IMPLEMENTASI DATA MINING DENGAN ALGORITMA *DECISION TREE* C4.5 UNTUK PREDIKSI KELULUSAN MAHASISWA DI UNIVERSITAS PANDANARAN

Abdul Rohman[1], Anief Rufiyanto[2]
[1,2]Program Studi Teknik Elektronika, Fakultas Teknik, Universitas Pandanaran
email: [1] abdulrohman@unpand.ac.id, [2]aniefrufiyanto@gmail.com

ABSTRAK

Memprediksi prestasi akademik mahasiswa sangat penting bagi penyelenggara pendidikan karena program strategis tersebut dapat direncanakan dalam meningkatkan atau mempertahankan kinerja mahasiswa selama masa studi di perguruan tinggi. Data mahasiswa menjadi hal yang sangat penting untuk mengambil suatu keputusan, jika data tersebut dianalisa dengan menggunakan *data mining*. Setiap kumpulan atau gudang data dapat memberikan pengetahuan penting yang menjadi informasi yang sangat berharga bagi perguruan tinggi. Pada perguruaan tinggi, suatu sistem informasi dapat digunakan untuk memperoleh informasi yang menunjang setiap pada pengambilan suatu keputusan. Data mining dengan algoritma *Decision Tree C4.5* dapat digunakan untuk menyusun sistem yang mempunyai kemampuan melihat pola kelulusan mahasiswa. Banyak penelitian tentang implementasi data mining untuk memprediksi kelulusan mahasiswa dengan menggunakan algoritma *Decision Tree*, dengan data mahasiswa reguler dan mayoritas statusnya belum bekerja. Sedangkan dalam penelitian ini mahasiswa Universitas Pandanaran, memiliki data mahasiswa kelas reguler dan mahasiswa kelas karyawan dan kebanyakan statusnya sudah bekerja. Tahapan yang dilakukan dalam penelitian ini yaitu; (1) pengumpulan data mahasiswa universitas pandanaran, (2) mengolah data mahasiswa dengan menggunakan klasifikasi data mining algoritma *Decision Tree* (3) eksperimen dan pengujian algoritma (4) evaluasi dan validasi hasil (5) Menghasilkan Pola/Model Kelulusan Mahasiswa yang dapat dimanfaatkan untuk sebuah keputusan diperguruan tinggi. Hasil dari penelitian ini menghasilkan 10 rule dengan nilai akurasi 65,98% dengan nilai AUC 0,874, dan dapat dikategorikan sebagai klasifikasi data yang baik. Maka hasil tersbut sangat penting untuk dijadikan pengambilan keputusan dalam lembaga.

Kata Kunci : *Data Mining, Decision Tree, C4.5*, Mahasiswa

1. PENDAHULUAN

Data mahasiswa menjadi hal yang sangat penting untuk mengambil suatu keputusan, jika data tersebut dianalisa dengan menggunakan *data mining*. Setiap kumpulan atau gudang data dapat memberikan pengetahuan penting yang menjadi informasi yang sangat berharga bagi perguruan tinggi. Pada perguruaan tinggi, suatu sistem informasi dapat digunakan untuk memperoleh informasi yang menunjang setiap pada pengambilan suatu keputusan [1].

Data mining dengan algoritma *Decision Tree* banyak digunakan untuk menyusun sistem yang mempunyai kemampuan melihat pola kelulusan mahasiswa, karena banyak kelebihan dengan algoritma yang lainnya. Kelebihan Algoritma *Decision Tree* dapat menghasilkan pohon keputusan yang mudah diinterprestasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik [2].

Banyak penelitian tentang implementasi data mining untuk memprediksi kelulusan mahasiswa dengan menggunakan algoritma *Decision Tree*, dengan data mahasiswa reguler dan mayoritas statusnya belum bekerja [3], [4], [5], [6], [7], Sedangkan dalam penelitian ini mahasiswa Universitas Pandanaran terutama jenjang pendidikan D3 di Fakultas Teknik, memiliki data mahasiswa kelas reguler dan mahasiswa kelas karyawan dan kebanyakan statusnya sudah bekerja. Maka dengan itu diperlukan penelitian untuk memprediksi kelulusan mahasiswa.

