Volume 5 No. 2, Agustus 2021 P-ISSN 1907-1205
Universitas Banten Jaya E-ISSN 2622-6391

IMPLEMENTASI DATA MINING MENGGUNAKAN METODE NAIVE BAYES DENGAN FEATURE SELECTION UNTUK PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU

Sukarna Royan[1], Ansori Yulian[2], Syaechurodji[3]
[1]Teknik Informatika, Institut Teknologi Tangerang Selatan
[2]Sistem Informasi, Universitas Budi Luhur
[3]Sistem Informasi, Universitas Primagraha

Email: *royan@itts.ac.id, julianopec@gmail.com, syaechurodji44@gmail.com

ABSTRACT

The Education Efficiency Rate (AEE) is one of the parameters of the quality of the education program. The quality is measured based on 7 main standards, one of which is students and graduates. Meanwhile, to predict students' graduation rates accurately based on manually owned data set characteristics is very difficult. Data Mining by Naïve Bayes method was chosen to find patterns in analyzing and predicting timely graduation of students. As for the test will be done by comparing the initial dataset and dataset characteristics using the algorithm attribute selector Gain Ratio Attribute with the help of tools WEKA. The results showed that there was a difference to the accuracy of the results, and the larger ROC or AUC curves on the dataset characteristics using the selector attribute by using the Gain Ratio Attribute, although not very significant. And the result of this research yield 81% accuracy level with precision equal to 83.563% and recall 88.41%. The method used is included in Good Classification and will become the reference of the college management side, to address the problems that may arise in the decrease of the quality of education (e.g. decrease ratio of lecturers with students).

Keyword: *Data Mining, Graduation Prediction, Naïve Bayes*

PENDAHULUAN

Educational Data Mining (EDM) menjadi trend sejak tahun 2002. Perkembangan era informasi menyebabkan hal ini masuk hingga ke aspek pendidikan, hanya saja pemanfaatannya masih belum optimal (Thakar et al., 2015).

Ada banyak algoritma klasifikasi yang dapat digunakan seperti Algoritma *Advance Learning Analytics* (Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D. Lytras, Farhat Abbas, 2017), *Decision Tree* (Ogunde & Ajibade, 2014), *Naive Bayes* (Bagus et al., 2017), *Neural Network* (Tekin, 2014), dan *Support Vector Machine* (K & Aljahdali, 2013).

Fakultas Teknik Informatika di Universitas Serang Raya pada tahun 2011 memiliki jumlah mahasiswa sebanyak 526, sedangkan yang dapat lulus tepat waktu yaitu sebanyak 271 mahasiswa, sehingga diperoleh prosentase kelulusan sebesar 68.71% pada

periodenya. Pada tahun 2012 memiliki jumlah mahasiswa sebanyak 661, sedangkan yang dapat lulus tepat waktu yaitu sebanyak 345 mahasiswa, sehingga diperoleh prosentase kelulusan sebesar 73.93% pada periodenya. Pada tahun 2013 memiliki jumlah mahasiswa sebanyak 695, sedangkan yang dapat lulus tepat waktu yaitu sebanyak 483 mahasiswa, sehingga diperoleh prosentase kelulusan sebesar 69.58% pada periodenya. Berdasarkan prosentase diatas dapat diketahui nilai rata-rata dari prosentase kelulusan selama 3 tahun yaitu sebesar 70.86%.

