

Prediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma C4.5 Pada Fakultas Dakwah Dan Ilmu Komunikasi UIN Raden Intan Lampung

Fiqih Satria¹, Zamhariri², M. Apun Syaripudin³

Dosen Universitas Islam Negeri Raden Intan Lampung^{1,2,3}

Jl. Letkol H. Endro Suratmin Sukarame, Bandar Lampung

Sur-el : fiqihatria@radenintan.ac.id¹, zamhariri@radenintan.ac.id²,
apunsyaripudin@radenintan.ac.id³

Abstract : Student graduation data is very important for universities because it is used in the accreditation process. Data continues to grow and is ignored because it is rarely used. Data of graduating students can provide useful information if processed optimally. This study processes data using data mining to obtain information in the form of a prediction of student graduation punctuality. The method used is the C4.5 algorithm. The criteria used are gender, regional origin, type of school origin, ranking and entry point. In its application, the C4.5 algorithm can be used in predicting student graduation times with a precision value of 70.70%, 60.4% recall, and 58.2% accuracy. In measuring the performance of the algorithm in pattern recognition or information retrieval it is recommended to use a minimum of two parameters namely precision and recall to detect bias, therefore in this study the F-Measure calculation is used. From the calculation of the F-Measure obtained a value of 71% which means that the C4.5 algorithm is considered good in classifying and predicting students who graduate on time.

Keywords: C4.5 algorithm, student data, data mining, decision trees.

Abstrak : Data kelulusan mahasiswa sangat penting bagi perguruan tinggi karena digunakan dalam proses akreditasi. Data terus bertambah dan terabaikan karena jarang digunakan. Data mahasiswa yang lulus dapat memberikan informasi yang bermanfaat jika diolah secara maksimal. Penelitian ini mengolah data menggunakan data mining untuk mendapatkan informasi berupa prediksi ketepatan waktu lulus mahasiswa. Metode yang digunakan adalah algoritma C4.5. Kriteria yang digunakan adalah jenis kelamin, asal daerah, jenis asal sekolah, peringkat dan jalur masuk. Dalam penerapannya, algoritma C4.5 dapat digunakan dalam memprediksi waktu lulus mahasiswa dengan nilai precision 70.70%, recall 60.4%, dan akurasi 58.2%. Dalam mengukur kinerja algoritma dalam pengenalan pola atau temu kembali informasi disarankan menggunakan minimal dua parameter yaitu precision dan recall untuk mendeteksi bias, oleh karena itu pada penelitian ini digunakan perhitungan F-Measure. Dari perhitungan F-Measure didapatkan nilai 71% yang memiliki arti bahwa algoritma C4.5 dinilai baik dalam mengklasifikasi dan melakukan prediksi terhadap mahasiswa yang lulus tepat waktu.

Kata kunci: Algoritma C4.5, data mahasiswa, data mining, pohon keputusan.

1. PENDAHULUAN

Persentase kelulusan mahasiswa tepat waktu merupakan salah satu instrumen penting di dalam penilaian akreditasi perguruan tinggi. Elemen penilaian pada lulusan perguruan tinggi salah satunya adalah perguruan tinggi memiliki angka efisien edukasi yang ideal dalam arti jumlah mahasiswa yang diterima memiliki

keseimbangan dengan jumlah mahasiswa yang lulus tepat waktu. Dengan persentase mahasiswa lulus tepat waktu yang besar akan meningkatkan penilaian akreditasi. Pada sisi lain, mahasiswa yang lulus tepat waktu tidak akan terbebani dengan biaya Uang Kuliah Tunggal (UKT) yang harus tetap dibayarkan selama mahasiswa belum dinyatakan lulus ujian skripsi. Namun masa studi mahasiswa belum dapat dideteksi secara

dini, sehingga dapat berdampak pada terlambatnya kelulusan mahasiswa.

Penelitian dengan algoritma *Naive Bayes Classifier* dilakukan oleh Mujib Ridwan dkk untuk evaluasi kinerja akademik mahasiswa.[1] Dalam penelitian tersebut mengungkapkan bahwa metode *Naive Bayes* berhasil melakukan prediksi dengan akurat, Pengujian pada data mahasiswa angkatan 2005 - 2009, algoritma *Naive Bayes Classifier* menghasilkan nilai *precision*, *recall*, dan *accuracy* masing-masing 83%, 50%, dan 70%.