2.TINJAUAN PUSTAKA

a. Kelulusan Mahasiswa

Mahasiswa sering disebut kelompok masyarakat yang memiliki ciri intelektualitas yang lebih luas dibandingkan dengan kelompok seusia mereka yang bukan mahasiswa ataupun kelompok usia lain yang dibawah mereka. Dengan intelektualitasnya mahasiswa akan mampu menghadapi dan mencari permasalahan secara sistematis yang nantinya diterapkan dalam kehidupan sehari-hari agar bisa bersaing dalam dunia kerja[8].

Kelulusan mahasiwa adalah hal yang penting diperhatikan, karena persentase jumlah kelulusan mempengaruhi penilaian pemerintah serta mempengaruhi status akreditasi program studi [9]. Faktor-faktor yang dapat mempengaruhi kelulusan mahasiswa antara lain adalah nilai akhir SMA, Indeks Prestasi Semester (IPS), gaji orang tua dan pekerjaan orang tua [6]. Indeks prestasi sering digunakan sebagai indikator penilaian akademik, banyak perguruan tinggi memberi standar minimum yang sulit di peroleh mahasiswa [10]. Adapun variabel yang dapat digunakan dalam prediksi kelulusan mahasiswa seperti umur, status pernikahan, jumlah

saudara [11]. Pada penenlitian ini parameter yang digunakan adalah nama jurusan, usia, jenis kelamin, status pekerjaan dan indeks prestasi semester satu sampai dengan indeks prestasi semester empat.

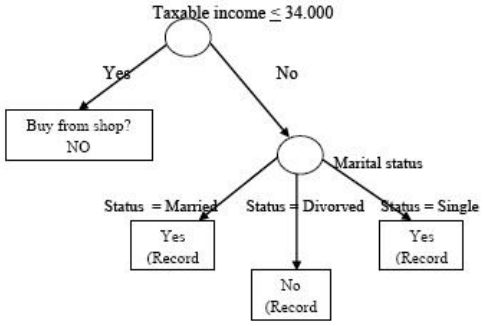
b. Klasifikasi Data Mining

Data mining adalah serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data [12]. *Data mining* akan memecahkan masalah dengan menganalisis data yang telah ada dalam basis data. Data mining, sering juga disebut *knowledge discovery in database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan pola keteraturan, pola hubungan dalam set data berukuran besar [13].

Dalam penelitian ini akan memanfaatkan *data mining* untuk mengklasifikasi data mahasiswa dengan jenjang pendidikan D3 di Fakultas Teknik Universitas Pandanaran.

c. Algoritma Decision Tree C4.5

Algoritma *decision tree C4.5* digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar [14]. *Decision Tree* atau pohon keputusan adalah model prediksi menggunakan struktur pohon atau hirarki dengan mengubah data menjadi pohon keputusan dan aturan-aturan keputusan.



Gambar 1. Contoh Konsep Keputusan Sederhana[14]

Beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma *Decision Tree C4.5* [14]. adalah sebagai berikut:

1. Mempersiapkan data *training*, dapat diambil dari data *history* yang pernah terjadi sebelumnya dan sudah dikelompokan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai index entropy, dengan rumus:

$$Entropy(i) = \sum_{j=1}^{m} f(i,j).2f[(i,j)] \quad (1)$$

3. Hitung nilai *gain* dengan rumus:

$$gain = -\sum_{i=1}^{p} {}^{ni}_{n} IE(i) \quad (2)$$

4. Untuk menghitung *gain ratio*perlu diketahui suatu term baru yang disebut *Split Inormation* dengan rumus:

$$Split\ Information = -\sum_{t=1}^{c} \frac{s1}{s} log_2 \frac{s1}{s} \quad (3)$$

5. Selanjutnya menghitung *ratio*

$$Gainratio(S.A) = \frac{Gain(S.A)}{SplitInformation(S.A)} \quad (4)$$

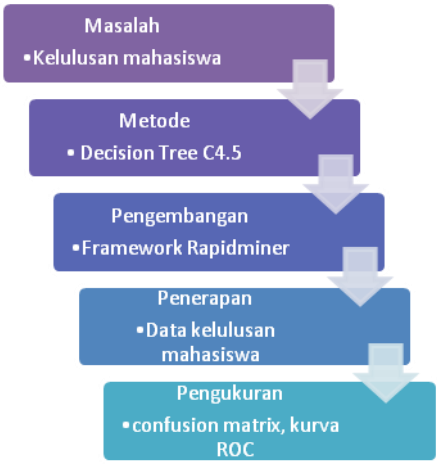
6. Mengulangi langkah ke-2 hingga semua *record* terpartisi

3.METODE PENELITIAN

Dalam penelitian ini menggunakan data kelulusan mahasiswa di Universitas Pandanaran dengan jenjang pendidikan D3 di Fakultas Teknik dari program studi Teknik Sipil, Teknik Mesin, Teknik Elektronika, Teknik Lingkungan, dan Teknik Kimia.

