

APLIKASI PREDIKSI KELULUSAN TEPAT WAKTU DENGAN ALGORITMA NAÏVE BAYES (Studi Kasus Program Studi Teknik Informatika, UMMI)

Indra Griha Tofik Isa¹

¹Program Studi Teknik Informatika, UMMI, email : igtiku@gmail.com

ABSTRAK

Penggunaan data mining sudah menjadi tren dalam pengolahan data karena ketersediaannya data dalam jumlah besar serta semakin banyaknya kebutuhan untuk mengubah data tersebut menjadi informasi maupun pengetahuan yang berguna. Selain sebagai alat bantu dalam mengekstraksi data, data mining juga digunakan sebagai pendukung keputusan, baik dalam bidang komersil maupun non-komersil. Dari sekian banyak algoritma yang digunakan dalam data mining, salah satunya adalah Algoritma Naïve Bayes, dimana dalam algoritma ini merupakan salah satu metode pada *Probabilistic Reasoning* yang bertujuan untuk melakukan klasifikasi data pada kelas tertentu. Dalam penelitian yang dilakukan dengan menerapkan tahapan-tahapan Algoritma Naïve Bayes pada perancangan aplikasi untuk memprediksi kelulusan tepat waktu mahasiswa berdasarkan parameter yang terdapat pada Penerimaan Mahasiswa Baru (PMB), Indeks Prestasi Kumulatif (IP) dan Bagian Keuangan. Data yang digunakan adalah data mahasiswa Program Studi Teknik Informatika Tahun Ajaran 2016/2017. Perancangan sistem menggunakan pemodelan UML dengan implementasi menggunakan bahasa pemrograman PHP dan basis data MySQL
Keywords—*Data Mining, Naïve Bayes, Aplikasi, Sistem Pendukung Keputusan*

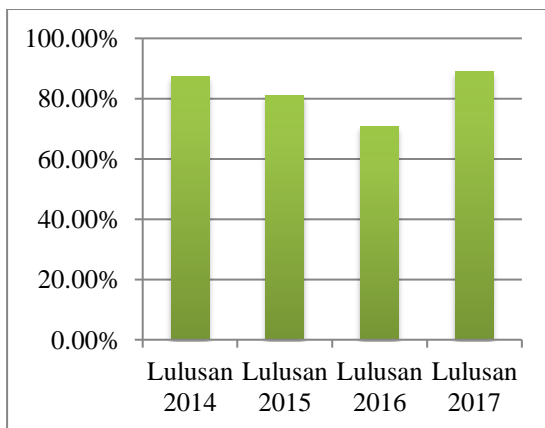
PENDAHULUAN

Data mining adalah bagaimana menerjemahkan fenomena-fenomena yang terjadi, dimana secara tidak sadar hal tersebut merupakan informasi dan pengetahuan yang sangat berguna. Saat ini apapun menggunakan media teknologi informasi dalam mengolah data, sehingga banyak data-data tersebar bahkan terbuang hanya sebatas digunakan sebagai output dari proses sebuah sistem. Padahal dari data-data tersebut dapat

digunakan sebagai informasi yang berguna bagi pihak yang berkepentingan. Tentu, untuk mengekstraksi data tersebut diperlukan sebuah metode, salah satunya Data Mining. Algoritma yang digunakan dalam data mining pun beragam antara lain[5]: C4.5, K-Means, Support Vector Machine, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, CART.

