

Perbandingan Kinerja Algoritma C4.5 dan *Naive Bayes* untuk Klasifikasi Penerima Beasiswa

Choirul Anam¹⁾, Harry Budi Santoso²⁾

^{1,2)}Magister Teknik Informatika, Universitas AMIKOM Yogyakarta
Jl. Ring Road Utara Condong Catur, Sleman, Yogyakarta 55283
Telp: (0274) 884201 - 207, Fax: (0274) 884208
E-mail : ch.a6.rowi@gmail.com¹⁾

Terima Naskah : 5 April 2018

Terima Revisi : 29 April 2018

ABSTRAK

Penentuan penerima beasiswa harus mempertimbangkan multifaktor sebagai penentu untuk memastikan bahwa pihak penerima memang layak mendapatkan beasiswa. Metode *data mining* untuk klasifikasi dapat digunakan untuk membantu meningkatkan kecepatan dan ketepatan dalam penentuan penerima beasiswa. Perbandingan kinerja algoritma C4.5 dan *Naive Bayes* bertujuan untuk mengukur tingkat akurasi dan lama waktu proses (*execution time*) dari masing-masing algoritma untuk mendapatkan algoritma terbaik yang akan diterapkan dalam membantu proses penentuan penerima beasiswa. Penelitian menggunakan data sekunder berupa daftar pemohon dan penerima beasiswa sebagai *data set* yang memiliki enam faktor penentu, yaitu semester, IPK, prestasi ko/ekstra kurikuler, penghasilan orang tua, beban biaya listrik dan jumlah tanggungan orang tua. Pengujian dengan *10-fold cross validation* sekaligus evaluasi kinerja model menggunakan *tool* RapidMiner. Hasil penelitian menunjukkan tingkat akurasi model algoritma C4.5 sebesar 96.40% lebih baik dari tingkat akurasi model algoritma *Naive Bayes* sebesar 95.11%, sedangkan waktu proses dari kedua model algoritma yang diteliti menunjukkan hasil 0 s. Hal ini disebabkan jumlah *data set* yang terlalu sedikit untuk dapat mengidentifikasi perbedaan waktu proses antar kedua model algoritma.

Kata Kunci: klasifikasi *data mining*, C4.5, *Naive Bayes*

ABSTRACT

Determination of scholarship recipients should consider multifactors as determinants to ensure that the recipient is indeed eligible for a scholarship. Data mining methods for classification can be used to help improve speed and accuracy in determining scholarship recipients. Comparison of performance of algorithm C4.5 and Naive Bayes aims to measure the level of accuracy and length of time of each algorithm to get the best algorithm that will be applied in assisting the process of determining the scholarship recipient. The study used secondary data in the form of list of applicants and scholarship recipients as data sets which have six determinants of semester, GPA, co-extracurricular achievement, parent income, electrical expense and parent dependent amount. Testing with 10-fold cross validation as well as model performance evaluation using RapidMiner tool. The results showed that the accuracy of C4.5 algorithm model is 96.40% better than the accuracy of Naive Bayes algorithm model 95.11%, while the execution time of the two algorithm models studied shows 0 s. This is due to the amount of data sets that are too small to be able to identify process time differences between the two algorithm models

Keywords: *data mining classification*, C4.5, *Naive Bayes*

PENDAHULUAN

Peraturan Pemerintah nomor 48 tahun 2008 tentang Pendanaan Pendidikan, bagian kelima, pasal 27 ayat (1), menyebutkan bahwa pemerintah dan pemerintah daerah sesuai kewenangannya memberi bantuan biaya pendidikan atau beasiswa kepada peserta didik yang orang tua atau walinya tidak mampu membiayai pendidikannya. Pasal 27 ayat (2) menyebutkan bahwa pemerintah dan pemerintah daerah sesuai dengan kewenangannya dapat memberi beasiswa kepada peserta didik yang berprestasi. Untuk menentukan mahasiswa yang memenuhi kualifikasi mendapatkan beasiswa dengan mempertimbangkan beberapa kriteria. Mengacu pada ketentuan pemerintah [1] yang mempersyaratkan mahasiswa yang dapat diberikan beasiswa adalah:

- a) Serendah-rendahnya pada semester 2 dan setinggi-tingginya pada semester 5 (untuk jenjang D3) dan pada semester 8 (untuk jenjang S1).
- b) Tidak menerima beasiswa/bantuan pendidikan lain dari sumber APBN/APBD.
- c) Memiliki IPK terbaik dengan syarat IPK minimal 3.0.
- d) Memiliki SKS terbanyak dalam satu angkatan.
- e) Memiliki prestasi pada kegiatan ko/ekstra kurikuler.
- f) Memiliki keterbasan kemampuan ekonomi.

Proses seleksi mahasiswa penerima beasiswa di Perguruan Tinggi mengacu pada ketentuan pemerintah tersebut dengan dijabarkan dan disesuaikan dengan karakteristik dan kondisi dari Perguruan Tinggi pelaksana. Dalam penelitian ini menggunakan enam faktor penentu yaitu semester, IPK, prestasi ko/ekstra kurikuler, penghasilan orang tua, beban biaya listrik dan jumlah tanggungan orang tua. Data mahasiswa pemohon beasiswa dibuat rekap dalam lembar kerja (MS Excel) dibahas dalam rapat seleksi untuk menentukan pemohon yang layak. Cara pengambilan keputusan seperti ini masih banyak membutuhkan waktu dan beresiko tidak konsisten akibat unsur subyektivitas. Dengan menggunakan algoritma klasifikasi *data mining* terhadap data yang telah ada akan diketahui pola pemetaan karakteristik pemohon ke keputusan yang diambil sehingga proses seleksi akan

menjadi lebih mudah, konsisten dan dapat menghindari unsur subyektivitas dari pengambil keputusan.

Penelitian-penelitian sebelumnya yang mengkaji kinerja algoritma klasifikasi *data mining* C4.5 dan *Naive Bayes*, penelitian [2] untuk mengetahui tingkat akurasi algoritma C4.5 dalam memprediksi kemampuan siswa sekolah menengah menunjukkan tingkat akurasi 50-60%. Penelitian lain [3] menguji kelayakan algoritma C4.5 sebagai pendukung keputusan dalam pengajuan penerima beasiswa dengan lima faktor penentu pada STIKOM Artha Buana menghasilkan tingkat akurasi 93.0556%, dengan pengujian yang sama tingkat akurasi algoritma *Naive Bayes* 90.2778%. Dalam penelitian [4] yang mengembangkan sistem aplikasi yang menggunakan algoritma *Naive Bayes* untuk penentuan calon penerima beasiswa dengan delapan faktor penentu pada SMK Pasim Plus Sukabumi menunjukkan tingkat akurasi sebesar 96.67%. Dalam penelitian yang membandingkan tingkat akurasi algoritma C4.5 dan *Naive Bayes* untuk ketepatan pemilihan konsentrasi mahasiswa [5] menunjukkan C4.5 lebih unggul atas *Naive Bayes*. Penelitian lain [6] membandingkan akurasi algoritma C4.5, *Neural Network* dan *Naive Bayes* dalam penerapan otentikasi uang kertas (*banknote authentication*) menunjukkan algoritma C4.5 menghasilkan akurasi 98.5% lebih baik dari *Naive Bayes* 85%. Hasil sebaliknya dalam tiga penelitian [7] mengkomparasi metode-metode untuk prediksi klinis dalam sistem penunjang keputusan untuk menghindari resiko-resiko yang terjadi pada proses persalinan, penelitian [8] membandingkan metode klasifikasi *Decision Tree* (model C4.5), *Naive Bayes* dan *k-Nearest Neighbor* untuk pencarian desain alternatif perangkat simulasi energi, dan penelitian [9] yang melakukan studi komparatif dalam memprediksi tingkat performa mahasiswa dengan menggunakan algoritma klasifikasi *Naive Bayes*, C4.5 dan ID3 menunjukkan hasil tingkat akurasi algoritma *Naive Bayes* lebih unggul dari algoritma C4.5.

