Tersedia di https://jtsiskom.undip.ac.id (28 April 2018)
DOI: 10.14710/jtsiskom.6.2.2018.76-83
Jurnal Teknologi dan Sistem Komputer, 6(2), 2018, 76-83

Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa

Dea Alverina[*)], Antonius Rachmat Chrismanto, R. Gunawan Santosa

Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana
Jalan Doktor Wahidin Sudirohusodo No. 5 – 25, Kotabaru, Gondokusuman, Yogyakarta, Indonesia

Cara sitasi: D. Alverina, A. R. Chrismanto, and R. G. Santosa, "Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa," Jurnal Teknologi dan Sistem Komputer, vol. 6, no. 2, Apr. 2018. doi: 10.14710/jtsiskom.6.2.2018.76-83, [Online].

Abstract – This research compared the accuracy of prediction of Grade Point Average (GPA) of the first semester students using C4.5 and CART algorithms in Faculty of Information Technology (FTI), Universitas Kristen Duta Wacana (UKDW). This research also explored various parameters such as numeric attribute categorization, data balance, GPA categories number, and different attributes availability due to the difference of data availability between Achievement Admission (AA) and Regular Admission (RA). The training data used to create decision tree were FTI students, 2008-2015 batch, while the testing data were FTI students, 2016 batch. The accuracy of prediction was measured by using crosstab table. In AA, the accuracy of both algorithms can be achieved about 86.86%. Meanwhile, in RA the accuracy of C4.5 is about 61.54% and CART is about 63.16%. From these accuracy result, both algorithms are better to predict AA rather than RA.

Keywords – *data mining accuracy; student grade prediction; prediction algorithms comparison*

Abstrak – Penelitian ini membandingkan akurasi prediksi kategori Indeks Prestasi (IP) semester pertama mahasiswa Fakultas Teknologi Informasi (FTI) Universitas Kristen Duta Wacana (UKDW) menggunakan algoritma C4.5 dan CART. Penelitian ini juga mengeksplorasi berbagai parameter seperti kategorisasi atribut numerik, keseimbangan data, jumlah kategori IP, dan ketersediaan atribut yang berbeda karena perbedaan ketersediaan data antara jalur prestasi dan jalur non-prestasi. Data mahasiswa FTI tahun 2008-2015 digunakan sebagai data latih sedangkan data uji menggunakan data tahun 2016. Akurasi kedua algoritma dalam memprediksi tersebut diukur dengan menggunakan tabel crosstab. Pada jalur prestasi, akurasi kedua algoritma mampu mencapai 86,86%. Pada jalur non-prestasi, akurasi algoritma C4.5 sebesar 61,54% dan algoritma CART sebesar 63,16%. Dilihat dari segi akurasinya, algoritma C4.5 dan CART lebih baik digunakan untuk memprediksi jalur prestasi daripada jalur non-prestasi.

Kata Kunci – *akurasi data mining; prediksi prestasi mahasiswa; perbandingan algoritma prediksi*

I. PENDAHULUAN

Performa akademis mahasiswa baru di semester pertama yang di bawah rata-rata dapat menimbulkan berbagai masalah, di antaranya adalah efek berantai performa rendah untuk semester-semester berikutnya. Terdapat beberapa faktor eksternal dan internal yang mempengaruhi tinggi rendahnya prestasi akademis mahasiswa baru. Faktor-faktor eksternal meliputi kategori asal SMA (pulau Jawa atau luar pulau Jawa), kategori SMA (SMA atau SMK), dan status SMA (Negeri atau Swasta). Faktor internal meliputi kemampuan spasial, kemampuan verbal, kemampuan numerik dan kemampuan analogi [1]. Lebih lanjut, Indriana dkk. [2] menyatakan bahwa ada perbedaan prestasi akademik mahasiswa yang bekerja dan yang tidak bekerja saat menempuh pendidikan.

Data-data dari faktor tersebut dapat diperoleh pada saat pendaftaran mahasiswa baru, namun tidak semua data dapat diperoleh. Hal tersebut disebabkan oleh adanya dua jalur penerimaan mahasiswa baru, yaitu jalur prestasi dan jalur non-prestasi. Penerimaan mahasiswa baru yang melalui jalur prestasi akan dilihat dari faktor eksternal dan yang melalui jalur non-prestasi akan dilihat dari faktor eksternal maupun faktor internal. Mahasiswa baru yang mendaftar melalui jalur prestasi dan jalur non-prestasi akan diuji kemampuan bahasa Inggrisnya yang dipetakan menjadi beberapa level, yaitu level 1, level 2, level 3, atau level ESP.