Universitas Muhammadiyah Sukabumi merupakan Universitas terbesar di Kota

Sukabumi dengan memiliki 7 Fakultas dan 20 Program Studi, dimana salah satunya adalah Program Studi Teknik Informatika (Prodi TI). Jumlah mahasiswa yang dimiliki oleh Prodi TI pada Tahun Ajaran 2016/2017 berjumlah 314 mahasiswa[1]. Ada beberapa parameter yang menjadi acuan dalam penilaian keberhasilan Program Studi, salah satu indikatornya adalah tingkat kelulusan tepat waktu, yang artinya adalah seorang mahasiswa menyelesaikan perkuliahannya dalam kurun waktu selama 7-8 semester. Bila dilihat dari tren angka lulusan tepat waktu, dengan mengambil selisih mahasiswa tingkat akhir dan jumlah lulusan, angka lulusan dari tahun 2014 hingga 2017 cenderung fluktuatif. Pada tahun 2014 tingkat kelulusan tepat waktu berjumlah 87.30%, tahun 2015 berjumlah 81.16%, tahun 2016 berjumlah 70.91%, tahun 2017 berjumlah 89.06%. Secara grafik dapat dilihat pada gambar 1.



Gambar 1. Tingkat Kelulusan Tepat Waktu Program Studi Teknik Informatika Tahun 2014-2017

Banyak faktor yang dapat mempengaruhi dari tingkat kelulusan tepat waktu tersebut. Data mahasiswa dapat dijadikan referensi dalam menentukan tingkat

kelulusan tepat waktu. Namun dibutuhkan suatu metode dalam menggali informasi dari data mahasiswa, salah satunya dengan menggunakan Algoritma Data Mining yakni Naïve Bayes. Kaitannya dalam penelitian ini adalah bagaimana menggali informasi mahasiswa berdasarkan NIM, Nama Mahasiswa, IPK, Pilihan Pertama Jurusan, Pilihan Kedua Jurusan, Persentasi Histori Pembayaran (tingkat ketepatan pembayaran) untuk memprediksi kelulusan tepat waktu. Hasil dari data tersebut dijadikan parameter dalam perancangan aplikasi prediksi kelulusan tepat waktu mahasiswa.

DATA MINING

Data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar [2]. Data mining diartikan juga sebagai proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar[4]. Proses penggalian nilai tambah berupa pengetahuan / *Knowledge Discovery in Database* (KDD) dalam Data Mining adalah:

1. *Selection*
2. *Pre-Processing / Cleaning*
3. *Transformation*
4. *Data Mining*
5. *Interpretation*

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu :

1. Deskripsi
2. Estimasi
3. Prediksi
4. Klasifikasi
5. Pengklusteran
6. Asosiasi

$P(H|X)$: Probabilitas hipotesis H berdasarkan kondisi X (*Posteriori Probability*)

$P(H)$: Probabilitas hipotesis H (*Prior Probability*)

$P(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$ Probabilitas X

ALGORITMA NAÏVE BAYES

Salah satu algoritma yang digunakan dalam data mining adalah Algoritma Naïve Bayes, dimana merupakan bagian dari teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang ditemukan oleh ilmuwan Inggris, yaitu Thomas Bayes dengan memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naïve dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naïve Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Berikut ini adalah persamaan dari teorema Bayes:

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

Keterangan:

- X : Data dengan class yang belum diketahui
- H : Hipotesis data X merupakan suatu class spesifik

METODOLOGI

Secara garis besar ada beberapa tahapan yang dilakukan dalam penelitian ini yaitu:

1. Pemilihan data yang digunakan
Ini merupakan tahapan awal, dimana penelitian ini menggali beberapa informasi mahasiswa yang diambil dari data Penerimaan Mahasiswa Baru (PMB), Data Akademik Mahasiswa (SIAK) dan Data Keuangan Mahasiswa periode PMB 2010-2013 dan data lulusan mahasiswa Tahun 2014-2017. Data yang dihasilkan merupakan data mentah yang akan diolah pada tahapan berikutnya.
2. Mereduksi data yang tidak digunakan
Dari data mentah pada tahapan sebelumnya, selanjutnya adalah penyortiran data yang akan digunakan, dari tahapan ini data yang digunakan adalah NIM, Nama Mahasiswa, IPK, Pilihan Pertama Jurusan, Pilihan Kedua Jurusan, Tahun Lulus, Persentasi Keterlambatan Pembayaran.
3. Implementasi Data Mining
Tahapan ini merupakan bagian dari implementasi data mining dari data yang sudah disortir. Pada tahapan ini algoritma Naïve Bayes diterapkan dalam aplikasi yang

dirancang. Pengembangan Aplikasi dengan menggunakan Microsoft Visual Studio 2012 dengan basis data MySQL. Data-data mentah diinputkan untuk diolah dengan algoritma naïve bayes yang sudah diterjemahkan ke dalam bahasa pemrograman, yang menghasilkan output berupa rekomendasi. Perancangan sistem menggunakan pendekatan berorientasi objek dengan pemodelan *Unified Modelling Language*.

