Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier

M. Syukri Mustafa*[1], I Wayan Simpen[2]
[1,2]Program Studi Teknik Informatika, STMIK Dipanegara, Makassar
E-mail: *[1]syukri@dipanegara.ac.id, [2]wayan@dipanegara.ac.id

Abstrak

Penelitian ini difokuskan untuk mengevaluasi kinerja akademik mahasiswa STMIK Dipanegara Makassar pada tahun ke-2 dan diklasifikasikan dalam kategori mahasiswa yang dapat lulus tepat waktu atau tidak. Input dari sistem ini adalah data induk mahasiswa dan data akademik mahasiswa. Sampel nilai yang digunakan untuk data latih dan testing adalah nilai mahasiswa angkatan 2008-2011 yang sudah dinyatakan lulus, sedangkan mahasiswa angkatan 2013-2014 dan belum lulus akan digunakan sebagai data target. Data input akan diproses menggunakan teknik data mining algoritma Naive Bayes Classifier (NBC) untuk membentuk tabel probabilitas sebagai dasar proses klasifikasi kinerja akademik mahasiswa yang kelulusannya akan diklasifikasikan dan memberikan rekomendasi untuk proses kelulusan tepat waktu yang paling tepat dengan nilai optimal berdasarkan histori nilai yang telah ditempuh mahasiswa. Hasil pengujian menunjukkan bahwa faktor yang paling berpengaruh dalam penentuan klasifikasi kinerja akademik mahasiswa yaitu Indeks Prestasi (IP) semester 1,2,3,4 dan jenis kelamin. Sehingga faktor-faktor tersebut dapat digunakan sebagai bahan evaluasi bagi pihak pengelola STMIK Dipanegara. Pengujian pada beberapa data mahasiswa angkatan 2008-2011 yang diambil secara acak, algoritma NBC menghasilkan nilai akurasi 92,3%.

Kata kunci — Algoritma Naive Bayes Classifier, Kinerja Akademik Mahasiswa.

Abstract

This study focused on evaluating the students' academic performance STMIK Dipanegara Makassar in year 2 and classified under the categories of students who can graduate on time or not. The input of this system is the master data of students and student academic data. Sample values used for training data and testing is the value of 2008-2011 generation students who already passed, while the younger students have not passed the 2013-2014 and will be used as target data. Input data will be processed using data mining techniques Naive Bayes classifier algorithm (NBC) to establish a probability table as the basis of the classification process academic performance of students whose graduation will be classified and provide recommendations for the process of graduation on time most appropriate to the optimum value based on historical values that have been taken college student. The test results showed that the most influential factor in determining the classification of students' academic performance is Grade (IP) semesters 1,2,3,4 and gender. So that these factors can be used as an evaluation for the manager STMIK Dipanegara. Tests on some data 2008-2011 generation students are captured at random, the algorithm NBC produces a value 92.3% accuracy.

Keywords— *Naive Bayes classifier algorithm, Student Academic Performance.*

1. Pendahuluan

Mahasiswa merupakan salah satu aspek penting dalam evaluasi keberhasilan penyelenggaran program studi pada suatu perguruan tinggi. Pemantauan mahasiswa yang masuk, peningkatan kemampuan mahasiswa, dan rasio kelulusan terhadap jumlah total mahasiswa, dan kompetensi lulusan semestinya mendapatkan perhatian yang serius untuk memperoleh kepercayaan *stakeholder* dalam menilai dan menetapkan kelulusannya.

Berdasarkan uraian di atas, pada penelitian ini akan dibuat sebuah sistem untuk mengklasifikasikan kelulusan mahasiswa dengan cara mengevaluasi kinerja pada tahun pertama dan tahun kedua. Pada penelitian ini, digunakan teknik data mining untuk menemukan pola kelulusan

mahasiswa yang sudah lulus, kemudian dijadikan dasar untuk mengklasifikasikan mahasiswa tahun kedua yang bisa lulus tepat waktu dan yang tidak.

Data mining adalah proses menemukan hubungan dalam data yang tidak diketahui oleh pengguna dan menyajikannya dengan cara yang dapat dipahami sehingga hubungan tersebut dapat menjadi dasar pengambilan keputusan. Teknik data mining yang akan digunakan dalam penelitian ini adalah algoritma NBC yang merupakan sebuah pengklasifikasian probabilitas sederhana yang mengaplikasikan teorema bayes. Algoritma NBC dapat mengolah data kuantitatif dan data diskrit yang hanya memerlukan sejumlah kecil data pelatihan untuk perhitungan estimasi peluang yang dibutuhkan untuk klasifikasi. Perhitungan Algoritma NBC dibandingkan dengan algortima klasifikasi yang lain lebih cepat karena hanya menguji probabilitas dengan menemukan class yang sama dari data training. Ide dasar dari Teorema Bayes adalah menangani masalah yang bersifat hipotesis yakni mendesain suatu klasifikasi untuk memisahkan objek.

Berdasarkan latar belakang yang telah diuraikan di atas, maka pokok permasalahannya yaitu bagaimana mengklasifikasikan mahasiswa yang telah menyelesaikan 4 semester yang dapat lulus tepat waktu dan yang tidak dapat lulus tepat waktu dengan melihat pola kelulusan mahasiswa STMIK Dipanegara Makassar beberapa periode sebelumnya menggunakan Algoritma *Naive Bayes Classifier ?*

2. Metode Penelitian

Metode Penelitian (bisa meliputi analisa, arsitektur, metode yang dipakai untuk menyelesaikan masalah, implementasi), dalam bahasan ini penulis bisa menguraikan bagaimana penelitian tersebut akan dilakukan.