Prediksi indeks prestasi (IP) mahasiswa semester satu perlu dilakukan secara dini untuk menanggulangi masalah-masalah yang mungkin akan ditimbulkan oleh mahasiswa di kemudian hari dan melakukan pembimbingan. Prediksi mahasiswa berprestasi telah dilakukan oleh Untari [3] menggunakan algoritma C4.5, sedangkan Kamagi dan Hansun [4] menggunakan algoritma tersebut untuk memprediksi tingkat kelulusan mahasiswa. Metode dan algoritma lain yang digunakan untuk prediksi adalah metode regresi logistik untuk

*) Penulis korespondensi (Dea Alverina)
Email: dea.alverina@ti.ukdw.ac.id

Dikirim: 19 Maret 2018; Direvisi: 3 April 2018, 17 April 2018; Diterima: 18 April 2018

Tabel 1. Jumlah data mahasiswa Fakultas Teknologi Informasi tahun 2008-2015

Tahun	Jalur Prestasi	Jalur Non-Prestasi	Total
2008	63	305	368
2009	11	249	260
2010	55	209	264
2011	144	107	251
2012	125	119	244
2013	125	80	205
2014	90	81	171
2015	193	61	254
Total	**806**	**1211**	**2017**

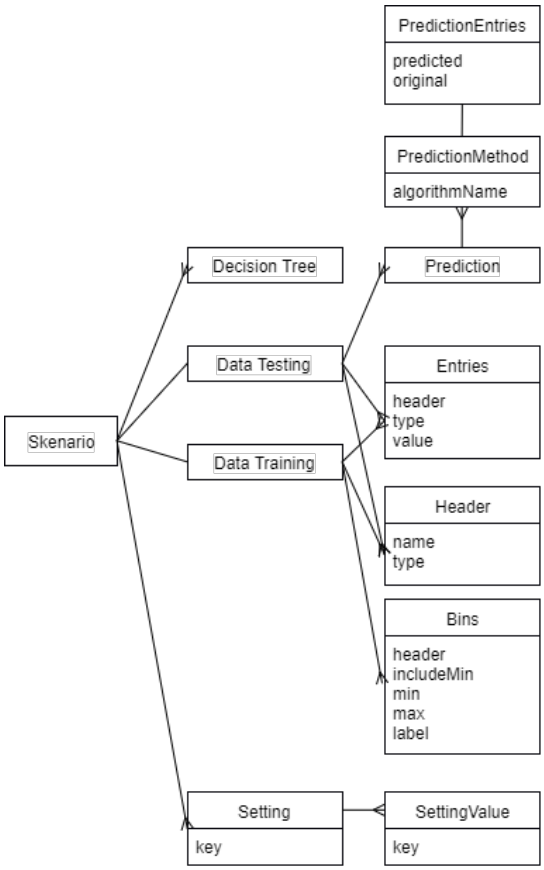
prediksi IP mahasiswa baru dari jalur prestasi [1] dan non-prestasi [5], algoritma K-Nearest Neighbor dan Naive Bayes untuk prediksi kategori IP mahasiswa [6] serta metode CART untuk klasifikasi ketepatan masa studi mahasiswa [7]. Perbandingan akurasi prediksi antar algoritma untuk klasifikasi dan analisis data nilai mahasiswa telah dilakukan, salah satunya adalah antara algoritma C4.5 dan CART [8], [9].

Prediksi dalam [3] dan [4] menggunakan data non-numerik dan kategori *output* yang hanya memiliki 1 macam kategori saja untuk mahasiswa aktif. Selain itu, ditemukan perbedaan dalam [8] yang menyatakan bahwa algoritma C4.5 cocok digunakan untuk klasifikasi data non-numerik daripada algoritma CART yang cocok untuk data numerik dan dalam [9] yang menyatakan bahwa performa CART lebih baik daripada C4.5 dimana data yang digunakan adalah data non-numerik. Penelitian ini bertujuan untuk menerapkan algoritma C4.5 dan CART untuk prediksi kategori IP mahasiswa baru semester satu dengan mengeksplorasi berbagai parameter, yaitu kategorisasi angka numerik, keseimbangan data, jumlah kategori IP dan ketersediaan atribut yang berbeda antara mahasiswa dari jalur prestasi dan jalur non-prestasi. Penelitian ini juga menganalisis dan membandingkan akurasi algoritma C4.5 dan CART untuk menentukan algoritma yang lebih baik dalam memprediksi kategori IP mahasiswa baru semester satu.