HASIL DAN PEMBAHASAN

1. Pemilihan Data yang digunakan

Data awal yang digunakan dalam penelitian ini berupa dataset excel yang terdapat pada bagian Penerimaan Mahasiswa Baru (PMB), Sistem Informasi Akademik (SIK) dan Data keuangan Mahasiswa. Data yang diolah merupakan data penerimaan mahasiswa baru Tahun Ajaran 2010/2011, 2011/2012, 2012/2013 dan 2013/2014. Data yang diolah sebanyak 60% dari data PMB, dengan jumlah total data sebanyak 436 dataset. Data yang diambil merupakan data mahasiswa yang aktif hingga 7-8 semester, untuk data mahasiswa PMB yang dengan status *Drop Out* tidak dimasukkan ke dalam dataset.

Tabel 1. Data PMB Periode 2010-2014

No	Nama	Tahun Daftar	PIL_1	PIL2
1.	Agustian	2012	TI	SASIN G
2	Fajar Ambar	2012	TI	ADBIS
3	Siti	2013	TI	TS

No	Nama	Tahun Daftar	PIL_1	PIL2
	Salwa			
4	Maria Mulfah	2013	TI	TS
5	Adi Fadilah	2010	TI	TS
...

Adapun data SIAK yang diambil merupakan mahasiswa semester 8 berdasarkan Tahun Angkatan mahasiswa tersebut menjadi mahasiswa Teknik Informatika UMMI. Jumlah dataset mahasiswa sesuai dengan jumlah dataset PMB yakni sebanyak 436 dataset. Data yang diambil dari SIAK adalah NIM, Nama Mahasiswa, IPK, Tahun Lulus (estimasi tepat waktu 8 semester). Sedangkan dari data Keuangan diperoleh data pembayaran, dengan menghitung jumlah histori keterlambatan pembayaran dibagi total masa pembayaran. Dataset yang diperoleh berupa NIM, Nama Mahasiswa, Persentasi Keterlambatan Pembayaran.

Tabel 2. Data Keuangan Periode 2010-2014

No	Nama	Cara Bayar	% Bayar
1.	Agustian	SMT	0%
2	Fajar Ambar	SMT	10%
3	Siti Salwa	SMT	0%
4	Maria Mulfah	SMT	0%
5	Adi Fadilah	SMT	20%
...

2. Mereduksi data yang tidak digunakan

Dari data PMB, SIAK dan Keuangan, tidak seluruhnya data digunakan, parameter yang dilibatkan adalah NIM, Nama Mahasiswa, Pilihan 1 Jurusan, Pilihan 2 Jurusan, Tahun Lulus dan Tingkat Keterlambatan Pembayaran (dalam %). Pada tahapan ini dilakukan pengujian dataset dengan algoritma Naïve Bayes untuk melihat pola dari data-data tersebut. Semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain. Apabila diberikan *k* atribut yang bebas (*independence*), nilai probabilitas dapat diberikan sebagai berikut:

$$P(x_1...x_k/C) = P(x_1|C)x_2 ... xP(x_k|C)(2)$$

Tahap awal dari perhitungan Naïve Bayes adalah dengan melakukan pengambilan data training dari data mahasiswa (PMB, SIAK, Keuangan). Adapun variable penentu yang digunakan dalam mengklasifikasikan data kelulusan tepat waktu mahasiswa yaitu:

a. Pilihan 1 Jurusan

Merupakan variable pilihan pertama mahasiswa baru dalam menentukan jurusan perkuliahan yang dipilihnya

b. Pilihan 2 Jurusan

Merupakan variable pilihan kedua mahasiswa baru dalam menentukan jurusan perkuliahan yang dipilihnya

c. Tahun Lulus

Tahun kelulusan mahasiswa, dengan estimasi 8 semester (4 tahun) dari pendaftaran mahasiswa baru.