2.1. Data mining

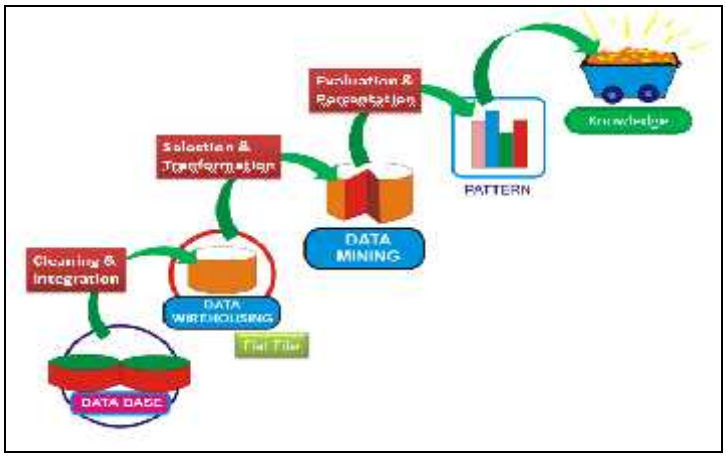
2.1.1. Definisi Data Mining

Data mining adalah adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual[1].

Istilah *data mining* memiliki hakikat sebagai dispilin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. *Data mining*, sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan pola atau hubungan dalam set data berukuran besar.

2.1.2. Definisi Data Mining

Data mining sangat perlu dilakukan terutama dalam mengelola data yang sangat besar untuk memudahkan aktifitas recording suatu transaksi dan untuk proses data warehousing agar dapat memberikan informasi yang akurat bagi penggunanya. Alasan utama mengapa data mining sangat menarik perhatian industri informasi dalam beberapa tahun belakangan ini adalah karena tersedianya data dalam jumlah yang besar dan semakin besarnya kebutuhan untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna karena sesuai fokus bidang ilmu ini yaitu melakukan kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran/berjumlah besar, informasi inilah yang nantinya sangat berguna untuk pengembangan. Langkah-langkah untuk melakukan data mining dapat dilihat pada gambar 1.



Gambar 1. Contoh Konsep *Data Mining* [2]

1. *Data cleaning* (untuk menghilangkan *noise* data yang tidak konsisten) Data *integration* (di mana sumber data yang terpecah dapat disatukan)
2. *Data selection* (dimana data yang relevan dengan tugas analisis dikembalikan ke dalam database)
3. *Data transformation* (dimana data berubah atau bersatu menjadi bentuk yang tepat untuk menambang dengan ringkasan performa atau operasi agresi)
4. *Knowledge Discovery* (proses esensial di mana metode yang intelejen digunakan untuk mengekstrak pola data)
5. *Pattern evolution* (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan atas beberapa tindakan yang menarik)
6. *Knowledge presentation* (di mana gambaran teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah ditambang kepada *user*).

2.1.3. Metode Data Mining

Ada banyak metode atau fungsi *data mining* yang bisa digunakan untuk menemukan, menggali dan menambang pengetahuan. Ada enam fungsi utama *data mining*, yaitu[3]:

1. *Description* (deskripsi), untuk memberi gambaran secara ringkas bagi sekumpulan data yang jumlahnya sangat besar dan banyak jenisnya. Termasuk dalam fungsi ini adalah metode *Decision Tree*, *Neural Network*, dan *Exploratory Data Analysis.*
2. *Estimation* (estimasi), untuk menerka sebuah nilai yang belum diketahui, misal menerka penghasilan seseorang ketika informasi mengenai orang tersebut diketahui. Metode yang digunakan antara lain *Point Estimation* dan *Confidence Interval Estimations*, *Simple Linear Regression* dan *Correlation*, dan *Multiple Regression.*
3. *Prediction* (prediksi), untuk memperkirakan nilai masa mendatang, misal memprediksi stok barang satu tahun ke depan. Fungsi ini mencakup metode *Neural Network*, *Decision Tree*, dan *k–Nearest Neighbor*.
4. *Classification* (klasifikasi), merupakan proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Metode yang digunakan antara lain *Neural Network, Decision Tree, k-Nearest Neighbor*, dan *Naive Bayes.*
5. *Clustering* (pengelompokan), yaitu pengelompokan mengidentifikasi data yang memiliki karakteristik tertentu. Metode dalam fungsi ini diantaranya *Hierarchical Clustering*, metode *K-Means*, dan *Self Organizing Map* (SOM)
6. *Association* (asosiasi), dinamakan juga analisis keranjang pasar dimana fungsi ini mengidentifikasi item-item produk yang kemungkinan dibeli konsumen bersamaan dengan produk lain. Metode atau algoritma dalam fungsi ini adalah *Apriori, Generalized Sequential Pattern (GSP), FP-Growth* dan *GRI algorithm*

2.2. Algoritma Naive Bayes Classifier

2.2.1. Definis Algoritma

Algoritma adalah (1) teknik penyusunan langkah-langkah penyelesaian masalah dalam bentuk kalimat dengan jumlah kata terbatas tetapi tersusun secara logis dan sistematis. (2) Suatu prosedur yang jelas untuk menyelesaikan suatu persoalan dengan menggunakan langkah-langkah tertentu dan terbatas jumlahnya. (3) Susunan langkah yang pasti, yang bila diikuti maka akan mentransformasi data input menjadi output yangberupa informasi[4].