Penelitian yang lain dilakukan oleh (Sri Anisa, 2018) yang berjudul *Klasifikasi Kelulusan Tepat Waktu Mahasiswa Informatika UMRAH Dengan Menggunakan Metode Classification And Regression Tree* (Studi Kasus: Mahasiswa Fakultas Teknik Jurusan Informatika Universitas Maritim Raja Ali Haji Tanjung pinang). Dari hasil penelitian tersebut menunjukkan bahwa tingkat keakuratan menggunakan metode *Classification and Regression Tree* dalam perhitungan ketepatan waktu kelulusan mahasiswa berdasarkan variabel yang digunakan sebesar 81,42%.[2]

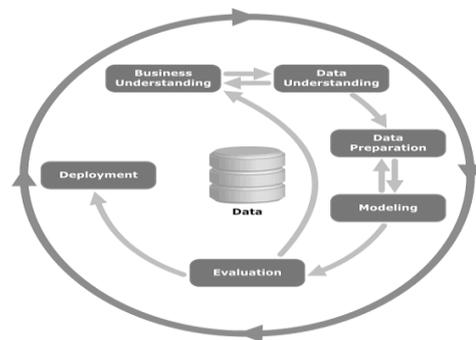
Penelitian ini menggunakan algoritma yang berbeda dari penelitian sebelumnya yaitu menggunakan algoritma C4.5 yang diharapkan dapat memberikan performa yang baik sesuai dengan acuan penilaian performansi algoritma.

Tujuan dari penelitian ini yaitu melakukan analisis ketepatan waktu lulus mahasiswa menggunakan teknik data mining sebagai dasar pertimbangan dalam mengambil keputusan lebih lanjut mengenai faktor-faktor apa saja yang mempengaruhi waktu kelulusan atau masa studi

mahasiswa menggunakan data pokok mahasiswa yang ada di dalam *database* Sistem Informasi Akademik. Metode *data mining* yang dapat digunakan untuk melakukan prediksi kelulusan mahasiswa di antaranya adalah metode klasifikasi. Salah satu metode klasifikasi yang akan digunakan dalam penelitian ini yaitu Algoritma C4.5.

2. METODOLOGI PENELITIAN

Data yang digunakan dalam penelitian ini, adalah data mahasiswa angkatan 2014 Fakultas Dakwah dan Ilmu Komunikasi UIN Raden Intan Lampung yang berjumlah 280 mahasiswa. Metode penelitian yang digunakan sesuai dengan tahapan-tahapan Model *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai berikut:



Gambar 1. Fase CRISP-DM model [3]

Model CRISP-DM terbagi dalam enam tahapan yaitu:

1. *Bussiness Understanding* (Pemahaman Bisnis)

Pemahaman bisnis Pada tahapan pemahaman bisnis ada beberapa hal yang dilakukan antara lain; menentukan tujuan bisnis,

menilai situasi, dan menentukan tujuan data mining.

2. *Data Understanding* (Pemahaman Data)

Pemahaman data Pada tahapan pemahaman data, ada beberapa hal yang dilakukan antara lain, mengumpulkan data awal, mendeskripsikan data, mengeksplorasi data, dan memverifikasi kualitas data.

3. *Data Preparation* (Pengolahan Data)

Persiapan Data Pada tahapan persiapan data, ada beberapa hal yang dilakukan antara lain, deskripsi data set, memilih data, membangun data, mengintegrasikan data dan membersihkan data.

4. *Modelling* (Pemodelan)

Pemodelan Pada penelitian merupakan tahapan yang secara langsung melibatkan data mining. Pemilihan teknik data mining, algoritma dan menentukan parameter dengan nilai yang optimal. Pada tahapan pemodelan, ada beberapa hal yang dilakukan antara lain, memilih teknik pemodelan, membangun model, dan menilai model. Tahapan yang dilakukan mengacu pada penelitian tersebut.