Salah satu tantangan yang dihadapi perguruan tinggi adalah untuk memperbaiki kualitas program pendidikannya. Hal yang menjadi krusial adalah dalam menentukan strategi dan perencanaan agar kualitas program pendidikan dapat ditingkatkan. Keberhasilan atau kegagalan seorang mahasiswa untuk menyelesaikan studi pada waktunya dapat dilakukan dengan evaluasi untuk terus meningkatkan kualitas perguruan tinggi baik dari segi manajemen, kualitas pendidikan dan akreditasi. Angka Efisiensi Edukasi (AEE) merupakan salah satu parameter dari kualitas program pendidikan. Kualitas pendidikan program studi yang terdapat di Perguruan Tinggi Indonesia diukur dengan akreditasi yang dilakukan oleh BAN PT (Badan Akreditasi Nasional Perguruan Tinggi) (BADAN AKREDITASI NASIONAL PERGURUAN TINGGI, 2008). Terdapat 7 (tujuh) standar utama yang digunakan untuk mengukur kualitas pendidikan, mahasiswa dan lulusan adalah salah satunya.

Rendahnya nilai AEE berimplikasi terhadap penilaian akreditasi dari Perguruan Tinggi ataupun Program Studi. Oleh karena itu, diperlukan suatu tindakan untuk mengantisipasi masalah ini. Maka penelitian ini bertujuan untuk memprediksi kelulusan dari mahasiswa yang tepat waktu sebagai penunjang pengambilan keputusan sebagai bagian dari upaya untuk meningkatkan AEE dengan implementasi dari *data mining* (Tekin, 2014).

Sebagai upaya untuk meningkatkan persentase kelulusan mahasiswa tepat waktu adalah dengan menganalisis pola dalam database akademik, untuk memprediksi tingkat kelulusan tepat waktu yang sulit untuk dianalisis secara manual (Tahyudin et al., 2013).

METODE

1. Data Mining

   Penggabungan bidang ilmu basis data, *machine learning*, statistika, *information retrieval*, dsb menghasilkan suatu ilmu baru yaitu *Data Mining. Data Mining*

digunakan untuk menggali informasi tersembunyi dalam basis data yang merupakan bagian dari tahap proses *Knowledge Discovery in Database (KDD)*. (Moertini, 2002). *Data Mining* dapat dartikan juga sebagai proses dalam eksplorasi dan analisis data yang dapat dilakukan dengan banyak metode yang memiliki kegunaan masing-masing. Atau dapat juga diartikan sebagai proses mengekstraksi informasi dari set data besar melalui penggunaan algoritma dan teknik yang diambil dari bidang statistik dan Manajemen Sistem Database (Ogunde & Ajibade, 2014).

2. Naïve Bayes

   Algoritma Naive Bayes (NB) meupakan metode yang sederhana dalam klasifikasi berdasarkan teori probabilitas yang dikemukakan pertama kali oleh ilmuwan inggris bernama Thomas Bayes. Disebut naif karena menyederhanakan masalah yang bergantung pada dua asumsi penting (Osmanbegović & Suljić, 2012).

   Keuntungan dari kalsifikasi Naïve Bayes adalah bahwa algoritma ini tidak membutuhkan data pelatihan dalam jumlah yang besar dalam proses klasifikasi (Bhardwaj & Pal, 2011). Klasifikasi Naïve Bayes telah terbukti dapat diaplikasikan dalam situasi nyata dan kompleks. Naïve bayes dapat didefinisikan sebagai berikut:

$$P(x) = \frac{P(c)P(c)}{P(x)}$$

   - $P(c|x)$ adalah *posterior probability* dari *class* (*target*) terhadap *predictor* (*attribute*).
   - $P(x|c)$ adalah *likelihood* yang mana merupakan kemungkinan dari *predictor* terhadap *class*.
   - $P(c)$ adalah *prior probability* dari *class*.
   - $P(x)$ adalah *prior probability* dari *predictor*.

3. Gain Ratio Attribute

   Dalam tugas fitur klasifikasi pola memainkan peran yang sangat penting (Wang, 2012). Oleh karena itu, pemilihan fitur yang sesuai diperlukan karena sebagian besar data mentah mungkin berlebihan atau tidak relevan dengan pengenalan pola. Dalam beberapa kasus, pengklasifikasi tidak dapat berfungsi dengan baik karena banyaknya fitur redundan (Blum & Langley, 1997). Fitur yang berbeda memainkan peran yang berbeda dalam mengelompokkan dataset. Fitur yang tidak diinginkan akan menghasilkan informasi kesalahan selama klasifikasi yang akan mengurangi presisi

klasifikasi. Sebagian besar pemilihan fitur tradisional dapat menghilangkan gangguan ini untuk meningkatkan kinerja klasifikasi.