2.2.2. Pengertian Algoritma NBC

Teorema Bayes adalah teorema yang digunakan dalam statistika untuk menghitung peluang untuk suatu hipotesis. Bayes Optimal Classifier menghitung peluang dari suatu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal.

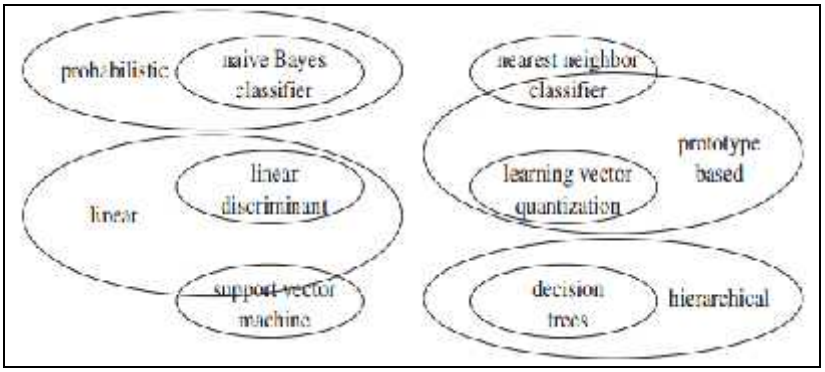
Pengklasifikasian menggunakan Teorema Bayes ini membutuhkan biaya komputasi yang mahal (waktu prosessor dan ukuran memory yang besar) karena kebutuhan untuk menghitung nilai probabilitas untuk tiap nilai dari perkalian kartesius untuk tiap nilai atribut dan tiap nilai kelas. Data latih untuk Teorema Bayes membutuhkan paling tidak perkalian kartesius dari seluruh kelompok atribut yang mungkin, jika misalkan ada 16 atribut yang masing-masingnya berjenis boolean tanpa missing value, maka data latih minimal yang dibutuhkan oleh Teorema bayes untuk digunakan dalam klasifikasi adalah 216 = 65.536 data. Untuk mengatasi kekurangan tersebut maka digunakan *Naive Bayes*.

Naive Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dari *Naive Bayes Classifier* ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi atau kejadian.

Naive Bayes untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari "master" tabel keputusan[5].

Naive Bayes Classifier bekerja sangat baik dibanding dengan model *classifier* lainnya. Hal ini dibuktikan oleh Xhemali, Hinde dan Stone dalam jurnalnya "*Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages*" mengatakan bahwa "*Naïve Bayes Classifier* memiliki tingkat akurasi yang lebih baik dibanding model *classifier* lainnya".

Pada gambar 2 dapat diliat skema yang sering digunakan dalam proses klasifikasi, yang tentunya juga menyertakan *Naive Bayes Classifier*.



Gambar 2 *Skema Klasifikasi Algoritma NBC*

Formula perhitungan *Naive Bayes Classifier* berdasarkan probabilitas ditunjukkan sebagai berikut :

$p(A|B).p(B) = p(B|A).p(A)$

$$p(A_i|B) = \frac{p(A_i).p(B|A_i)}{\sum_{j=1}^{n} p(A_j).p(B|A_j)}$$

Dengan mengubah nilai A_i dan A_j kedalam vector "x" maka didapatkan bentuk formula sebagai berikut :

$$p(x|i) = \frac{p(i|x).p(x)}{\sum_{j=1}^{n} p(j).p(x|j)}$$

Adapun perhitungan *Naive Bayes Classifier* untuk data kontinu menggunakan *Distribusi Gauss* sebagai berikut:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Keterangan :
$p(x|i)$ = Probabilitas hipotesis x jika diberikan fakta atau *record* i (*Posterior probability).*
$p(i|x)$ = Mencari nilai parameter yang memberi kemungkinan yang paling besar *(Likelihood)*
$p(x)$ = *Prior probability* dari I (*Prior Probability*)
$p(i)$ = Jumlah *probability tuple* yang muncul
g = *Distribusi Gauss*
μ = Rata-rata
σ = Standar Deviasi

Bila *p(x|i)* dapat diketahui melalui perhitungan diatas, maka kelas (label) dari data sampel *X* adalah kelas (label) yang memiliki *p(x|i) * p(i)* maksimum.

3. Hasil dan Pembahasan

Pada penelitian ini digunakan berbagai macam data, diantaranya data training, data testing, dan juga data *history* matakuliah.

3.1 Data Training

Data Training digunakan untuk membentuk sebuah model *classifier*. Model ini merupakan representasi pengetahuan yang akan digunakan untuk prediksi kelas data baru yang belum pernah ada. Data ini akan digunakan sebagai proses *mining* berupa 1115 sampel data induk akademik mahasiswa STMIK Dipanegara angkatan 2008-2011 yang sudah dinyatakan lulus, dengan pengklasifikasian 781 Mahasiswa Lulus Tepat Waktu, dan 334 Mahasiswa Tidak Lulus Tepat Waktu. Data ini memiliki atribut NIM, nama, jenis kelamin, IPS 1, IPS 2, IPS 3, IPS 4, IPK semester 4, dan keterangan lulus. Beberapa sampel data training dapat dilihat pada tabel 1 berikut.