5. *Evaluation* (Evaluasi)

Tahapan evaluasi adalah fase interpretasi terhadap hasil data mining. Evaluasi dilakukan secara mendalam dengan tujuan agar hasil pada tahap modelling sesuai dengan sasaran yang ingin dicapai dalam tahap business understanding. Pada tahapan evaluasi, ada beberapa hal yang dilakukan antara lain:

- a. Mengevaluasi hasil (*Evaluate Results*)
- b. Meninjau Proses (*Review Process*)
- c. Menentukan Langkah selanjutnya (*Determine Next Steps*)

6. *Deployment* (Penyebaran)

Menggunakan model yang dihasilkan dan dipresentasikan atau proses *knowledge presentation*.

3. HASIL DAN PEMBAHASAN

Langkah-langkah penelitian ini sesuai alur pada Model CRISP-DM (*Cross Industry Standard Process for Data Mining*):

1. *Pemahaman Bisnis (Business Understanding)*

Institusi Perguruan tinggi saat ini dituntut untuk memiliki keunggulan dan memiliki daya saing yaitu dengan meningkatkan mutu dan akreditasi. Pencapaian akreditasi yang baik, diantaranya dipengaruhi oleh jumlah kelulusan mahasiswa per tahun. Banyak mahasiswa yang tingkat kelulusannya berbeda, tepat waktu dan tidak tepat waktu. Mahasiswa yang lulus tidak tepat waktu menjadi catatan penting untuk program studi, saat penilaian akreditasi.

2. *Pemahaman Data (Data Understanding)*

Dataset yang digunakan pada penelitian ini dikumpulkan dari Sistem Informasi Akademik (SIKAD) UIN Raden Intan Lampung. Data yang digunakan adalah data mahasiswa angkatan 2014 Fakultas Dakwah dan Ilmu Komunikasi yang terdiri dari 4 Program Studi yaitu Komunikasi dan Penyiaran Islam (KPI), Pengembangan Masyarakat Islam (PMI), Manajemen Dakwah (MD) dan Bimbingan dan Konseling Islam (BKI). Tidak semua atribut dari data mahasiswa digunakan,

dalam penelitian ini hanya menggunakan 5 atribut yang terdiri dari Jenis Kelamin, Asal Daerah, Jenis Asal Sekolah, Peringkat dan Jalur Masuk.

Tabel 1. Pemilihan Atribut Data

No	Atribut	Keterangan
1	NPM	×
2	Nama	×
3	Judul Skripsi	×
4	IPK	×
5	No. HP	×
6	Tahun Masuk	×
7	Program Studi	×
8	Tempat Tanggal Lahir	×
9	Jenis Kelamin	✓
10	No. KTP	×
11	Pekerjaan	×
12	Perkawinan	×
13	Asal Daerah	✓
14	Jenis Asal Sekolah	✓
15	Nama Sekolah	×
16	Peringkat	✓
17	Jalur Masuk	✓
18	Nama Bapak	×
19	Nama Ibu	×
20	Nama Wali	×
21	Gaji Bapak	×
22	Gaji Ibu	×
23	Kelas	×

Keterangan :

✓ : Atribut yang dipilih

× : Atribut yang tidak dipilih

Sumber : Data Alumni FDIK UIN Raden Intan Lampung Tahun Angkatan 2014

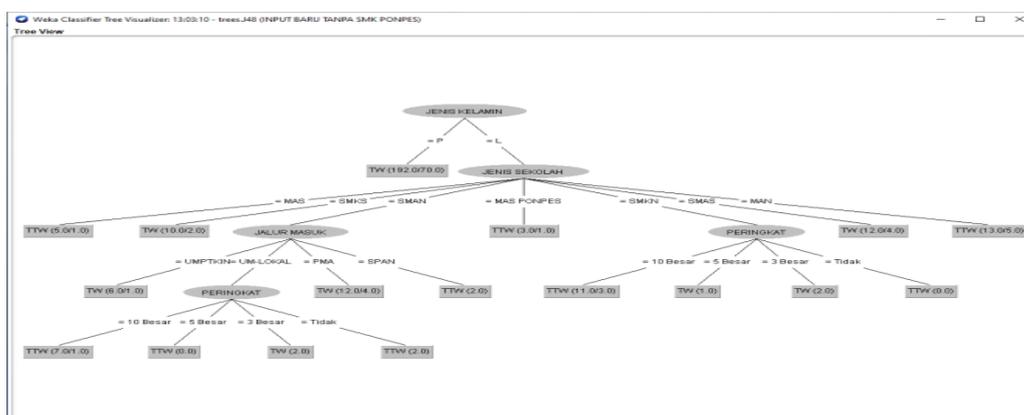
3. Pengolahan Data (*Data Preparation*)