Pada penelitian ini penulis akan membandingkan kinerja dari dua algoritma klasifikasi *data mining*, yaitu C4.5 dan *Naive Bayes* untuk mendapatkan hasil pengujian yang paling akurat dan tercepat dalam mengklasifikasi mahasiswa yang berhak mendapat beasiswa

sehingga dapat digunakan dalam sistem pendukung keputusan untuk memperbaiki sistem seleksi penentuan mahasiswa penerima beasiswa.

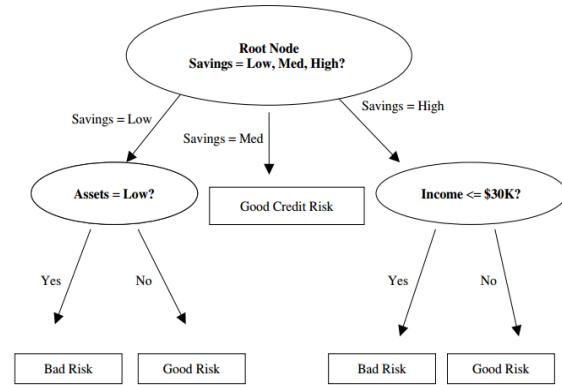
Data Mining

Data mining adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [10]. *Data mining* adalah 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 [11], sehingga sering disebut *Knowledge Discovery in Database* (KDD). Dalam penerapannya data mining memerlukan berbagai perangkat lunak analisis data untuk menemukan pola dan relasi data agar dapat digunakan untuk membuat prediksi dan klasifikasi dengan akurat [12].

Untuk mendukung penelitian ini pembahasan difokuskan pada fungsi klasifikasi *data mining* khusus metode-metode yang digunakan dalam penelitian:

- **Algoritma pohon keputusan (*decision tree*)**

Merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon, dimana setiap node merepresentasikan atribut, dan daun (*leaf*) merepresentasikan kelas. *Node* yang paling atas dari pohon keputusan disebut sebagai akar (*root*) [13], tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu. *Internal node* merupakan *node* percabangan, hanya terdapat satu *input* dan mempunyai *output* minimal dua. Pada *node* daun hanya terdapat satu *input* dan tidak mempunyai *output*. Contoh sebuah pohon keputusan sederhana diperlihatkan pada Gambar 1.



Gambar 1. Pohon keputusan sederhana[10]

Pada pohon keputusan contoh di atas, keputusan kelasnya adalah *Good Risk* atau *Bad Risk*. Apabila *Savings = Low & Assets = Low* (*Bad Risk*), *Savings = Low & Assets = Not Low* (*Good Risk*), *Savings = Medium* (*Good Risk*), *Savings = High & Income <= \$30K* (*Bad Risk*), *Savings = High & Income > \$30K* (*Good Risk*).

Termasuk dalam metode pohon keputusan yang digunakan dalam penelitian ini adalah algoritma C4.5, digunakan untuk membentuk pohon keputusan yang merepresentasikan aturan dalam klasifikasi. Langkah-langkah untuk membentuk pohon keputusan dengan algoritma C4.5 yaitu:

1. Menyiapkan *data training* yang diambil dari data histori atau data masa lampau yang dibuat ke dalam kelas-kelas tertentu.
2. Menghitung akar dari pohon. Memilih atribut sebagai akar didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada [12]. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai *entropy* dengan rumus:

$$Entropy(S) = \sum_{i=1}^n -p_i \log_2(p_i) \tag{1}$$

Keterangan:

- S : himpunan kasus
- n : jumlah partisi S
- p_i: proporsi dari S_i terhadap S

3. Menghitung nilai *gain* dengan persamaan:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \tag{2}$$

Keterangan:

- S : himpunan kasus
- A : atribut
- n : jumlah partisi S
- |S_i| : jumlah kasus pada partisi ke-i
- |S| : jumlah kasus dalam S

4. Dari atribut sebagai akar yang didapat dari langkah 3 dibuat cabang untuk tiap-tiap nilai.
5. Pada setiap cabang yang belum menunjuk pada suatu kelas tertentu diulangi lagi langkah 1 s/d langkah 3 hingga semua cabang telah menunjuk pada suatu kelas dan proses selesai.