Modifikasi dari *Information Gain* adalah *Gain Ratio*, yang diperuntukan mengurangi bias atribut yang memiliki banyak cabang. Sifat dari *Gain ratio* ialah sebagai berikut:

a. Bila data menyebar rata, *Gain Ratio* akan memiliki nilai besar.
b. Nilai akan menjadi kecil apabila semua data masuk dalam satu cabang.

Gain ratio didefinisikan sebagai:

$$GainRatio\ (A) = \frac{Gain(A)}{SplitInfo(A)}$$

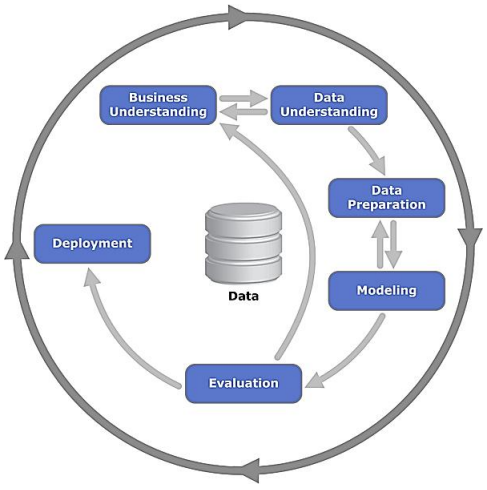
Dimana rumus *split info* seperti pada rumus diatas dengan *m* menyatakan banyaknya *split*. Jenis *split* yang dipilih adalah *split* yang memiliki nilai *Gain Ratio* yang terbesar. Bisa dikatakan nilai informasi split mewakili informasi potensial yang dihasilkan dengan membagi data pelatihan yang ditetapkan D ke v partisi, sesuai dengan hasil v pada atribut A.

$$SplitInfo\ \ _A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x log_2\left(\frac{|D_j|}{D}\right)$$

HASIL DAN PEMBAHASAN

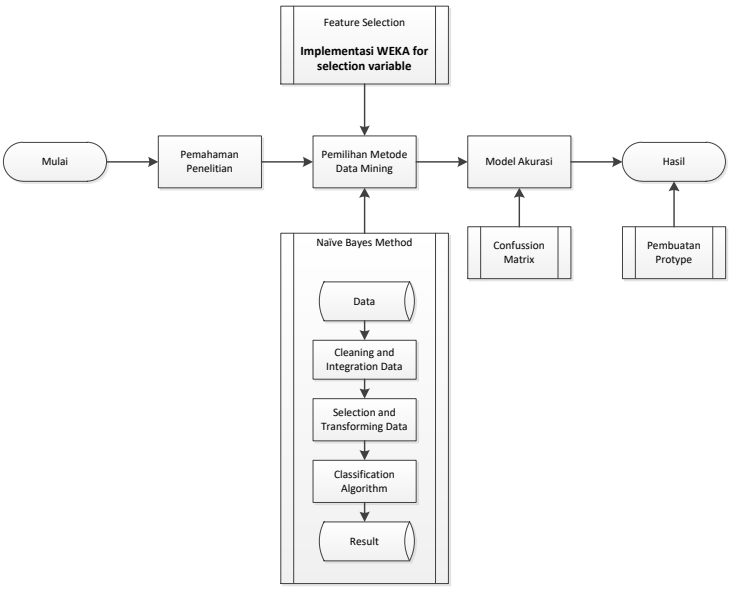
1. Langkah-langkah Penelitian

Langkah-langkah yang digunakan dalam penelitian ini penulis mengadopsi model CRISP-DM (*Cross Standard Industries Process for Data Mining*), dimana model ini terdapat 6 tahapan (Chapman et al., 2000), yaitu:



Gambar 1. Siklus CRISP-DM

Sedangkan, kerangka konsep dalam penelitian dapat digambarkan sebagai berikut :



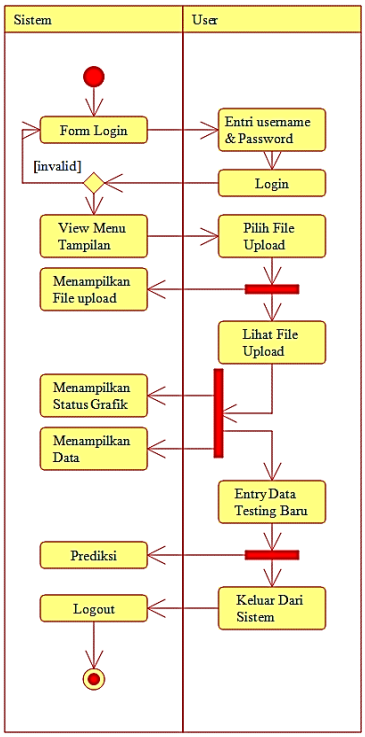
Gambar 2. Kerangka Konsep

Pola pikir yang tergambar diatas, dapat dijelaskan sebagai berikut:

a. Proses pemahaman penelitian ini merupakan fase dalam pengumpulan data yang mendukung untuk penelitian. Data yang diambil ialah data dari penerimaan mahasiswa baru, dan data profil akademik mahasiswa.
b. Proses klasifikasi, pada fase ini algoritma dari klasifikasi data mining dipilih dan diterapkan. Kemudian dikomparasi dengan algoirtma atribut selektor menggunakan bantuan dari WEKA *tools*.
c. Proses validasi, fase ini berfungsi untuk mengukur tingkat akurasi dari sebuah model yang telah dibuat. Kurva ROC (*Receiver Operating Characteristic*) akan digunakan untuk mengukur AUC (*Area Under Curve*) yang bersumber dari dari atribut dataset asli dengan atribut dataset yang berasal dari algoritma *attribute selector* oleh WEKA *tools*.
d. Pembuatan prototipe, fase ini membangun sebuah prototipe yang akan digunakan untuk memprediksi data kelulusan mahasiswa yang tepat waktu.

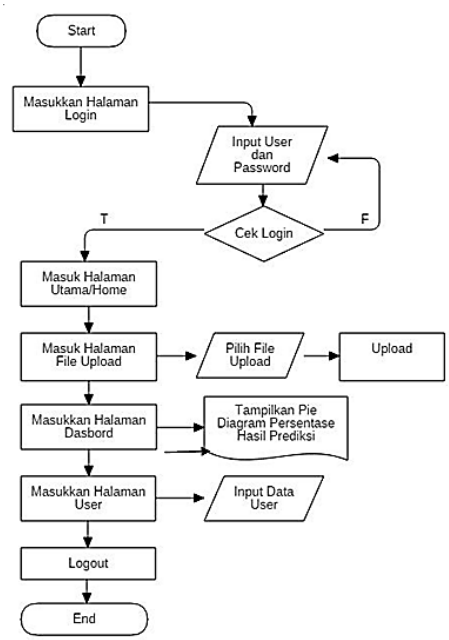
2. *Prototype* Sistem

Berikut adalah alur dari *Activity Diagram, Use Case, Class Diagram* dan *Flow Chart* yang diimplementasikan kedalam *prototype* prediksi:

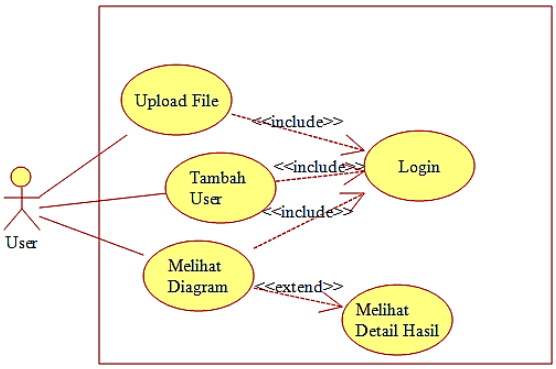


Gambar 3. *Activity Diagram*

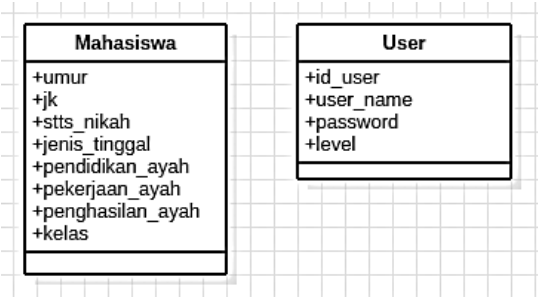
Pengguna membuka aplikasi dengan memasukkan login pada sistem dan akan menampilkan menu tampilan dan menampilkan pilihan untuk upload file yang dipilih user pada saat yang sama sistem juga dapat menampilkan data dari file yang di upload yang dipilih user, selanjutnya system akan import upload file dan memberikan summary dari data upload. User akan mengupload lagi pilihan file data testing dan menampilkan grafik prediksi dari data yang dipilih. Setelah selesai user akan keluar (*logout*) dari aplikasi.



Gambar 4. *Flowchart Prototype*



Gambar 5. *Use Case Prototype*



Gambar 6. *Class Diagram*

IMPLEMENTASI

Untuk menerapkan metode yang digunakan dibuat 2 (dua) buah dataset yang merupakan dataset awal dan dataset yang didapatkan dari proses *feature selection* menggunakan algoritma *Gain Ratio Attribute*. Data yang dipakai secara umum dibagi dalam 3 (tiga) jenis data, yaitu:

1. Data mahasiswa
2. Data orang tua
3. Data profil akademik

Dataset awal memiliki 14 (empat belas) buah atribut yang terdiri dari 13 (tiga belas) atribut *predictor* dan 1 atribut hasil. Masing-masing atribut dan nilai dapat dilihat dalam tabel dibawah ini:

Tabel 1. Atribut Dataset Awal

No	Atribut	Nilai	Nilai Baru
1	KOTA	Dalam Serang	Lokal
		Luar Serang	Non Lokal
2	UMUR	13-16	Remaja Awal
		17-25	Remaja Akhir
		26-35	Dewasa Awal
		36-45	Dewasa Akhir
3	JK	L	L
		P	P
4	PERNIKAHAN	Menikah	Menikah
		Belum Menikah	Belum Menikah
5	KELAS	R1	R1
		R2	R2
6	JENIS TINGGAL	Bersama Orang Tua	Non Kost
		Kost	Kost
7	TRANSPORTASI	Kendaraan Pribadi	Kendaraan Pribadi

		Angkutan Umum	Angkutan Umum
		Lainnya	Lainnya
8	PENDIDIKAN AYAH	Tidak Sekolah/SD	Dasar
		SMP/SMA	Menengah
		D1/D2/D3/D4/S1/S2/S3	Atas
9	PEKERJAAN AYAH	PNS	PNS
		Wiraswasta	Wiraswasta
		Lainnya	Lainnya
10	PENGHASILAN AYAH	0 - 1 jt	Rendah
		1jt - 5jt	Sedang
		5jt - 20jt	Tinggi
11	PENDIDIKAN IBU	Tidak Sekolah/SD	Dasar
		SMP/SMA	Menengah
		D1/D2/D3/D4/S1/S2/S3	Atas
12	PEKERJAAN IBU	PNS	PNS
		Wiraswasta	Wiraswasta
		Lainnya	Lainnya
13	PENGHASILAN IBU	0 - 1 jt	Rendah
		1jt - 5jt	Sedang
		5jt - 20jt	Tinggi
14	STATUS	LULUS	TEPAT
		AKTIF	TERLAMBAT