d. Cara Bayar

Merupakan variable cara pembayaran perkuliahan, dikategorikan dalam 3 kategori yakni Bulanan, Semester dan Tahunan

e. % Bayar

Merupakan variable tingkat keterlambatan pembayaran mahasiswa, diambil dari histori intensitas keterlambatan pembayaran mahasiswa dibagi dengan intensitas pembayaran per semester

f. Tepat Waktu

Merupakan selisis antara Tahun Lulus dengan Tahun Daftar. Tepat waktu dikategorikan ke dalam 2 kategori yakni TEPAT dan TIDAK TEPAT

TABEL 3. DATA SET MAHASISWA

No	Nama Mahasiswa	Tahun Daftar	Pil1_ Jur	Pil2_ Jur	Thn Lulus	IPK	Cara Bayar	% Bayar	Tepat Waktu
1	Agustian	2012	TI	SASING	2016	3.45	SMT	0%	TEPAT
2	Fajar Ambar	2012	TI	ADBIS	2016	3.11	SMT	10%	TEPAT
3	Siti Salwa	2013	TI	TS	2017	3.70	SMT	0%	TEPAT
4	Maria Mulfah	2013	TI	TS	2017	3.26	SMT	0%	TEPAT
5	Adi Fadilah	2010	TI	TS	2016	3.01	SMT	20%	TIDAK

No	Nama Mahasiswa	Tahun Daftar	Pil1_Jur	Pil2_Jur	Thn Lulus	IPK	Cara Bayar	% Bayar	Tepat Waktu
									TEPAT
6	Shandra Seplyani	2012	TS	TI	2017	2.75	BLN	10%	TIDAK TEPAT
7	Wandra A	2011	TI	ADPUB	2015	3.04	BLN	0%	TEPAT
8	Radous Andhika	2010	TI	AKT	2014	3.26	THN	0%	TEPAT
9	Jihan Fideyna	2013	TI	AKT	2017	2.86	THN	10%	TEPAT
10	Feri Andriansyah	2012	TI	SASING	2016	2.93	THN	0%	TIDAK TEPAT
...

3. Implementasi Data Mining

Dalam perhitungan algoritma Naïve bayes, diperlukan penerjemahan dataset tabel 2 ke dalam bentuk biner (1 dan 0), sehingga disimpulkan terdapat 4 variabel yaitu Pilihan1 Jurusan, IPK, % Bayar dan Tepat Waktu.

Setiap variabel memiliki 2 nilai dan variabel class juga memiliki 2 nilai, seperti pada penjelasan di bawah ini:

Variabel: **Pilihan 1 Jurusan**

- a. Jurusan TI : 1
- b. Jurusan Non TI : 0

Variabel: **IPK**

- a. $IPK \geq 3$: 1
- b. $IPK < 3$: 0

Variabel: **% Bayar**

- a. Bila 0% : 1
- b. Bila > 0% : 0

Variabel: **Tepat Waktu**

- a. TEPAT : 1
- b. TIDAK TEPAT : 0

Hasil Konversi dataset dapat dilihat pada Tabel 4 di bawah ini:

Tabel 4. Hasil Konversi Dataset'

Nama Mahasiswa	Pil1_Jur	IPK	% Bayar	Tepat Waktu
Agustian	1	1	1	1
Fajar Ambar	1	1	0	1
Siti Salwa	1	1	1	1
Maria Mulfah	1	1	1	1
Adi Fadilah	1	1	0	0
Shandra Seplyani	0	0	0	0
Wandra A	1	1	1	1
Radous Andhika	1	1	1	1
Jihan Fideyna	1	0	0	1
Feri Andriansyah	1	0	1	0
...