- Algoritma Naive Bayes

Merupakan teknik probabilistik klasifikasi berdasarkan teorema Bayes dengan asumsi independensi diantara variabel prediktor. Secara sederhana, pengelompokan *Naive Bayes* menganggap adanya suatu fitur tertentu dalam sebuah kelas tidak terkait dengan adanya fitur lainnya. Teorema Bayes memberikan cara untuk menghitung *posterior probability* P (C|X) dari P (C), P (X) dan P (X|C) dengan persamaan berikut:

$$P(C|X) = \frac{P(X|C) P(C)}{P(X)} \quad (3)$$

Keterangan:

- P(C|X) : *posterior probability* kelas (C, target) yang diberikan prediktor (X, atribut)
- P(C) : probabilitas kelas sebelumnya
- P(X|C) : Kemungkinan yang merupakan probabilitas prediktor kelas yang diberikan
- P(X) : probabilitas prediktor sebelumnya

Evaluasi Kinerja Klasifikasi Data Mining

Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. Evaluasi dimaksudkan untuk menguji model klasifikasi data mining untuk mengetahui kinerja sistem. Metode yang cukup populer untuk evaluasi kinerja klasifikasi *data mining* adalah:

- k-fold Cross Validation

Merupakan salah satu teknik memvalidasi akurasi sebuah model yang dibangun berdasarkan *data set* tertentu. Metode ini membagi *data set* menjadi dua bagian, yaitu *data training* dan *data testing*. *Data set* dibagi

menjadi sejumlah *k*-buah partisi secara acak. Kemudian dilakukan sejumlah *k*-kali proses klasifikasi, dimana masing-masing proses pengujian menggunakan data partisi ke-*k* sebagai *data testing* dan memanfaatkan sisa partisi lainnya sebagai *data training* [14].

- Confusion Matrix

Confusion matrix mengandung informasi mengenai kelas sebenarnya dan kelas prediksi dari suatu proses klasifikasi. Pada dasarnya *confusion matrix* membandingkan hasil klasifikasi yang dilakukan oleh suatu sistem dengan hasil klasifikasi yang sebenarnya [15]. Tabel 1 menunjukkan *confusion matrix* klasifikasi dua kelas.

Tabel 1. *Confusion Matrix*

		Kelas Prediksi	
		Negatif	Positif
Kelas Sebenarnya	Negatif	a	b
	Positif	c	d

Keterangan:

- a: jika hasil prediksi negatif dan nilai sebenarnya negatif.
- b: jika hasil prediksi positif sedangkan nilai sebenarnya negatif.
- c: jika hasil prediksi negatif sedangkan nilai sebenarnya positif.
- d: jika hasil prediksi positif dan nilai sebenarnya positif.

Confusion matrix melakukan perhitungan dengan empat keluaran yaitu:

- *Recall* adalah proporsi kasus positif yang diidentifikasi dengan benar.

$$Recall = d/(c+d) \quad (4)$$

- *Precision* adalah proporsi kasus dengan hasil positif yang benar.

$$Precision = d/(b+d) \quad (5)$$

- *Accuracy* adalah proporsi kasus yang diidentifikasi benar terhadap jumlah semua kasus.

$$Accuracy = (a+c)/(a+b+c+d) \quad (6)$$

Nilai *accuracy* menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar.