Jumlah data yang diolah yaitu 235 mahasiswa, yaitu terdiri dari; 1) Teknik Sipil 53 mahasiswa, Teknik Mesin 95 mahasiswa, Teknik Elektronika 45 mahasiswa, Teknik Lingkungan 25 mahasiswa dan Teknik Kimia 17 mahasiswa. Dari data tersebut terdapat 151 mahasiswa yang lulus tepat waktu dan 84 mahasiswa yang terlambat.

Kerangka pemikiran dalam penelitian ini adalah:



Gambar 2. Kerangka Pemikiran

Tabel 1. Dataset Kelulusan Mahasiswa

jurusan	umur	jk	pekerjaan	ipk1	ipk2	ipk3	ipk4	label
DIII Teknik Sipil	23	Laki-laki	bekerja	3.62	3.76	3.67	3.5	tepat
DIII Teknik Sipil	28	Laki-laki	bekerja	3.29	3.19	3.05	3.2	terlambat
DIII Teknik Sipil	29	Laki-laki	bekerja	3.29	3.14	2.76	3.5	terlambat
DIII Teknik Sipil	35	Laki-laki	bekerja	3.19	3.1	2.9	3.1	terlambat
DIII Teknik Sipil	35	Laki-laki	bekerja	3.62	3.52	3.62	3.3	terlambat
DIII Teknik Sipil	27	Laki-laki	bekerja	2.86	2.86	3	3.1	terlambat
DIII Teknik Sipil	23	Laki-laki	bekerja	3.05	3.19	3.29	3.4	terlambat
DIII Teknik Sipil	33	Perempuan	bekerja	3.29	3.05	3.48	3.3	terlambat
DIII Teknik Sipil	21	Laki-laki	bekerja	3.48	3.14	3.38	3.3	terlambat
DIII Teknik Sipil	22	Laki-laki	bekerja	3.23	3.24	3.29	3.1	terlambat
DIII Teknik Sipil	25	Perempuan	bekerja	3.57	3.9	3.17	3.4	tepat
DIII Teknik Sipil	39	Laki-laki	bekerja	3.67	3.57	3.48	3	tepat
DIII Teknik Sipil	27	Laki-laki	bekerja	3.48	3.67	3.48	3	tepat
DIII Teknik Sipil	33	Laki-laki	bekerja	3.48	3.1	3.05	3.5	tepat
DIII Teknik Sipil	34	Laki-laki	bekerja	3.24	2.86	2.57	2.8	tepat
DIII Teknik Sipil	23	Laki-laki	bekerja	2.9	3	3	3	terlambat
DIII Teknik Sipil	30	Laki-laki	bekerja	3.19	3.24	3.57	3.1	tepat
DIII Teknik Sipil	31	Laki-laki	bekerja	3.48	3.14	2.86	2.9	tepat
DIII Teknik Sipil	28	Laki-laki	bekerja	3.43	3.57	2.76	2.8	tepat
DIII Teknik Sipil	24	Laki-laki	bekerja	3.29	3.29	2.57	2.7	tepat
DIII Teknik Sipil	43	Laki-laki	bekerja	3.38	3.24	2.48	2.4	tepat
DIII Teknik Sipil	25	Laki-laki	bekerja	3.38	3.57	2.38	2.4	tepat
DIII Teknik Sipil	30	Laki-laki	bekerja	3.19	3.24	3.57	3.1	tepat
DIII Teknik Sipil	30	Laki-laki	bekerja	3.05	3.43	3.19	3.5	terlambat
DIII Teknik Sipil	26	Laki-laki	bekerja	2.86	2.95	2.62	3.3	terlambat
DIII Teknik Sipil	27	Laki-laki	bekerja	3.05	3.35	3	3.2	terlambat
DIII Teknik Sipil	26	Laki-laki	bekerja	2.9	3.26	2.84	3.2	terlambat
DIII Teknik Sipil	28	Laki-laki	bekerja	3.29	3.26	2.84	3.3	terlambat

dst….