Seluruh data yang berhasil dikumpulkan

sebanyak 281 data. Data terlebih dahulu dilakukan pra processing data. Pada tahapan ini data dibersihkan. Atribut yang dianggap tidak penting untuk digunakan dalam proses analisis (contoh: atribut NPM, Nama Mahasiswa), maka atribut tersebut akan dihilangkan. Pada tahap ini adalah pemilihan atribut data yang akan digunakan untuk pemodelan. Tabel 1. menjelaskan mengenai atribut-atribut yang akan digunakan dalam analisis. Berikut adalah rincian dari atribut yang digunakan beserta nilai (*value*) yang digunakan:

Tabel 2. Atribut dan Value

No	Atribut	Value
1	Jenis Kelamin	Laki-laki Perempuan
2	Asal Daerah	Dalam Kota Luar Kota
3	Jenis Asal Sekolah	MAN MAS MAS PONPES SMAN SMAS SMKN SMKS
4	Peringkat	Tiga Besar Lima Besar Sepuluh Besar Tidak termasuk
5	Jalur Masuk	SPAN UMPTKIN UM-LOKAL PMA



Gambar 2. Pohon Keputusan

4. Pemodelan (*Modelling*)

Metode yang digunakan dalam penelitian ini adalah Algoritma C4.5 yang dapat menghasilkan output berupa decision tree (pohon keputusan). Salah satu kelas klasifikasi pada aplikasi data mining WEKA yang mengimplementasikan algoritma C4.5 adalah Algoritma J48 [4].

Algoritma C4.5 akan menghasilkan output sebuah pohon keputusan dan confusion matrix. Pohon keputusan diubah dalam bentuk rule atau ditulis dalam bentuk notasi *frist order logic*. Konversi dilakukan dengan melihat *leaf* yang bernilai hasil dari target atributnya. Lakukan penelusuran melalui simpul *root*, sampai ditemukannya *leaf node* yang berisi target atribut [5].

Sedangkan *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya [6]. Berikut pohon keputusan yang dihasilkan dari aplikasi WEKA dapat dilihat pada Gambar 2.

Dari pohon keputusan yang dihasilkan oleh aplikasi WEKA di atas maka dapat ditarik suatu rule (aturan) untuk memudahkan dalam analisis pohon keputusan yaitu menentukan mana yang menjadi root sampai ditemukannya leaf node yang berisi target atribut. Rule yang dihasilkan dari pohon keputusan adalah sebagai berikut :

```
JENIS KELAMIN = P: TW
JENIS KELAMIN = L
|   JENIS SEKOLAH = MAS: TTW
```

```
|   JENIS SEKOLAH = SMKS: TW
|   JENIS SEKOLAH = SMAN
|   |   JALUR MASUK = UMPTKIN: TW
|   |   JALUR MASUK = UM-LOKAL
|   |   |   PERINGKAT = 10 Besar:
TTW
|   |   |   PERINGKAT = 5 Besar: TTW
|   |   |   PERINGKAT = 3 Besar: TW
|   |   |   PERINGKAT = Tidak: TTW
|   |   JALUR MASUK = PMA: TW
|   |   JALUR MASUK = SPAN: TTW
|   JENIS SEKOLAH = MAS PONPES: TTW
|   JENIS SEKOLAH = SMKN
|   |   PERINGKAT = 10 Besar: TTW
|   |   PERINGKAT = 5 Besar: TW
|   |   PERINGKAT = 3 Besar: TW
|   |   PERINGKAT = Tidak: TTW
|   JENIS SEKOLAH = SMAS: TW
|   JENIS SEKOLAH = MAN: TTW
```

Keterangan :

TW : Tepat Waktu
TTW : Tidak tepat Waktu

5. Evaluasi (*Evaluation*)

Selain pohon keputusan, aplikasi WEKA juga menghasilkan confusion matrix seperti pada tabel 3.