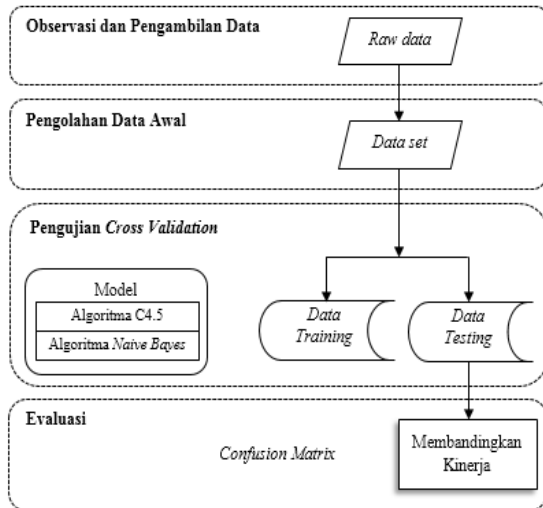
- *Error Rate* adalah kasus yang diidentifikasi salah dari semua kasus.

$$Error Rate = (b+c)/(a+b+c+d) \quad (7)$$

METODE

Alur Penelitian

Alur penelitian dimulai dengan observasi dan pengambilan data (*raw data*), pengolahan *raw data* menjadi *data set*, pengujian *cross validation* setiap model, evaluasi dan perbandingan kinerja. Alur penelitian diperlihatkan pada Gambar 2.



Gambar 2. Alur Penelitian

Metode Pengumpulan Data

a) Observasi

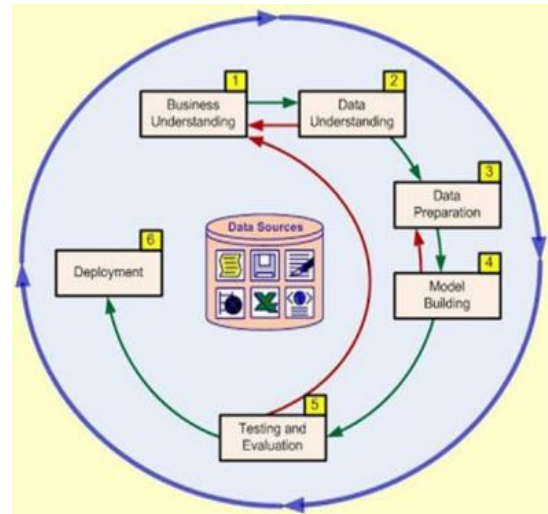
Melakukan pengamatan langsung pada bagian yang menangani urusan seleksi beasiswa, penulis mendapatkan data sekunder yang berupa rekap daftar mahasiswa pemohon dan penerima beasiswa. Pada data mentah ini kemudian dilakukan pengolahan data awal menjadi *data set*.

b) Wawancara

Melakukan wawancara dengan koordinator dan petugas administrasi seleksi penerima beasiswa untuk mendapatkan penjelasan tentang kebijakan dan proses seleksi penerima beasiswa.

Metode Analisis Data

Metode analisis data mengacu pada tahapan proses CRISP-DM (*Cross-Industry Standard Process for Data Mining*) sebagai proses standar dalam *data mining* yang dapat diaplikasikan di berbagai sektor Industri. Gambar 3 menjelaskan tentang siklus hidup pengembangan *data mining* yang telah ditetapkan dalam CRISP-DM.



Gambar 3. Enam Tahapan Proses CRISP Dalam Data Mining[10]

1. Memahami sistem dan permasalahan dalam seleksi penerimaan beasiswa.
2. Menggunakan data sekunder berupa rekap daftar pemohon dan penerima beasiswa.
3. Melakukan proses awal terhadap data mentah yang didapat sehingga menjadi *data set*.
4. Membangun model dengan pelatihan algoritma yang digunakan dengan *data training*.

Pengujian setiap model dengan *data testing* dan membandingkan kinerja antar model.

HASIL DAN PEMBAHASAN

Data set

Pengolahan data awal terhadap *raw data* menghasilkan *data set* dengan enam atribut faktor penentu seperti pada Tabel 2.