Tabel 1. Data Training

No	Stambuk	Nama	Jenis Kelamin	IPS1	IPS2	IPS3	IPS4	IPS5	Kelulusan
1	92008	HARTVIG BATURANTE	Laki - Laki	3,285714	0,7142857	2,571429	2,25	2,166667	Terlambat
2	92010	IDIAM	Laki - Laki	3,666667	3,904762	3,181818	3,1	3,464286	Tepat Waktu
3	92011	SYAIFUL SYAH ALAM	Laki - Laki	0	0	1,153846	1,833333	0,5271186	Terlambat
4	92012	MUH. IKRAL S.	Laki - Laki	2,666667	3,428571	2,444444	2,705882	2,831169	Tepat Waktu
5	92018	MIIS NANIA	Laki - Laki	3,476191	3,476191	3,4	3,05	3,353658	Tepat Waktu
6	92052	THEO XAVER YANSEN	Laki - Laki	2,714286	2,476191	1,666667	1,666667	2,376344	Terlambat
7	92053	HERDY BUDYANTO HALIK, S	Laki - Laki	3,476191	3,476191	3,611765	3,611765	3,210382	Tepat Waktu
8	92054	MUHAMMAD IMRAN L.	Laki - Laki	2,619048	2,238095	2,006624	2,5	2,329897	Tepat Waktu
9	92057	SISWANA ASIS	Perempuan	3,571429	3,904762	3,905238	3,388889	3,568627	Tepat Waktu
10	92058	ARDY MAULANA WAHAB	Laki - Laki	1,952381	2,761905	0,5714286	0,6875	1,477273	Terlambat
11	92059	ANDY PUTRA ANUGRA. A	Laki - Laki	2,428571	2,476191	2,682353	2,3	2,404737	Tepat Waktu
12	102119	YUNI MARSELINA PANGGESO	Perempuan	1,857143	2,071429	2,611111	3	2,315217	Terlambat
13	102121	NITSAR PALECHE	Laki - Laki	0,5714286	0,125	2,136364	1,529412	1,173333	Terlambat
14	112052	VERONICA WIJAYA	Perempuan	3,428571	3,571429	3,818182	3,882353	3,817647	Tepat Waktu
15	112055	ELSYE KIONG	Laki - Laki	3,285714	3,071429	1,111111	1,25	2,518868	Terlambat

3.2 Data Training

Data *Testing* digunakan untuk mengukur sejauh mana algoritma NBC berhasil melakukan klasifikasi dengan benar. Karena itu, data yang ada pada testing set seharusnya tidak boleh ada pada

training set sehingga dapat diketahui apakah model classifier sudah “pintar” dalam melakukan klasifikasi. Data ini akan digunakan untuk pengujian, berupa data akademik mahasiswa angkatan 2013-2014. Data ini memiliki atribut STB, nama, jenis kelamin, IPS 1, IPS 2, IPS 3, IPS 4, dan IPK semester 4. Setelah proses *mining*, data ini akan memiliki kelas berdasarkan probabilitas yang diperoleh dari data *training*. Beberapa sampel data testing dan juga hasil klasifikasinya dapat dilihat pada tabel 2 dan tabel 3 berikut.

Tabel 2. *Data Testing*

No	Stambuk	Nama	Jenis Kelamin	IPS1	IPS2	IPS3	IPS4	IPK
1	132011	HASNIATI PAMULA	Perempuan	2,857143	2,190476	2,380952	2	2,37037
2	132012	YOSIA TANDIALO	Laki - Laki	3,095238	3,333333	2,952381	3	3,095238
3	132013	SIGIT ADRIAN RH.	Laki - Laki	3,428571	3,047619	3,523809	3,454545	3,364706
4	142026	MIA AUDINA M	Perempuan	3,5	3,454545	3,333333	3,142857	3,365079
5	142028	FALMASARI SABBAN	Perempuan	3,1	2,863636	3,190476	1,809524	2,738095
6	142029	RESKI AMALIA	Laki - Laki	3,536364	3,9	3,333333	3,523809	3,68254

Tabel 3. Hasil Klasifikasi Data Testing

No	Stambuk	Nama	Jenis Kelamin	IPS1	IPS2	IPS3	IPS4	IPK	Hasil Klasifikasi	Rekomendasi
1	132011	HASNIATI PAMULA	Perempuan	2,857143	2,190476	2,380952	2	2,37037	Tepat Waktu	Tidak
2	132012	YOSIA TANDIALO	Laki - Laki	3,095238	3,333333	2,952381	3	3,095238	Tepat Waktu	Tidak
3	132013	SIGIT ADRIAN RH.	Laki - Laki	3,428571	3,047619	3,523809	3,454545	3,364706	Tepat Waktu	Tidak
4	142026	MIA AUDINA M	Perempuan	3,5	3,454545	3,333333	3,142857	3,365079	Tepat Waktu	Tidak
5	142028	FALMASARI SABBAN	Perempuan	3,1	2,863636	3,190476	1,809524	2,738095	Tepat Waktu	Tidak
6	142029	RESKI AMALIA	Laki - Laki	3,536364	3,9	3,333333	3,523809	3,68254	Tepat Waktu	Tidak

3.3 Data Riwayat Matakuliah

Data ini digunakan untuk mengevaluasi *data testing* ketika diklasifikasikan terlambat. Data ini akan dianalisis untuk memberikan rekomendasi dalam perkuliahan semester berikutnya. Data Riwayat matakuliah dapat dilihat pada tabel 4.