Tabel 3. Confusion Matrix

True Positive (TP)	False Positive (FP)
145	22
False Negative (FN)	True Negative (TN)
95	18

Dari hasil *confusion matrix* di atas, dapat dijelaskan sebagai berikut:

- **True Positive (TP):** kasus dimana 145 mahasiswa diprediksi lulus tepat waktu, memang benar (True) lulus tepat waktu.
- **True Negative (TN):** kasus dimana 18 mahasiswa diprediksi lulus tidak tepat waktu dan sebenarnya mahasiswa tersebut memang (True) lulus tidak tepat waktu.

- **False Positive (FP):** kasus dimana 22 mahasiswa yang diprediksi lulus tepat waktu, ternyata lulus tidak tepat waktu. Prediksinya salah (False).
- **False Negatif (FN):** kasus dimana 95 mahasiswa yang diprediksi lulus tidak tepat waktu, tetapi ternyata sebenarnya (True) lulus tepat waktu. Prediksinya salah (False).

a. Pengukuran Kinerja Algoritma C4.5

Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data.

Berdasarkan nilai *True Negative* (TN), *False Positive* (FP), *False Negative* (FN), dan *True Positive* (TP) dapat diperoleh nilai akurasi, presisi dan *recall*. Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar.

Nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data. Nilai akurasi dapat diperoleh dengan Persamaan 1.

Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif. Presisi dapat diperoleh dengan Persamaan 2.

Recall menunjukkan berapa persen data kategori positif yang terklasifikasikan dengan benar oleh sistem. Nilai *recall* diperoleh dengan Persamaan 3.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

$$\text{Presisi} = \frac{TP}{FP+TP} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{TP}{FN+TP} \times 100\% \quad (3)$$

b. Perhitungan Akurasi, Presisi dan Recall

$$\begin{aligned} \text{Akurasi} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \\ &= \frac{145 + 18}{145 + 18 + 22 + 95} \times 100\% \\ &= \frac{163}{280} \times 100\% \\ &= 58,2\% \end{aligned}$$

$$\begin{aligned} \text{Presisi} &= \frac{TP}{FP + TP} \times 100\% \\ &= \frac{145}{22 + 145} \times 100\% \\ &= \frac{145}{167} \times 100\% \\ &= 86,8\% \end{aligned}$$

$$\begin{aligned} \text{Recall} &= \frac{TP}{FN + TP} \times 100\% \\ &= \frac{145}{95 + 145} \times 100\% \\ &= \frac{145}{240} \times 100\% \\ &= 60,4\% \end{aligned}$$

Dengan perhitungan di atas algoritma C4.5 menghasilkan akurasi, presisi dan recall masing-masing 58,2%, 86,8% dan 60,4%.

c. Acuan performansi algoritma

Akurasi sangat bagus digunakan sebagai acuan performansi algoritma JIKA dataset memiliki jumlah data *False Negative* (FN) dan *False Positive* (FP) yang sangat mendekati. Namun jika jumlahnya tidak mendekati, maka gunakan *F-Measure (F1 Score)* sebagai acuan.

Menggunakan presisi atau akurasi saja dalam pengukuran kinerja dari sebuah algoritma/metode dapat menimbulkan bias yang sangat fatal. Sebagai contoh pengujian menggunakan 100 ikan salmon dan 900 ikan lain, ternyata mesin hanya memisahkan 1 ikan salmon dan setelah dicek oleh manusia, 1 ikan

tersebut benar merupakan ikan salmon. Pengujian ini dapat dihitung sebagai berikut:

Tabel 4. Contoh hasil Pengujian

True Positive (TP) 1	False Positive (FP) 0
False Negative (FN) 99	True Negative (TN) 900

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \\
 &= \frac{1 + 900}{1 + 900 + 0 + 99} \times 100\% \\
 &= \frac{901}{1000} \times 100\% \\
 &= 90,1\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Presisi} &= \frac{TP}{FP + TP} \times 100\% \\
 &= \frac{1}{0 + 1} \times 100\% \\
 &= \frac{1}{1} \times 100\% \\
 &= 100\%
 \end{aligned}$$

$$\begin{aligned}
 \text{Recall} &= \frac{TP}{FN + TP} \times 100\% \\
 &= \frac{1}{99 + 1} \times 100\% \\
 &= \frac{1}{100} \times 100\% \\
 &= 1\%
 \end{aligned}$$