Penelitian ini adalah penelitian experiment yang melibatkan penyelidikan tentang perlakuan pada parameter dan variabel yang semuanya tergantung pada peneliti itu sendiri. Software dan hardware sebagai alat bantu dalam penelitian ini dapat dilihat pada tabel 2.

Tabel 2. Spesifikasi Hardware dan Sofware

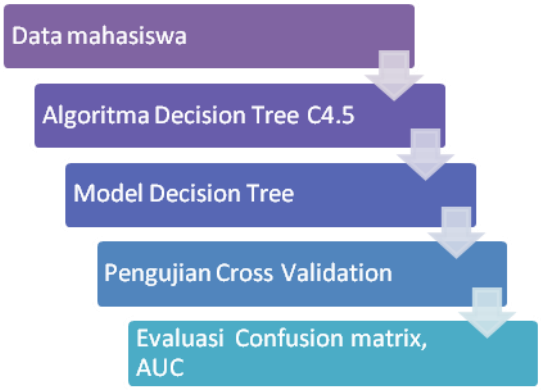
Software	*Hardware*
Sistem Operasi: Windows 10 64 bit *Data Mining: RapidMiner Studio 9,3*	*CPU: CORE I3 2,00 Ghz* *Ram 4 GB, Hdd 1 TB*

Model yang diusulkan pada penelitian ini adalah menggunakan algoritma klasifikasi *data mining Decision Tree C4.5*. Setelah data diolah dan menghasilkan model, maka dilanjutkan dengan pengujian menggunakan *k-fold cross validation*, kemudian dilakukan evaluasi dan validasi hasil dengan *confusion matrix* dan kurva *ROC* dengan AUC *(Area Under Curve)*.

Performance keakurasian AUC [14], dapat diklasifikasikan menjadi lima kelompok yaitu:

a. 0.50 – 0.60 = klasifikasi salah
b. 0.60 – 0.70 = klasifikasi buruk
c. 0.70 – 0.80 = klasifikasi cukup
d. 0.80 – 0.90 = klasifikasi baik
e. 0.90 – 1.00 = klasifikasi sangat baik

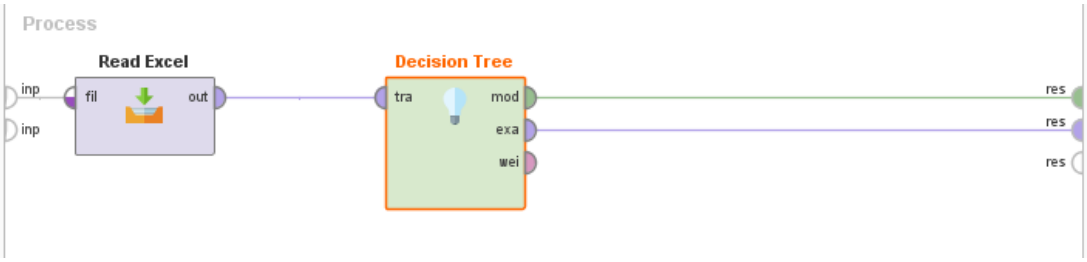
Dalam tahapannya seperti gambar dibawah ini:



Gambar 3. Metode yang diusulkan

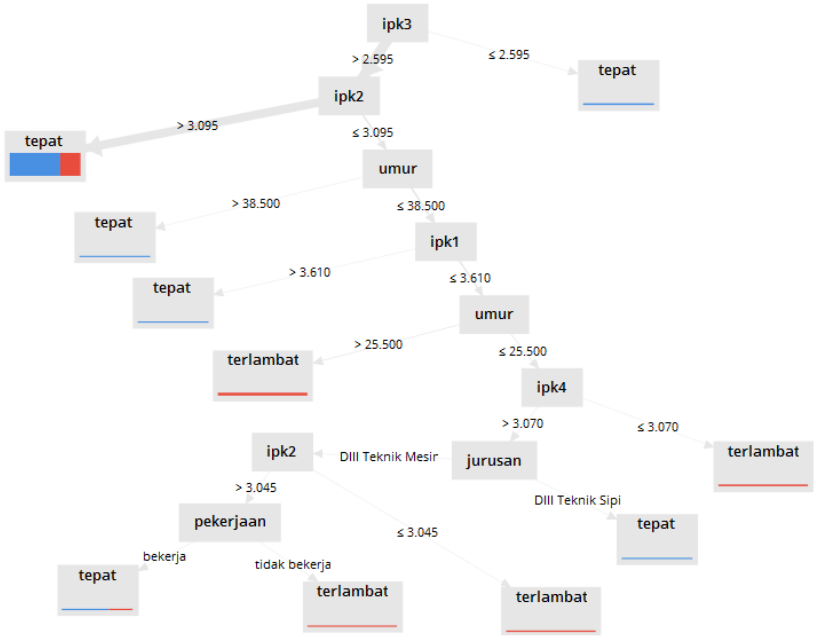
4.HASIL DAN PEMBAHASAN

Dari data kelulusan mahasiswa D3 Fakultas Teknik Universitas Pandanaran yang terkumpul sebanyak 235 data dengan 151 adalah lulus tepat waktu dan 84 adalah terlambat., maka data tersebut diolah dengan menggunakan algoritma *Decision Tree* C4.5.



Gambar 4. Proses Olah Data dengan *Decision Tree C4.5*

Sehingga menghasilkan model pohon keputusan yang dapat dilihat pada gambar 5.

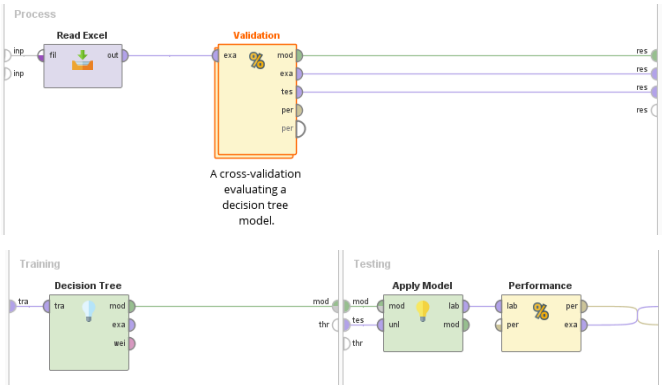


Gambar 5. Model Pohon Keputusan Kelulusan Mahasiswa

Tujuan utama dari menganalisis data dengan menggunakan algoritma *Decision Tree* ini adalah ingin mendapatkan rule [12], yang akan dimanfaatkan untuk pengambilan keputusan pada data baru. Adapun rule yang didapat adalah:

1. Jika ipk3 <= 2,59 maka mahasiswa lulus tepat waktu
2. Jika ipk2 > 3,095 maka mahasiswa lulus tepat waktu
3. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur > 38,5 maka mahasiswa lulus tepat waktu
4. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 maka mahasiswa lulus tepat waktu
5. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 dan umur > 25,5 maka mahasiswa terlambat
6. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 dan umur < 25,5 dan ipk4 >3,070 dan jurusan teknik mesin dan ipk2 > 3,045 dan kelas karyawan maka mahasiswa lulus tepat waktu
7. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 dan umur < 25,5 dan ipk4 >3,070 dan jurusan teknik mesin dan ipk2 > 3,045 dan kelas reguler maka mahasiswa terlambat
8. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 dan umur < 25,5 dan ipk4 <= maka mahasiswa terlambat
9. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 dan umur < 25,5 dan ipk4 >3,070 dan jurusan teknik sipil maka mahasiswa lulus tepat waktu
10. Jika ipk3 >2,595 dan ipk2 >= 3,095 dan umur < 38,5 dan ipk1 > 3,610 dan umur < 25,5 dan ipk4 >3,070 dan jurusan teknik mesin dan ipk2 <= 3,045 maka mahasiswa terlambat

Setelah mendapatkan model dan rule, langkah selanjutnya pengujian algoritma *Decision tree* terhadap data kelulusan mahasiswa dengan *K-Fold Cross Validation*. Dalam pengujian K-*Fold Cross Validation* algoritma *Decision Tree C4.5*, peneliti menggunakan 10 kali percobaan dengan sampling type Stratified (bertingkat-tingkat).