Tabel 4. Data Riwayat Matakuliah

No	Stambuk	Nama	Mata Kuliah	Nilai	No	Stambuk	Nama	Mata Kuliah	Nilai
1	142020	MIA AUDINA M	ALGORITMA DAN PEMROGRAMAN	A	21	142020	MIA AUDINA M	PRAKTIKUM ELEKTRONIKA DIGITAL	A
2	142020	MIA AUDINA M	PRAK. ALGORITMA DAN PEMROGRAMAN	A	22	142020	MIA AUDINA M	PRAK. SISTEM MANAJEMEN DATA	C
3	142020	MIA AUDINA M	PENGEMBANGAN DIRI	A	23	142020	MIA AUDINA M	ANALISIS DAN DESAIN SISTEM	A
4	142020	MIA AUDINA M	KALKULUS I	B	24	142020	MIA AUDINA M	MIKROKONTROLLER	A
5	142020	MIA AUDINA M	PRAK. PENGINAPLAN TEKNOLOGI INFORMASI	A	25	142020	MIA AUDINA M	MULTIMEDIA DAN KOMUNIKASI	A
6	142020	MIA AUDINA M	PRAK. ELEKTRONIKA ANALOG	A	26	142020	MIA AUDINA M	PEMROGRAMAN VISUAL .Net	B
7	142020	MIA AUDINA M	BAHASA INGGRIS I	B	27	142020	MIA AUDINA M	PENGANTAR INTELEGENSI BUATAN	C
8	142020	MIA AUDINA M	PENGINAPLAN TEKNOLOGI INFORMASI	A	28	142020	MIA AUDINA M	PRAK. MIKROKONTROLLER	A
9	142020	MIA AUDINA M	PENDIDIKAN AGAMA ISLAM	B	29	142020	MIA AUDINA M	PRAK. VISUAL .NET	B
10	142020	MIA AUDINA M	ELEKTRONIKA ANALOG	C	30	142020	MIA AUDINA M	SISTEM OPERASI KOMPUTER	B
11	142020	MIA AUDINA M	BAHASA INGGRIS II	C	31	142020	MIA AUDINA M	ETIKA PROFESI	B
12	142020	MIA AUDINA M	ELEKTRONIKA DIGITAL	A	32	142020	MIA AUDINA M	DATA WAREHOUSE & DATA MINING	B
13	142020	MIA AUDINA M	HUKUM DAN KOMPUTER	A	33	142020	MIA AUDINA M	JARINGAN KOMPUTER	B
14	142020	MIA AUDINA M	INTERAKSI MANUSIA DAN KOMPUTER	A	34	142020	MIA AUDINA M	METODE NUMERIK	C
15	142020	MIA AUDINA M	KALKULUS II	C	35	142020	MIA AUDINA M	ORGANISASI DAN ARSITEKTUR KOMPUTER	A
16	142020	MIA AUDINA M	KECAKAPAN ANTAR PERSONAL	A	36	142020	MIA AUDINA M	PEMROGRAMAN WEB	B
17	142020	MIA AUDINA M	KEWIRAUSAHAAN	A	37	142020	MIA AUDINA M	PRAK. JARINGAN KOMPUTER	A
18	142020	MIA AUDINA M	MISTIK DAN ETIKA	B	38	142020	MIA AUDINA M	PRAK. PEMROGRAMAN WEB	C
19	142020	MIA AUDINA M	PANCASILA DAN KEWARGANEGARAAN	A	39	142020	MIA AUDINA M	KEAMANAN PERANGKAT LUNAK	A
20	142020	MIA AUDINA M	PENGANTAR FORENSIK TEKNOLOGI INFORMASI	A	40	142020	MIA AUDINA M	STRUKTUR DATA	B

3.4 Implementasi Algoritma Naïve Bayes Classifier

Implementasi algoritma *Naïve Bayes Classifier* dalam penelitian ini menggunakan Data Testing yang diberikan dapat dilihat pada tabel 5.

Tabel 5. Data Testing

No	NIM	JKL	IPS 1	IPS 2	IPS 3	IPS 4	IPK
1	142037	Laki – Laki	3	2	1,047619	2,173913	2,214286
2	142038	Perempuan	2,272727	2,65	2,285714	2,857143	2,587301

Langkah-langkah Algoritma NBC :

1. Menentukan *Prior Probability* (P)

KELULUSAN	JUMLAH KEJADIAN
Tepat Waktu	781
Terlambat	334
JUMLAH	**1115**

$$P(Tepat\ Waktu) = \frac{781}{1115} = 0.70044 \qquad P(Terlambat) = \frac{334}{1115} = 0.29955$$

2. Menentukan probabilitas kemunculan setiap nilai untuk atribut Jenis Kelamin (X1).

JENIS KELAMIN (X1)	JUMLAH KEJADIAN	
	Tepat Waktu	Terlambat
Laki-Laki	576	294
Perempuan	205	40
JUMLAH	**781**	**334**

3. Menentukan *Mean* dan Standar Deviasi untuk atribut IPS1 (X2), IPS2 (X3), IPS3 (X4), IPS4 (X5), IPK (X6), karena nilai untuk atribut berupa data diskrit, yang nantinya akan dimasukkan kedalam rumus Distribusi Gauss.