Pada pengujian data sample yang berisikan 104 data dengan 13 atribut predictor dan 1 atribut hasil, dengan menggunakan algoritma atribut selector Gain Ratio Attribute melalui tools WEKA diambil 8 atribut dengan ranking teratas, yaitu:

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 STATUS):
        Gain Ratio feature evaluator

Ranked attributes:
 0.3949991    4 PERNIKAHAN
 0.1948099    2 UMUR
 0.1317787    6 JENIS TINGGAL
 0.1207707    5 KELAS
 0.1093551    3 JK
 0.064027    10 PENGHASILAN AYAH
 0.0462722    8 PENDIDIKAN AYAH
 0.0431264    9 PEKERJAAN AYAH
 0.0264052   13 PENGHASILAN IBU
 0.0201508    7 TRANSPORTASI
 0.0179726   11 PENDIDIKAN IBU
 0.0079813   12 PEKERJAAN IBU
 0.0000257    1 KOTA
```

Gambar 7. Output Atribut *selector* dari *tools WEKA*

Atribut yang dipilih disajikan dalam bentuk tabel berikut ini:

Tabel 2. Atribut Hasil dari *Tools WEKA*

No	Ranking	Atribut
1	0.3949991	Pernikahan
2	0.1948099	Umur
3	0.1317787	Jenis Tinggal
4	0.1207707	Kelas
5	0.1093551	Jk
6	0.064027	Penghasilan Ayah
7	0.0462722	Pendidikan Ayah
8	0.0431264	Pekerjaan Ayah

Pengujian model dataset menggunakan *k-fold cross validation*, dengan k=10. Hasil dari pengujian metode yang telah dilakukan yaitu dengan algoritma *Naïve Bayes*, dilakukan pengujian tingkat akurasi dengan menggunakan bantuan *software Rapid Miner* untuk mencari *confusion matrix* dan kurva ROC/AUC (*Area Under Cover*). Setelah itu dilakukan pengujian hasil dari pemilihan atribut dengan menggunakan algoritma *Gain Ratio Attribute Evaluator* dari *tools WEKA*.

Tabel 3. *Confusion Matrix* Dataset Awal

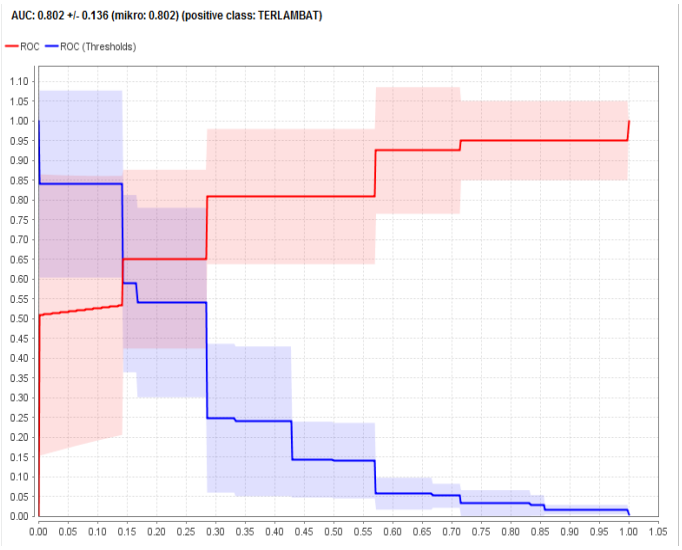
	true TEPAT	true TERLAMBAT	class precision

pred. TEPAT	61	13	82.43%
pred. TERLAMBAT	8	22	73.33%
class recall	88.41%	62.86%	

Dari pengujian model yang telah dilakukan dengan menggunakan *software rapid miner*, didapatkan *performance vector* dan kurva ROC sebagai berikut :