Tabel 2. Atribut *data set*

Atribut	Nilai
Semester	Bawah/Atas
IPK	Cukup/Tinggi
Prestasi ko/ekstra kurikuler	Tidak/Ada
Penghasilan orang tua	Rendah/Sedang/Tinggi
Biaya listrik	Rendah/Sedang/Tinggi
Jumlah tanggungan	Rendah/Sedang/Tinggi

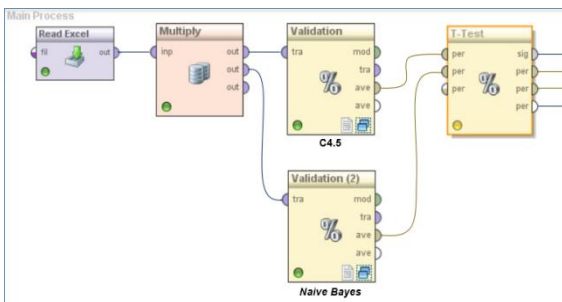
Dataset yang dihasilkan sebanyak 164 record dan digunakan dalam penelitian seperti terlihat pada Tabel 3.

Tabel 3. Potangan data set

No	Semester	IPK	Prestasi	Penghasilan	Biaya listrik	Jumlah tang	Keputus
1	Atas	Cukup	Tidak	Rendah	Rendah	Sedang	Ya
2	Atas	Tinggi	Tidak	Rendah	Rendah	Sedang	Ya
3	Atas	Tinggi	Tidak	Sedang	Sedang	Sedang	Ya
4	Atas	Cukup	Ada	Rendah	Rendah	Sedang	Ya
5	Atas	Cukup	Tidak	Tinggi	Sedang	Sedang	Tidak
6	Atas	Cukup	Tidak	Rendah	Rendah	Sedang	Ya
7	Atas	Cukup	Tidak	Sedang	Rendah	Rendah	Tidak
8	Atas	Tinggi	Tidak	Rendah	Rendah	Sedang	Ya
9	Atas	Cukup	Tidak	Sedang	Sedang	Sedang	Tidak
10	Atas	Tinggi	Tidak	Sedang	Rendah	Sedang	Ya
11	Atas	Cukup	Tidak	Rendah	Rendah	Rendah	Tidak
12	Atas	Cukup	Tidak	Rendah	Rendah	Sedang	Ya
13	Atas	Cukup	Tidak	Sedang	Sedang	Sedang	Tidak
14	Atas	Cukup	Tidak	Sedang	Sedang	Tinggi	Ya
15	Atas	Cukup	Tidak	Rendah	Rendah	Sedang	Ya

Pengujian dan Evaluasi dengan RapidMiner

Pengujian model klasifikasi dengan data set menggunakan tool RapidMiner dengan 10-fold cross validation. Konfigurasi proses pengujian pada RapidMiner seperti terlihat pada Gambar 4.



Gambar 4. Screenshot konfigurasi proses pengujian pada RapidMiner

Perbandingan tingkat akurasi dan waktu proses dari pengujian model klasifikasi seperti terlihat pada Tabel 4.

Tabel 4. Perbandingan kinerja model yang diuji

Model	Accuracy	Execution Time
C4.5	96.4%	0 s
Naive Bayes	95.11%	0 s

Analisis Hasil

Hasil pengujian terhadap model klasifikasi menunjukkan tingkat akurasi dari model algoritma C4.5 sebesar 96.4% dengan waktu proses 0s dan tingkat akurasi model algoritma

Naive Bayes sebesar 95.11% dengan waktu proses 0s. Perbandingan tingkat akurasi sejalan dengan hasil penelitian sejenis di tempat lain [3] dimana tingkat akurasi model algoritma C4.5 lebih tinggi dari model algoritma Naive Bayes. Waktu proses (execution time) kedua model terdeteksi 0s disebabkan jumlah data set terlalu sedikit untuk dapat mengidentifikasi waktu yang diperlukan untuk proses klasifikasi.

SIMPULAN

Berdasarkan hasil implementasi algoritma C4.5 dan Naive Bayes untuk model klasifikasi penerima beasiswa pada penelitian ini menunjukkan model algoritma C4.5 mempunyai kinerja yang lebih baik dari Naive Bayes. Hal ini sejalan dengan hasil penelitian sejenis di tempat lain [3]. Hasil perbandingan kinerja sebaliknya terjadi yang mana kinerja model algoritma Naive Bayes lebih baik dari C4.5 dalam penelitian dengan obyek yang berbeda [7], [8] dan [9].