	IPS1 (X2)	
	Tepat Waktu	Terlambat
	3,666667	3,285714
	2,666667	0
	.....	.....
	2,904762	3
Mean	**3,069460907**	**2,233810369**
Stdv	**0,522886664**	**0,831114459**

	IPS2 (X3)	
	Tepat Waktu	Terlambat
	3,904762	0,7142857
	3,428571	0
	.....	.....
	2,285714	3,142857
Mean	**3,14080341**	**2,358325649**
Stdv	**0,477485491**	**0,766993131**

	IPS3 (X4)	
	Tepat Waktu	Terlambat
	3,181818	2,571429
	2,444444	1,153846
	.....	.....
	2,888889	2,45
Mean	**2,994895206**	**1,959281693**
Stdv	**0,534886209**	**0,813286468**

	IPS4 (X5)	
	Tepat Waktu	Terlambat
	3,1	2,25
	2,705882	1,833333
	.....	.....
	3,315789	2,555556
Mean	**2,977060345**	**1,698225088**
Stdv	**0,568959766**	**0,883503517**

	IPK (X6)	
	Tepat Waktu	Terlambat
	3,464286	2,166667
	2,831169	0,6271186
	.....	.....
	2,835443	2,85
Mean	**3,103901497**	**2,154534084**
Stdv	**0,384092914**	**0,531462498**

4. Menghitung Probabilitas dari setiap atribut
   a. Pengujian data testing mahasiswa ke-1 :

   X1 = "Laki - Laki", X3 = 2, X5 = 2,173913,
   X2 = 3, X4 = 1,047619, X6 = 2,214286

   P (X1 = "Laki - Laki" | Tepat Waktu) = 576/781 = 0,737516005

   P (X1 = "Laki – Laki"| Terlambat) = 294/334 = 0,880239520

   P (X2 = 3| Tepat Waktu)

   $$= \frac{1}{\sqrt{2\pi}\,(0{,}522886664)} e^{\frac{-(3-3{,}069460907)^2}{2(0{,}522886664)^2}} = 0{,}756258986518607 1$$

   P (X2 = 3| Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0,821114455)} e^{\frac{-(2-2,222210855)^2}{2(0,821114455)^2}} = 0,31388617128865404$$

P (X3 = 2 | Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0,477485451)} e^{\frac{-(2-2,840208431)^2}{2(0,477485451)^2}} = 0,048130909257542674$$

P (X3 = 2 | Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0,766598121)} e^{\frac{-(2-2,258225806)^2}{2(0,766598121)^2}} = 0,46663630964137604$$

P (X4 = 1,047619 | Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0,524888205)} e^{\frac{-(1,047619-2,594885206)^2}{2(0,524888205)^2}} = 9,87616404341 \times 10^{-7}$$

P (X4 = 1,047619 | Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0,813226488)} e^{\frac{-(1,047619-1,559221898)^2}{2(0,813226488)^2}} = 0,2617033308008$$

P (X5 = 2,173913 | Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0,568555766)} e^{\frac{-(2,173913-2,977080645)^2}{2(0,568555766)^2}} = 0,25890099899$$

P (X5 = 2,173913 | Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0,822502517)} e^{\frac{-(2,173913-1,858225082)^2}{2(0,822502517)^2}} = 0,3906193371335$$

P (X6 = 2,214286 | Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0,384092514)} e^{\frac{-(2,214286-2,105513497)^2}{2(0,384092514)^2}} = 0,0710525685$$

P (X6 = 2,214286 | Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0,521462458)} e^{\frac{-(2,214286-2,154524084)^2}{2(0,521462458)^2}} = 0,7459206464469$$

Sehingga:

***Likelihood* Tepat Waktu**

$$= 0,7375168005 \times 0,75825898518807 1 \times 0,048130909257542674 \times 9,87616404341 \times 10^{-7} \times 0,25890099899 \times 0,0710525685 \times 0,7044$$

$$= 3,4168976744141945 \times 10^{-7}$$

***Likelihood* Terlambat**

$$= 0,8802239520 \times 0,31388617128865404 \times 0,46663630964137604 \times 0,2617033308008 \times 0,3906193371335 \times 0,7459206464469 \times 0,29955$$

$$= 0,002942762778336564$$

Kesimpulan kelulusan = **Terlambat**,
karena nilai *Likelihood* Terlambat Lebih besar dari *Likelihood* Tepat Waktu

Gambar 3. Hasil Pengujian Data Testing Nim 142037

b. Pengujian data testing mahasiswa ke-1

X1 = "Perempuan", X4 = 2,285714,
X2 = 2,272727, X5 = 2,857143
X3 = 2,65, X6 = 2,587301