Dari perhitungan di atas didapatkan precision sebesar 100% dan accuracy sebesar 90,1 %. Sekilas tampak baik, namun perhatikan nilai recall yang hanya 1%. Hal ini menunjukkan bahwa sistem hanya dapat memisahkan ikan salmon dalam jumlah yang sedikit sekali dan masih banyak ikan-ikan salmon yang lolos dari pemisahan. Jadi dalam mengukur kinerja dari sebuah algoritma/metode dalam pengenalan pola atau temu kembali informasi disarankan menggunakan minimal dua parameter yaitu precision dan recall untuk mendeteksi bias pada contoh kasus di atas. Oleh karena itu pada

penelitian ini digunakan perhitungan F-Measure dengan persamaan sebagai berikut :

$$\begin{aligned}
 \text{F - Measure} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
 &\quad \times 100\% \\
 &= 2 \times \frac{0,868 \times 0,604}{0,868 + 0,604} \times 100\% \\
 &= 2 \times \frac{0,524}{1,472} \times 100\% \\
 &= 2 \times \frac{0,524}{1,472} \times 100\% \\
 &= 71,1\%
 \end{aligned}$$

Dari perhitungan *F-Measure* didapatkan nilai 71 % yang memiliki arti bahwa algoritma C.45 dinilai baik dalam mengklasifikasi dan melakukan prediksi terhadap mahasiswa yang lulus tepat waktu.

4. KESIMPULAN

Data mining dapat digunakan salah satunya untuk memprediksi ketepatan waktu lulus mahasiswa. Pada penelitian ini, *software WEKA* dengan menggunakan algoritma C4.5 menjadikan jenis kelamin root pada pohon keputusan, kemudian dilanjutkan dengan atribut jenis sekolah, jalur masuk dan peringkat.

Dari pengukuran kinerja algoritma yang telah dilakukan, dapat disimpulkan dalam kasus ini algoritma C4.5 memiliki nilai akurasi, presisi dan recall masing-masing 58,2%, 86,8% dan 60,4%.

Nilai akurasi yang kecil tidak membuat algoritma dinilai tidak baik. Akurasi sangat bagus digunakan sebagai acuan performansi

algoritma JIKA *dataset* memiliki jumlah data *False Negative* (FN) dan *False Positive* (FP) yang sangat mendekati. Namun jika jumlahnya tidak mendekati, maka menggunakan *F-Measure* (*F1 Score*) sebagai acuan karena menggunakan *precision* atau *accuracy* saja dalam pengukuran kinerja dari sebuah sistem/metode dapat menimbulkan bias yang sangat fatal.

Dari perhitungan *F-Measure* didapatkan nilai 71 % yang memiliki arti bahwa algoritma C.45 dinilai baik dalam mengklasifikasi dan melakukan prediksi terhadap mahasiswa yang lulus tepat waktu.

Faktor yang menentukan kelulusan mahasiswa tidak hanya dilihat dari data diri mahasiswa saja, masih banyak faktor lain seperti faktor akademis, faktor keluarga, faktor psikologis dan faktor situasi akademik. Jadi penelitian selanjutnya dapat dikembangkan lagi dengan menambahkan atribut lain dan menggunakan algoritma yang lain untuk mendapatkan hasil yang lebih baik lagi dalam memprediksi waktu kelulusan mahasiswa tepat waktu.

DAFTAR PUSTAKA

- [1] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," Vol. 7, No. 1, pp. 59–64, 2013.
- [2] Sri Anisa, "Klasifikasi Kelulusan Tepat Waktu Mahasiswa Informatika UMRAH Dengan Menggunakan Metode Classification And Regression Tree (Studi Kasus : Mahasiswa Fakultas Teknik Jurusan Informatika Universitas Maritim Raja Ali Haji Tanjungpinang)," *Repository Universitas Maritim Raja Ali Haji*. [Online]. Available:

- <http://repository.umrah.ac.id/1601/>.
[Accessed: 02-Feb-2020].
- [3] P. C. Ncr et al., "Crisp-dm 1.0," 2000.
 - [4] Y. N. Kunang et al., "Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang," Vol. 2013, No. Semnasif, pp. 1–8, 2013.
 - [5] Suyanto, "Artificial Intelligence (Searching, Reasoning, Planning dan Learning)", Informatika Bandung, 2007
 - [6] E. Prasetyo, *Data Mining: Konsep dan Aplikasi menggunakan Matlab*, 1 ed. Yogyakarta: Andi Offset, 2012.