.2. Transformasi Data

Tujuan utama dari transformasi data ini adalah untuk mengubah skala pengukuran data asli menjadi bentuk lain sehingga data dapat memenuhi asumsi-asumsi yang mendasari analisis.

2.3. Seleksi Fitur

Seleksi fitur yang digunakan dalam penelitian ini menggunakan information gain dengan mengambil nilai info gain berdasarkan nilai threshold yang telah ditentukan.

2.4. Data mining

Dengan menggunakan teknik naïve bayes classifier yang diaplikasikan pada data untuk mencari pola atau informasi dari data informasi yang ingin dicari yaitu kelas tepat waktu dan terlambat.

2.5. Evaluasi Pola

Tahapan ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir data mining adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami pengguna.

3. HASIL DAN PEMBAHASAN

3.1 Klasifikasi Naive Bayes Classifier

Ketika atribut X_i bertupe kuantitatif maka peluang P($X_i|Y$) akan sangat kecil sehingga membuat persamaan tersebut tidak dapat diandalkan untuk permasalahan atribut bertipe kuantitatif. Maka untuk menangani atribut kuantitatif, ada beberapa cara pendekatan yang dapat digunakan seperti distribusi normal (Gaussian) merupakan salah satu distribusi probabilitas yang paling banyak digunakan dalam analisis statistika.dengan formulasi berikut.

$$P(Xi = xi \mid Y = yj) = \frac{1}{\sqrt{2\pi}\,\sigma_{ij}} exp^{-\frac{(Xi-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(1)}$$

Berikut adalah sampel data masukan yang berjumlah 82 mahasiswa sebagai data latih dengan TW adalah tepat waktu dan TR adalah terlambat berikut pada Tabel 3.1

Tabel 1. Data Masukan

No	NIM	Alpro	P.Alpro	Matdas	⋯	S.Distribusi	Kelas
1	3411061003	4	4	2	⋯	3	TR
2	3411071010	2	2	3	⋯	3	TW
3	3411071019	3	2	2	⋯	4	TR
4	3411081001	3	3	3	⋯	4	TW
5	3411081018	2	2	2	⋯	4	TW
6	3411081032	2	2	2	⋯	4	TW
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
81	3411071041	3	2	2	⋯	3	TW
82	3411107047	2	2	2	⋯	4	TR

Berikut adalah pengujian data yang dilakukan terhadap data kasus lama sebagai pengetahuan dari penelitian ini berkit adalah sebagian data uji dari 35 mahasiswa dalam penelitian ini pada Tabel 2.

Tabel 2. Data Uji

NIM	Alpro	P.Alpro	matda	⋯	Teori_game	Sist_distribusi	Kelas
34111????	4	?	?	. . .	?	?	TW
34111????	4	?	?	. . .	?	?	TR

Sebelum menentukan probabilitas setiap kelas terhadap masing-masing mahasiswa menggunakan distribusi normal terlebih dahulu dilakukan pemisahan antara likelihood tepat waktu dan likelihood terlambat dengan persamaan 1 untuk Likelihood tepat waktu P (MK001) = 4 | Kelas = Tepat waktu.

$$= \frac{1}{\sqrt{2\ x\ 3.14}\ x\ 0.6893} e^{-\frac{(4-2.8864)^2}{2\ x\ (0.6893)^2}}$$

$$= \frac{1}{1.7273} 2.718282^{-\frac{1.2401}{0.9502}}$$

$$= 0.579\ \text{x}\ 2.718282^{-1.3050}$$

= 0.579 x 0.2711

= 0.1570

Likelihood terlambat P (MK002) = 4 | Kelas = Terlambat

$$= \frac{1}{\sqrt{2\ x\ 3.14}\ x\ 0.6017} e^{-\frac{(4-2.4474)^2}{2\ x\ (0.6017)^2}}$$

$$= \frac{1}{1.5078} 2.718282^{-\frac{2.4105}{0.7240}}$$

$$= 0.6632\ \text{x}\ 2.718282^{-3.3294}$$

= 0.6632 x 0.0358

= 0.0237

Berikut adalah contoh hasil normalisasi setelah melalui proses menghitung menggunakan distribusi normal di setiap matakuliah pada data uji.