Untuk memilih model klasifikasi terbaik dari algoritma-algoritma yang akan dipilih untuk diimplementasikan disarankan mengacu pada hasil penelitian lain yang telah membandingkan kinerja model-model algoritma yang akan dipilih jika ada. Untuk membandingkan kinerja dari segi waktu proses (execution time) disarankan menggunakan data set dengan jumlah data yang memadai untuk identifikasi waktu proses.

DAFTAR PUSTAKA

- [1] Kementerian Riset Teknologi Dan Pendidikan Tinggi. Pedoman Teknis Beasiswa Peningkatan Prestasi Akademik (PPA) Tahun 2017.
- [2] Rizki Haqmanullah Pambudi, Budi Darma Setiawan, Indriati. Penerapan Algoritma C4.5 Dalam Program Untuk Memprediksi Kinerja Siswa Sekolah Menengah. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. E-ISSN: 2548-964X Vol. 2, No. 7, 2017.
- [3] Jose Augusto Duarte Guterres. Kelayakan Algoritma C4.5 Sebagai Pendukung Keputusan Dalam Pengajuan Penerima Beasiswa. Seminar Nasional Teknologi Informasi 2015

- [4] Rizal Amegia Saputra, Shinta Ayuningtias. Penerapan Algoritma *Naive Bayes* Untuk Penentuan Calon Penerima Beasiswa pada SMK Pasim Plus Sukabumi. SWABUMI Vol. IV, No. 2, 2016.
- [5] Wiwit Supriyanti, Kusriani, Armadyah Amborowati. Perbandingan Kinerja Algoritma C4.5 Dan *Naive Bayes* Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa. Jurnal INFORMA Politeknik Indonusa Surakarta ISSN: 2242-7942 Vol. 1 No. 3, 2016.
- [6] Khairul Sani, Wing Wahyu Winarno, Silmi Fauziati. Analisis Perbandingan Algoritma *Classification* Untuk *Authentication* Uang Kertas (Studi kasus: *Banknote Authentication*). Jurnal Informatika Vol. 10, No. 1, 2016.
- [7] Hilda Amalia dan Evicienna. Komparasi Metode *Data Mining* Untuk Penentuan Proses Persalinan Ibu Melahirkan. Jurnal Sistem Informasi (*Journal of Information System*). 2/13 (2017), 103-109
- [8] Ahmad Ashari, Iman Paryudi, A Min Tjoa. *Performance Comparison between Naive Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. (IJACSA) International Journal of Advanced Computer Science and Application*, Vol. 4, No. 11, 2013
- [9] A. Rajalakshmi M.Sc., K. Sivaranjani, M.Sc., M.Phil. *A Comparative Study on Student Performance Prediction. International Conference on Engineering Trends and Science & Humanities (ICETSH-2015)*
- [10] Daniel T. Larose. 2005. *DISCOVERING KNOWLEDGE IN DATA. An Introduction to Data Mining*. John Wiley & Sons, Inc.
- [11] Efraim Turban, Jay E. Aronson, Ting-Peng Liang. 2005. *Decision Support Systems and Intelligent Systems*. Prentice-Hall, Inc.
- [12] Kusriani, Emha Taufiq Luthfi. 2009. ALGORITMA DATA MINING. Penerbit ANDI.
- [13] Gorunescu, F. 2011. *Data Mining Concept Model and Techniques*. Berlin: Springer. ISBN 978-3-642-19720-8.
- [14] Yaya Winardi. *K-Folds Cross Validation*. <http://blog.yayaw.web.id/riset/k-folds-cross-validation>, 2 April 2013
- [15] Eko Prasetyo. 2012. *Data Mining, Konsep Dan Aplikasi Menggunakan MATLAB*. Penerbit ANDI Yogyakarta.