Gambar 6. Proses Pengujian

Hasil dari proses pengujian diatas menghasilkan *confision matrix* sebagai berikut:

accuracy: 65.98% +/- 5.09% (micro average: 65.96%)

	true tepat	true terlambat	class precision
pred. tepat	143	72	66.51%
pred. terlambat	8	12	60.00%
class recall	94.70%	14.29%	

Gambar 7. *confision matrix*

Jumlah *True Positive* (TP) adalah 143 *record* diklasifikasikan sebagai TEPAT terpilih dan *False Negative* (FN) sebanyak 72 *record* diklasifikasikan sebagai TEPAT terpilih tetapi TERLAMBAT terpilih. Berikutnya 12 *record* untuk *True Negative* (TN) diklasifikasikan sebagai TERLAMBAT terpilih, dan 8 *record False Positive* (FP) diklasifikasikan sebagai TERLAMBAT terpilih ternyata TEPAT. Sehingga nilai akurasinya adalah 65,98%.

Dan dari confision matrix diatas juga menghasilkan *performance* keakurasian dengan nilai AUC (*Area Under Curve*) sebesar 0.874 dengan nilai akurasi Baik.

5 .KESIMPULAN

Dalam penelitian ini dilakukan pengolahan data kelulusan mahasiswa Universitas Pandanaran dengan menggunakan algoritma *Decision Tree* sehingga menghasilkan model dan rule. Dan dari hasil model kelulusan

mahasiswa tersebut dilakukan evaluasi algoritma sehingga menghasilkan nilai akurasi 65,98 % dengan nilai AUC 0,874 dan termasuk klasifikasi data Baik.

UCAPAN TERIMA KASIH

Kami sangat berterima kasih kepada Kemenristek-Dikti atau Kementerian Riset dan Pendidikan Tinggi Indonesia (DPRM-DIKTI) yang membiayai penelitian.

DAFTAR PUSTAKA

[1] Y. Asriningtias *et al.*, 2014, Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa, *J. Inform.*, vol. 8, no. 1, hal 837–848.

[2] M. Ridwan, H. Suyono, and M. Sarosa, 2013, Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, hal 59–64.

[3] D. Himawan, 2014. Aplikasi Data Mining Menggunakan Algoritma ID3 Untuk Mengklasifikasi Kelulusan Mahasiswa Pada Universitas Dian Nuswantoro Semarang, *Fak. Ilmu Komput.*, Semarang.

[4] Muhammad Sholeh, 2014, Penerapan algoritma C4.5 Untuk Klasifikasi Prediksi Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta , *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, Yogyakarta, 15 November 2014.

[5] H. Romadhona, Agus, suprapedi dan himawan, 2017, Prediksi Kelulusan Tepat Waktu Mahasiswa Stmik-Ymi, *J. Teknol. Inf*, vol. 13, no. 1, hal 69–83,.

[6] E. Marselina Silvia Suhartinah, 2010, Graduation Prediction Of Gunadarma University Students Using Algorithm And Naive Bayes C4.5 Algorithm, *Fac. Ind. Technol. Gunadarma Univ*, Jakarta.

[7] E. S. Siska Haryati, Aji Sudarsono, 2015, Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, hal 130–138.

[8] Azwar, 2014, *Penyusunan Skala Psikologi*, Pustaka Pelajar, Yogyakarta.

[9] S. T. Karamouzis and A. Vrettos, 2008, "An Artificial Neural Network for Predicting Student Graduation Outcomes," *Lect. Notes Eng. Comput. Sci.*, vol. 2173, no. 1, hal 991–994.

[10] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, 2010, Application of k Means Clustering algorithm for prediction of Students Academic Performance, vol. 7, hal 292–295.

[11] J. Yingkuachat, P. Praneetpolgrang, and B. Kijsirikul, 2007, An Application of the Probabilistic Model to the Prediction of Student Graduation Using Bayesian Belief Networks," *Electr. Eng. Comput. Telecommun. Inf. Technol. Assoc. Thail. (ECTI Thailand)*, vol. 3, no 1, hal 63–71.

[12] I. H. Witten, 2007, *Data Mining Data Mining Complications : Overfitting Statistical modeling One attribute does all the work*, Morgan Kaufmann Publisher, Burlington.

[13] Budi Santosa, 2007, *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, Yogyakarta.

[14] F. Gorunescu, 2011, *Data Mining: Concepts, Models and Techniques (Intelligent Systems Reference Library)*, Springer, Berlin.