Likelihood Tepat Waktu :

= 0.1570 x 0.220 x 0.417 x ….. x 0.379 = 2.31213 x 10^{-29}

Likelihood Terlambat :

= 0.035 x 0.099 x 0.118 x …… x 0.247 = 1.04468 x 10-^{51}

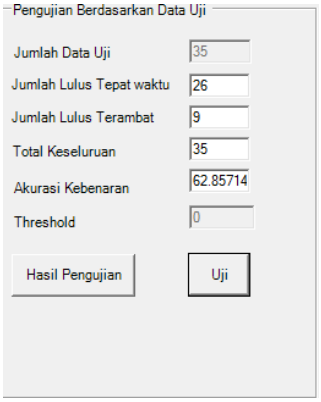
Untuk menghasilkan nilai Probabilitas maka dilakukan normalisasi terhadap *likelihood* tepat waktu dan terlambat.

$$= \frac{2.31213\ x\ 10^{-29}}{2.31213\ x\ 10^{-29} + 1.04468\ x\ 10^{-51}} = 1$$

$$= \frac{1.04468\ x\ 10^{-51}}{2.31213\ x\ 10^{-29} + 1.04468\ x\ 10^{-51}} = 4.51828\ x\ 10^{-23}$$

Diketahui likelihood tepat waktu dengan nilai probabilitas 1 yaitu bilai dipersentasikan adalah 100% sehingga mahasiswa tersebut akan diprediksikan lulus tepat waktu. Disebabkan memiliki data yang sangat banyak terdiri dari 65 kolom dan 82 baris sehingga perlu adanya bantuan sistem diperoleh nilai akurasi keseluruhan sebelum menggunakan seleksi fitur dengan pemahaman konsep pada sebelumnya adalah 62 % berikut pada Gambar 1.

ISBN 978-602-99334-5-1



Gambar 1. Akurasi Sistem Setelah Seleksi Fitur

3.2 Information Gain (Seleksi Fitur)

Entrophy merupakan ukuran ketidak pastian dimana semakin tinggi entrophy maka semakin tinggi ketidakpastian. Rumus dari entrophy (Slocum,2012) :

$E(S) = -\sum_{j=1}^{n} f_s(j) log_2 f_s(j)$.....(2)

Dimana:
E(S) adalah total informasi entrophy untuk atrubut S
N adalah junlah nilai pada atribut S
Fs(j) adalah frekuensi dari nilai S terhdap J atau kelas
Log2 adalah logaritma biner

Information gain dari output data atau variable dependent Y (kelas) yang dikelompokan berdasarkan atribut A, dinotasikan dengan gain(y,A). Information gain (y,A) dari atribut A relative terhadap output data y adalah (Santosa, 2007) :

$gain(y, A) = Entrophy(y) - \sum_{c \in nilai(A)} \frac{y_c}{y} entrophy(y_c)$.....(3)

Dimana nilai(A) adalah semua nilai yang mungkin dari atribut A, dan *yc* adalah subset dari *y* dimana A mempunyai nilai c. Term yang pertama pada rumus information gain di atas adalah entrophy total y dan term kedua adalah entrophy sesudah dilakukan pemisahan data berdasarkan atribut A. Pada penelitian ini untuk kelas tepat waktu dan 38 untuk kelas terlambat sehingga hasil perhitungan untuk 2 kelas tepat waktu dan terlambat adalah sebagai berikut menggunakan persamaan 3 :

E = ((-44/82)log2(44/82)) + ((-38/82)log2(38/82))=0.996.....(4)

Jika pada algoritma dan pemrograman terdapat nilai 2 maka hitung jumlah 2 pada algoritma & pemrograman untuk kelas tepat waktu dan untuk kelas terlambat maka menghasilkan nilai sejumlah 13 untuk kelas tepat waktu dan 27 untuk kelas terlambat berikut adalah peroleh entrophy setiap pada setiap nilainya yaitu menggunakan persamaan 4 :

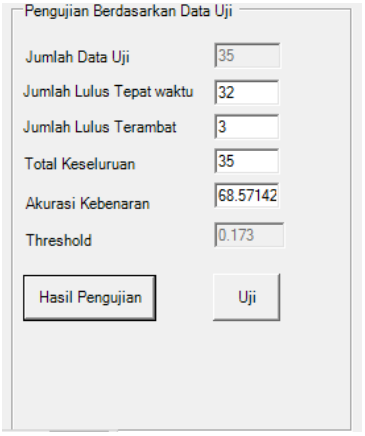
Entrophy (2) = *Entrophy* [13,23] = $-\frac{13}{40} log_2 \frac{13}{40} log_2 - \frac{23}{40} log_2 \frac{23}{40} = 0.944$

Entrophy (3) = *Entrophy* [23,13] = $-\frac{23}{36} log_2 \frac{23}{36} log_2 - \frac{13}{36} log_2 \frac{13}{36} = 0.994$

Entrophy (4) = *Entrophy* [8,2] = $-\frac{8}{12} log_2 \frac{8}{12} log_2 - \frac{2}{12} log_2 \frac{2}{12} = 0.72$

Diketahui nilai akurasi sebelum menggunakan seleksi fitur adalah 62 % kemudian setelah mengetahui setiap nilai information gain pada setiap matakuliah pada kasus lama kemudian dihilangkan satu persatu dengan menghilangkan nilai information gain yang paling kecil yang

dijadikan sebagai nilai trhreshold sehingga diperoleh akurasi terbaik setelah menggunakan information gain pada **Gambar 2**.