Terdapat dua jenis peluang yang perlu dihitung dari dataset untuk tabel di atas, yakni *Class Probabilities* (Tepat Waktu) dan *Conditional Probabilities*. Perhitungan *Class Probabilities* (Tepat Waktu) dilakukan dengan:

- a. $P(\text{Tepat Waktu}=1) = \frac{\text{Jumlah}(\text{Tepat Waktu}=1)}{\text{Jumlah}(\text{Tepat Waktu}=1) + \text{Jumlah}(\text{Tepat Waktu}=0)}$
- b. $P(\text{Tepat Waktu}=0) = \frac{\text{Jumlah}(\text{Tepat Waktu}=0)}{\text{Jumlah}(\text{Tepat Waktu}=1) + \text{Jumlah}(\text{Tepat Waktu}=0)}$

Maka didapatkan:

- a. $P(\text{Tepat Waktu}=1) = 7 / (3+7) = 0.7$
- b. $P(\text{Tepat Waktu}=0) = 3 / (3+7) = 0.3$

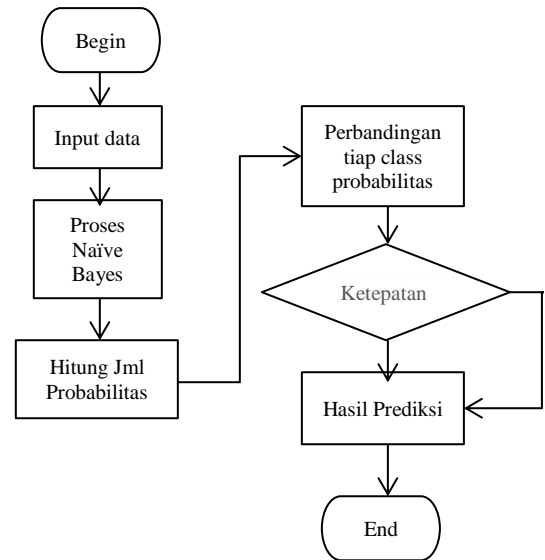
Berikutnya adalah menentukan peluang dari masing-masing variabel tersebut, misalnya untuk variabel IPK, maka aturan peluang berdasarkan formula Naïve Bayes salah satunya adalah:

$$P(\text{IPK} \geq 3 | \text{Tepat Waktu}) = \frac{\sum_{\text{IPK} \geq 3 \text{ and } \text{Tepat Waktu}}}{\sum_{\text{Tepat waktu}}} \dots\dots\dots(3)$$

Begitupun dengan variabel-variabel lainnya, diperlukan perhitungan untuk menentukan probabilitas dari masing-masing variabel.

4. Perancangan dan Implementasi Sistem

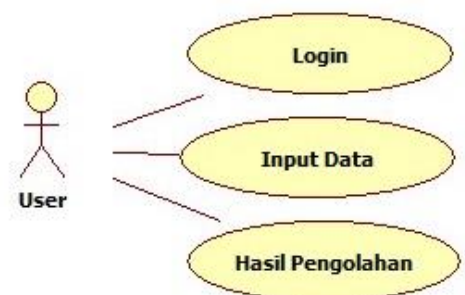
Dalam perancangan sistem yang dibuat, alur logika atau *flowchart* dari aplikasi yang dibuat terdapat pada Gambar 2:



Gambar 2. Flowchart Sistem yang diusulkan

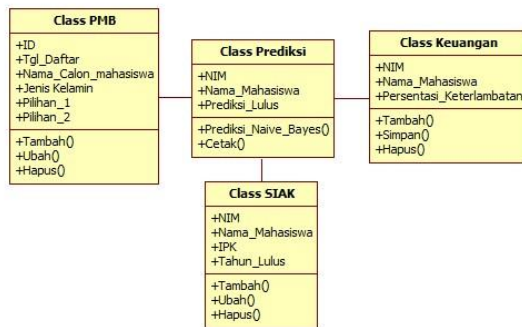
Pada *flowchart* di atas dimulai dengan input data dengan variabel-variabel yang sudah ditentukan, selanjutnya dilakukan proses perhitungan *Naïve Bayes*, dengan mengkalkulasikan jumlah probabilitas dan perbandingan tiap kelas dari variabel data yang diinput sebelumnya. Sehingga pada hasil akhir muncul prediksi kelulusan dari variabel mahasiswa yang diinput tersebut