P (X1 = "Perempuan" | Tepat Waktu) = 205 / 781 = 0,262483994

P (X1 = "Perempuan"| Terlambat) = 40 / 334 = 0,119760479

P (X2 = 2,272727| Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}522886664)} e^{\frac{-(2{,}272727-2{,}089480907)^2}{2(0{,}522886664)^2}} = 0{,}2389711749794$$

P (X2 = 2,272727| Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}821114459)} e^{\frac{-(2{,}272727-2{,}285714285)^2}{2(0{,}821114459)^2}} = 0{,}4794826990530$$

P (X3 = 2,65 | Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}477485491)} e^{\frac{-(2{,}65-2{,}340280481)^2}{2(0{,}477485491)^2}} = 0{,}4926295622806 7$$

P (X3 = 2,65 | Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}766592121)} e^{\frac{-(2{,}65-2{,}852285645)^2}{2(0{,}766592121)^2}} = 0{,}4838556637577135$$

P (X4 = 2,285714| Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}524886209)} e^{\frac{-(2{,}285714-2{,}994285106)^2}{2(0{,}524886209)^2}} = 0{,}3096904086669$$

P (X4 = 2,285714| Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}812286468)} e^{\frac{-(2{,}285714-1{,}959231898)^2}{2(0{,}812286468)^2}} = 0{,}4525679618 23$$

P (X5 = 2,857143, | Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}668555766)} e^{\frac{-(2{,}857143-2{,}977060845)^2}{2(0{,}668555766)^2}} = 0{,}6857760797 4E97$$

P (X5 = 2,857143, | Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0{,}882502537)} e^{\frac{-(2{,}857143-1{,}698823542)^2}{2(0{,}882502537)^2}} = 0{,}1910163809 2668$$

P (X6 = 2,587301| Tepat Waktu)

$$= \frac{1}{\sqrt{2\pi}\,(0,284052514)} e^{\frac{-(2,587301-2,30651497)^2}{2(0,284052514)^2}} = 0,420595299$$

P (X6 = 2,587301| Terlambat)

$$= \frac{1}{\sqrt{2\pi}\,(0,521462458)} e^{\frac{-(2,587301-2,15454084)^2}{2(0,521462458)^2}} = 0,538830867$$

Sehingga:

***Likelihood* Tepat Waktu**

$= 0,262483594 \times 0,236971174574 \times 0,4926295622867 \times$

$= 0,3$[illegible] $\times 0,6$[illegible] $\times 0,42$[illegible] $\times 0,70044$

$= 0,0019324615231346462$

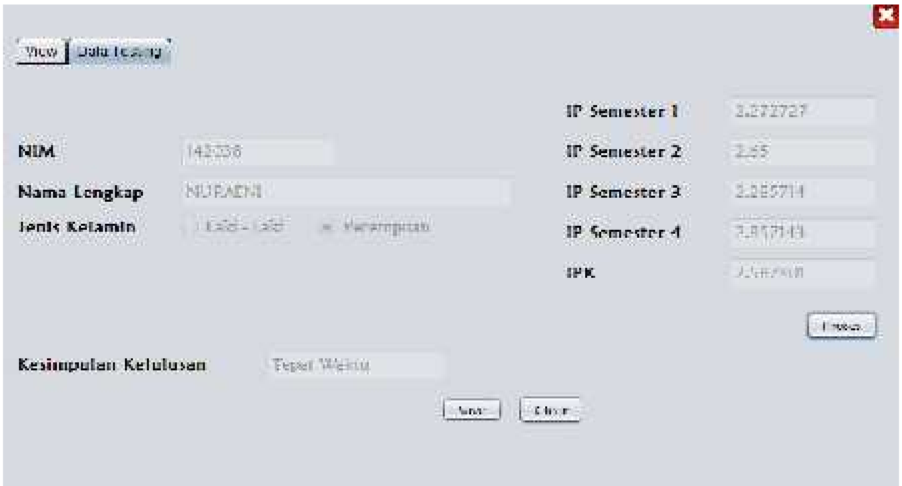
***Likelihood* Terlambat**

$= 0,115760475 \times 0,4754826$[illegible] $\times 0,4$[illegible]$577135 \times$

$= 0,45256756182$[illegible] $\times 0,151$[illegible] $\times 0,538830867 \times 0,29955$

$= 3,8768675256774254 \times 10^{-4}$

Kesimpulan kelulusan = **Tepat Waktu**, karena nilai Likelihood Tepat Waktu
Lebih besar dari Likelihood Tepat Waktu



Gambar 4. Hasil Pengujian Data Testing Nim 142038

3.5 Pengujian dan Akurasi Data

Dalam menguji akurasi dan ketepatan hasil pengklasifikasian sistem ini, digunakan 26 data alumni yang diambil secara acak. 26 data tersebut tidak terdapat di dalam data training, hal ini dimaksudkan agar hasil pengklasifikasian kelulusan dari sistem yang dirancang dapat dibandingkan dengan hasil kelulusan yang sesuai dengan data alumni mahasiswa STMIK Dipanegara Makassar.