Gambar 2. Hasil Akurasi Setelah Menggunakan Seleksi Fitur

Dari hasil penseleksian fitur akurasi 68 % adalah nilai akurasi terbaik yang diperoleh dengan fitur yang terpilih hanyalan 10 dari 65 matakuliah yang di inputkan sebagai pengukuran prediksi kelulusan mahasiswa.

4. KESIMPULAN

Berdasarkan hasil dari penelitian yang dilakukan untuk mengklasifikasi mahasiswa konsentrasi IF dengan data latih yang digunakan dari lulusan tahun 2012 – 2014 dan data uji sebayak 35 mahasiswa untuk lulusan tahun 2015 kemudian diuji cobakan terhadap data uji baru dengan akurasi kebenaran 62 % ini di sebabkan jumlah data uji untuk pengklasifikasian hanya 35 mahasiswa dan terdapat beberapa faktor internal lain seperti pergantian matakuliah yang menyebabkan pengujian tidak maksimal karena tidak terdapat pada kasus lama kemudian untuk meningkatkan akurasi digunakan teknik information gain hanya 10 matakuliah dari 65 matakuliah yang digunakan untuk pengklasifikasian kemudian hasil seleksi diuji cobakan terhadap data uji baru dan hasilnya tidak jauh beda yakni 68 % dengan ini disimpulkan bahwa dalam memprediksi kelulusan menggunakan atribut matakuliah tidak harus semua matakuliah diproses hanya beberapa matakuliah saja yang sebenarnya baik untuk dijadikan sebuah pengklasifikasian dalam memprediksi kelulusan mahasiswa tepat waktu dan terlambat.

Untuk Penelitian selanjutnya dalam menguji pelatihan sebaiknya menggunakan metode khusus seperti *K-fold corss validation* dan untuk data sebaiknya di tambahkakan faktor eksternal yaitu diluar faktor nilai matakuliah atau perkuliahan serta bukan satu konsentrasi saja.

DAFTAR PUSTAKA

Arifin Muhammad, *IG-KNN Untuk Prediksi Customer Churn Telekomunikasi*, SIMETRIS, ISSN: 2252-4983,vol 6, no 1 April 2015.

D. P. K. Muhammad Yunus, *Desain Model Prediksi Kelulusan Mahasiswa Dengan Algoritma Decision Tree*, Matrik, vol. 2, no. 13, pp. 1-5, 2015.

Kusumadewi Sri, *Klasifikasi Status Gizi Menggunakan Naïve Bayesian Classification*, commIT, vol.3 no. 1,pp. 6-11, 2009.

Laily Diana, Fithri, darmanto Eko,*Sistem Pendukung Keputusan Untuk Memprediksi Kelulusan Mahasiswa menggunakan Metode Naïve Bayes*,Proseding SNATIF Ke-1, ISBN:978-602-1180, pp. 319 – 324, 2014.

Mujib Ridwan, Suryono Hadi, M.Sarosa,*Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma Naïve Bayes*,EECCIS, vol 7 no. 1,pp. 59 – 64, 2013.

ISBN 978-602-99334-5-1

Purwarianti Ayu, *Eksperimen pada Sistem Klasifikasi berita Hoax Berbahasa Indonesia Berbasis Pembelarajan mesin*, Rasywir Erissya, Cybermatika, Vol 3, no.2, pp 1- 8, 2015.

W. D. Hera Wasiati, *Sistem Pendukung Keputusan Penentuan Kelayakan Calon Tenaga Kerja Inddonesia Menggunakan Metode Naive Bayes (Studi Kasus: Di P.T. Karyatama Mitra Sejati Yogyakarta)*, IJNS, vol. 3, no. 2, pp. 45 - 51, 2014.

W. Maharani, *Klasifikasi Data Menggunakan JST Backpropagation Momentum Dengan Adaptive Learning Rate*, semnasIF, pp. D-25 - D-31, 2009.