Adapun dalam perancangan sistem dilakukan dengan pemodelan *Unified Modelling Language* (UML), dengan *Use Case diagram* pada Gambar 3 dan *class diagram* terdapat pada Gambar 4 :



Gambar 3. Use Case Diagram Aplikasi Prediksi Kelulusan Tepat Waktu Program Studi Teknik

Informatika Universitas Muhammdiyah
Sukabumi



Gambar 4. Class Diagram Aplikasi Prediksi Kelulusan Tepat Waktu Program Studi Teknik Informatika Universitas Muhammdiyah Sukabumi

Perancangan Basis Data

Berikut ini adalah perancangan dari basis data yang sudah dibuat :

Tabel 5. Tabel PMB

Field	Tipe Data	Length
ID	Varchar	20
NIM	Varchar	20
Nama_Mhs	Varchar	50
Tgl_Daftar	Date/Time	Short
Pilihan_1	Varchar	20
Pilihan_2	Varchar	20

Tabel 6. Tabel SIAK

Field	Tipe Data	Length
NIM	Varchar	20
Nama_Mhs	Varchar	50
IPK	Double	
Tahun_Lulus	Integer	

Tabel 7. Tabel Keuangan

Field	Tipe Data	Length
NIM	Varchar	20

Field	Tipe Data	Length
Nama_Mhs	Varchar	50
Cara_Bayar	Varchar	20
Persentasi_Bayar	Integer	

Tabel 8. Tabel Prediksi

Field	Tipe Data	Length
NIM	Varchar	20
Nama_Mhs	Varchar	50
Prediksi_Lulus	Varchar	20

HASIL PENGUJIAN

Pengujian dilakukan dengan menginput beberapa variabel baru sejumlah 53 data mahasiswa, dari hasil perhitungan data mining terdapat 1 data yang tidak sesuai dengan data riil kelulusan mahasiswa. Sehingga bila diterjemahkan ke dalam tingkat eror data sejumlah 1.88% yang didapatkan dari :

$$\frac{\text{Jumlah Data Error}}{\text{Jumlah Seluruh Variabel}} \times 100\% \dots\dots\dots(4)$$

Jumlah data error sejumlah 1 data, dan jumlah seluruh variabel yang diinput berjumlah 53 data.

KESIMPULAN

Naïve Bayes dapat dijadikan sebuah metode dalam menggali informasi dari data yang belum dijadikan pengetahuan/informasi yang berguna bagi yang berkepentingan. Hasil akhir dari penelitian ini adalah aplikasi prediksi kelulusan tepat waktu bagi

mahasiswa Teknik Informatika Universitas Muhammadiyah Sukabumi. Berdasarkan hasil pengujian didapatkan tingkat error data sebanyak 1.88%. Sehingga disarankan perlu dilakukan analisis perbandingan dengan algoritma data mining lainnya untuk melihat tingkat error pada algoritma-algoritma tersebut. Sehingga akan mengoptimalkan proses ekstraksi data warehouse melalui data mining.

DAFTAR PUSTAKA

- [1] Borang Akreditasi Sarjana Program Studi Teknik Informatika – Universitas Muhammadiyah Sukabumi. 2017
- [2] Hermawati, Fajar Astuti, “Data Mining”, Penerbit Andi, Yogyakarta, 2013
- [3] Larose, Daniel T, “*Data Mining Methods and Models*”, John Wiley & Sons, Inc., New Jersey, 2006
- [4] Turban, E, dkk, “*Decision Support Systems and Intelligent Systems*”, Andi Offset, 2005.
- [5] <https://www.ieee.org/index.html>