Metode pengujian yang digunakan adalah *Confusion Matrix*. *Confusion matrix* adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Rumus ini melakukan perhitungan dengan 4 keluaran, yaitu: *recall, precision, acuraccy dan error rate.*

1. *Recall* adalah proporsi kasus positif yang diidentifikasi dengan benar. Rumus dari *recall* = D/(C+D)
2. *Precision* adalah proporsi kasus dengan hasil positif yang benar. Rumus dari *Precision* = D/(B+D)
3. *Accuracy* adalah perbandingan kasus yang diidentifikasi benar dengan jumlah semua kasus. Rumus dari *accuracy*= (A+D)/(A+B+C+D)
4. *Error Rate* adalah kasus yang diidentifikasi salah dengan sejumlah semua kasus.
   Rumus dari *Error Rate* = (B+C)/(A+B+C+D)

Keterangan:

A = jika hasil prediksi Terlambat dan data sebenarnya Terlambat.
B = jika hasil prediksi Tepat Waktu sedangkan nilai sebenarnya Terlambat.
C = jika hasil prediksi Terlambat sedangkan nilai sebenarnya Tepat Waktu.
D = jika hasil prediksi Tepat Waktu dan nilai sebenarnya Tepat Waktu.

A = 9, B = 1, C = 1, D = 15

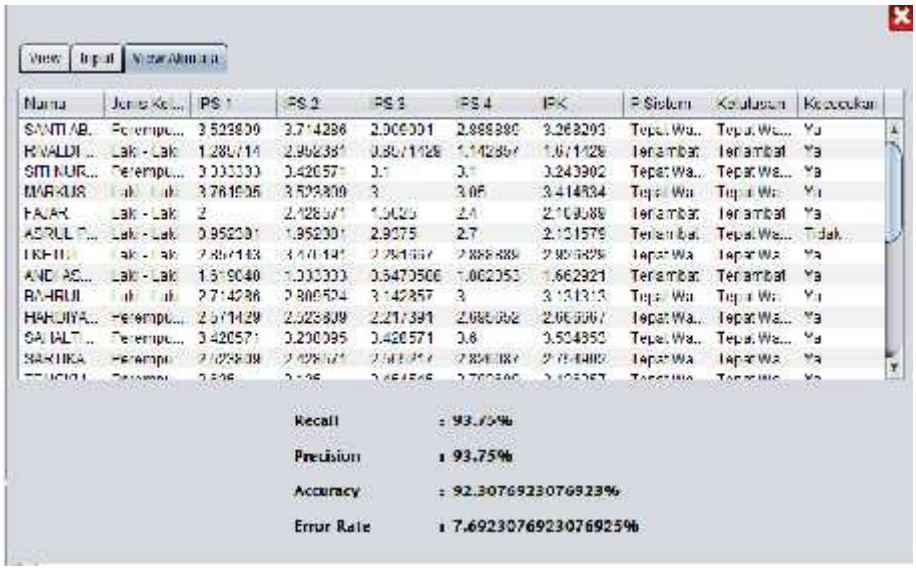
$$Recall = \frac{15}{(1+15)} = 0{,}9375$$

$$Precision = \frac{15}{(1+15)} = 0{,}9375$$

$$Accuracy = \frac{(9+15)}{(9+1+1+15)} = 0{,}923076923$$

$$Error\ Rate = \frac{(1+1)}{(9+1+1+15)} = 0{,}076923077$$

Hasil pengujian menunjukkan *accuracy* sebesar 92.3%. Detail perbandingannya dapat dilihat pada gambar 5. di bawah ini.



Gambar 5. Hasil Pengujian Akurasi

4. Kesimpulan

Berdasarkan hasil yang diperoleh pada penelitian ini, maka dapat disimpulkan bahwa :

1. Aplikasi data *mining* ini dapat mengklasifikasikan mahasiswa STMIK Dipanegara Makassar yang dapat lulus tepat waktu dan yang tidak dapat lulus tepat waktu dengan menggunakan Algoritma *Naive Bayes Classifier*.
2. Semakin banyak data *training* yang digunakan, maka kecerdasan sistem akan semakin baik, serta akan meningkatkan akurasi penelitian.
3. Berdasarkan hasil pengujian akurasi, bahwa ada faktor-faktor yang mempengaruhi kelulusan mahasiswa STMIK Dipanegara Makassar, bukan hanya dari faktor akdemik saja, tetapi faktor non-akademik juga mempengaruhi.

5. Saran

Dalam bahasan ini memuat saran untuk menutup kekurangan penelitian. Tidak memuat saran-saran selain untuk penelitian yang lebih lanjut.

Aplikasi ini hanya menggunakan satu metode klasifikasi data mining saja, yaitu menggunakan Algoritma Naive Bayes Classifier, jadi disarankan untuk pengembangan yang lebih lanjut bisa dibuat beberapa metode klasifikasi data mining yang lain, agar hasil dari beberapa metode tersebut dapat dibandingkan keakuratannya.

Daftar Pustaka

[1] Pramudiono, I, 2006, "*Apa Itu Data Mining?"*, Penerbit Andi, Yogyakarta.

[2] Eko Prasetyo, 2013, "*Data Mining Konsep dan Aplikasi Menggunakan Matlab*", Penerbit Andi, Yogyakarta.

[3] Susanto, Sani dan Suryadi, Dedy, 2010, "*Pengantar Data Mining Menggali dari Bongkahan Data*", Penerbit Andi, Yogyakarta.

[4] Suarga, 2012,"*Algoritma dan Pemrograman*", Penerbit Andi, Yogyakarta.

[5] David L. Olsen dan Dursun Delen, 2008, "Advanced Data Mining Techniques", Penerbit pringer, USA.