```
PerformanceVector:
accuracy: 80.00% +/- 12.07% (mikro: 79.81%)
ConfusionMatrix:
True:   TEPAT   TERLAMBAT
TEPAT:  61      13
TERLAMBAT:      8       22
precision: 74.17% +/- 22.50% (mikro: 73.33%) (positive class: TERLAMBAT)
ConfusionMatrix:
True:   TEPAT   TERLAMBAT
TEPAT:  61      13
TERLAMBAT:      8       22
recall: 62.50% +/- 20.50% (mikro: 62.86%) (positive class: TERLAMBAT)
ConfusionMatrix:
True:   TEPAT   TERLAMBAT
TEPAT:  61      13
TERLAMBAT:      8       22
AUC (optimistic): 0.804 +/- 0.136 (mikro: 0.804) (positive class: TERLAMBAT)
AUC: 0.802 +/- 0.136 (mikro: 0.802) (positive class: TERLAMBAT)
AUC (pessimistic): 0.800 +/- 0.137 (mikro: 0.800) (positive class: TERLAMBAT)
```

Gambar 8. *Performance Vector* Dataset Awal



Gambar 9. Kurva ROC Dataset Awal

Pengujian *confusion matrix* untuk atribut dataset dari *tools WEKA* yang diolah menggunakan algoritma *Naive Bayes* dengan 104 data Training dapat dilihat pada tabel dibawah ini:

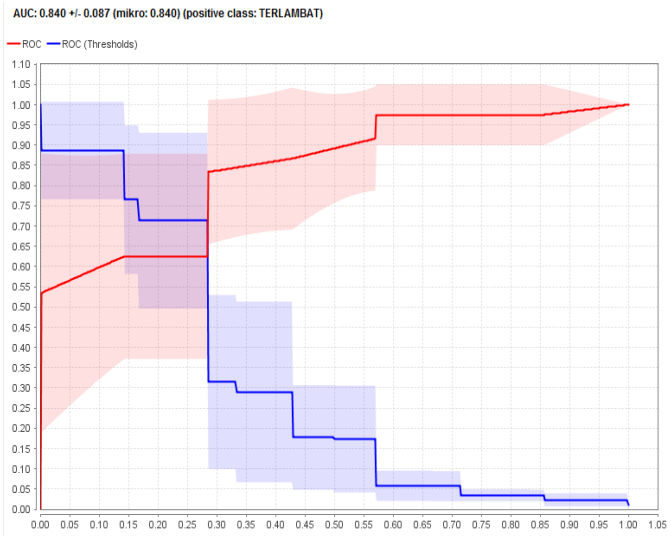
Tabel 4. *Confusion Matrix* Dataset *Feature Selection*

	true TEPAT	true TERLAMBAT	class precision
pred. TEPAT	61	12	83.56%
pred. TERLAMBAT	8	23	73.33%
class recall	88.41%	65.71%	

Dari pengujian model yang telah dilakukan dengan menggunakan *software rapid miner*, didapatkan *performance verctor* dan kurva ROC sebagai berikut:

```
PerformanceVector:
accuracy: 81.00% +/- 14.98% (mikro: 80.77%)
ConfusionMatrix:
True:   TEPAT   TERLAMBAT
TEPAT:  61      12
TERLAMBAT:      8       23
precision: 75.67% +/- 25.08% (mikro: 74.19%) (positive class: TERLAMBAT)
ConfusionMatrix:
True:   TEPAT   TERLAMBAT
TEPAT:  61      12
TERLAMBAT:      8       23
recall: 65.83% +/- 23.41% (mikro: 65.71%) (positive class: TERLAMBAT)
ConfusionMatrix:
True:   TEPAT   TERLAMBAT
TEPAT:  61      12
TERLAMBAT:      8       23
AUC (optimistic): 0.855 +/- 0.081 (mikro: 0.855) (positive class: TERLAMBAT)
AUC: 0.840 +/- 0.087 (mikro: 0.840) (positive class: TERLAMBAT)
AUC (pessimistic): 0.826 +/- 0.095 (mikro: 0.826) (positive class: TERLAMBAT)
```

Gambar 10. *Performance Vector* Dataset *Feature Selection*



Gambar 11. Kurva ROC Dataset *Feature Selection*

Dataset yang digunakan memiliki 104 buah *instance* dengan model awal 14 atribut mendapatkan nilai *precision* sebesar 82.43%, *recall* 88.41%, *accuracy* 80.00%, dan nilai ROC sebesar 0.802. Sedangkan model dengan dataset yang menggunakan atribut hasil dari *feature selection Gain Ratio* mendapatkan 9 buah atribut dengan 1 atribut *predictor*. Adapun nilai precision yang diraih sebesar 83.56%, *recall* 88.41%, *accuracy* 81.00%, dan nilai ROC sebesar 0.840. Sehingga model dataset yang dibuat tergolong kedalam klasifikasi yang baik (*good classification*).

KESIMPULAN

Dalam paper ini, metode yang diusulkan yaitu menggunakan *feature selection* dengan algoritma *Gain Ratio Attribute* mendapatkan hasil yang baik dan mengalami peningkatan akurasi. Pada penelitian selanjutnya sangat dimungkinkan untuk mengimplementasikan *feature selection* kedalam algoritma klasifikasi yang lain, seperti KNN, C45, dsb.

DAFTAR PUSTAKA

Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D. Lytras, Farhat Abbas, J. S. A. (2017). Predicting student performance using Advanced Learning Analytics. *Pakistan Saudi Arab2018-01-11*, *C*, 415–421. https://doi.org/10.1145/3041021.3054164

BADAN AKREDITASI NASIONAL PERGURUAN TINGGI. (2008). *Akreditasi Program Studi Sarjana Buku VI Pedoman Penilaian Akreditasi Program Studi Sarjana* (BAN-PT (ed.)). BAN-PT. https://banpt.or.id/download_instrumen

Bagus, I., Peling, A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. *International Journal of Engineering and Emerging Technology*, *2*(1), 53–57.

Bhardwaj, B. K., & Pal, S. (2011). Data Mining : A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security*, *9*(4), 136–140.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1–2), 245–271.

https://doi.org/10.1016/S0004-3702(97)00063-5

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth,

R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76. https://doi.org/10.1109/ICETET.2008.239

K, A. A., & Aljahdali, S. (2013). Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques. *International Journal of Computer Applications*, *69*(11), 12–16.

Moertini, V. S. (2002). Data Mining Sebagai Solusi Bisnis. *Integral*, *7*(1), 44–56.

Ogunde, & Ajibade. (2014). A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm. *Computer Science and Information Technology*, *2*(1), 21–46. http://www.ejer.com.tr/index.php?git=22&kategori=103&makale=925

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. *Journal of Economics and Business*, *X*(1), 3–12.

Tahyudin, I., Utami, E., & Amborowati, A. (2013). Comparing Clasification Algorithm Of Data Mining to Predict the Graduation Students on Time. *Information Systems International Conference (ISICO), December*, 2–4.

Tekin, A. (2014). Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach. *Eurasian Journal of Educational Research*, *14*(54), 207–226. https://doi.org/10.14689/ejer.2014.54.12

Thakar, P., Mehta, A., & Manisha. (2015). Performance analysis and prediction in educational data mining: A research travelogue. *International Journal of Computer Applications*, *110*(15), 60–68.

Wang, H. (2012). An Empirical Study on the Stability of Feature Selection for Imbalanced Software Engineering Data. *International Journal of Advanced Computer Research*, *2*(3), 1–5. https://doi.org/10.1109/ICMLA.2012.60