II. METODE PENELITIAN

A. Pengumpulan Data

Data yang digunakan pada penelitian ini diambil dari penelitian yang telah dilakukan oleh Santosa & Rachmat [1]. Tabel 1 menunjukkan tabel jumlah data berdasarkan jumlah mahasiswa FTI dari tahun 2008 sampai dengan tahun 2015, yaitu 2.017 data. Data mahasiswa tersebut akan dikategorikan berdasarkan jalur penerimaan, yaitu jalur prestasi dan jalur non-prestasi. Jalur prestasi memiliki atribut kategori (SMU atau SMK), status (negeri atau swasta), lokasi (Jawa atau luar Jawa), dan level (1, 2, 3, atau ESP). Jalur non-prestasi memiliki atribut kategori (SMU atau SMK), status (negeri atau swasta), lokasi (Jawa atau luar Jawa), level (1, 2, 3, atau ESP), nilai numerik, nilai verbal, nilai spasial dan nilai analogi.



Gambar 1. Rancangan struktur basis data

B. Perancangan Sistem

Jalannya program diawali dengan menampilkan kumpulan Skenario dengan struktur objek ditunjukkan pada Gambar 1. Sebuah Skenario menyimpan data latih (*training*), data uji (*testing*), *settings*, dan *decision tree*. Data latih memiliki *entries*, *header*, dan *bins*. Data uji memiliki *entries*, *header*, dan *prediction* (hasil prediksi). *Settings* digunakan untuk menyimpan berbagai parameter, di antaranya ip_kategori dan atribut yang diinferensi. Data *decision tree* yang disimpan pada skenario dalam bentuk *tree*, memanfaatkan fitur NoSQL. Basis data yang digunakan pada program prediksi kategori IP adalah NoSQL yang dapat menyimpan data yang memiliki struktur dinamis [10].

Input data Skenario berupa data pelatihan dan jumlah kategori IP. Sebelum dilakukan klasifikasi, dilakukan *preprocessing,* meliputi *cleaning* untuk membersihkan data yang tidak lengkap dan *binning* terhadap atribut IP dan atribut numerik. *Binning* dilakukan pada atribut numerik non-*output* dengan jumlah bin sebesar 5, hanya jika pengguna memilih untuk melakukan *binning* pada atribut numerik tersebut. Pada atribut pasti, *binning* dilakukan dengan jumlah bin sebesar jumlah kategori IP.

C. Proses Pelatihan

Data latih yang digunakan adalah data mahasiswa angkatan 2008-2014. Sebelum melakukan pelatihan data latih, dilakukan tahap *preprocessing*, yaitu meliputi

transformasi data, *data cleaning* serta *binning* dan *threshold*. Transformasi data memodifikasi atribut-atribut kosong dari atribut yang tersedia pada data, misalnya dari atribut Sekolah "SMA Negeri 3 Yogyakarta", dapat disimpulkan bahwa statSMU adalah Negeri, katSMU adalah SMU, dan lokSMU adalah Jawa. *Data Cleaning* digunakan untuk membersihkan data dari *entry* yang memiliki *missing value* yang tidak bisa dikembalikan pada tahap transformasi data.

Binning dan *threshold* dilakukan pada skenario yang menggunakan data mahasiswa non-prestasi. Pada data mahasiswa jalur prestasi, tidak ada atribut numerik sehingga tidak dilakukan *binning* pada prediksi jalur prestasi. *Threshold* digunakan untuk membuat *child node* dari atribut numerik, *split* terbaik akan dihitung dengan mengintegrasi *threshold* di antara *unique value* dari atribut tersebut. Untuk mengetahui nilai dari suatu *split*, digunakan rumus *Information Gain* untuk algoritma C4.5 atau *Gini Impurity* untuk algoritma CART seperti pada *split* atribut non-numerik, namun dengan dua *child*, yaitu lebih besar dari *threshold* dan lebih kecil dari *threshold*.

Kategori atribut-atribut numerik (num, ver, ana, dan spa) yang dinyatakan ke dalam 5 kategori ditunjukkan dalam Tabel 2. *Binning* juga dilakukan pada setiap skenario untuk atribut ips1. Tabel 3 menunjukkan 2 kategori IP, yaitu A yang berarti IP tersebut rendah dan B yang berarti IP tersebut tinggi. Tabel 4 menunjukkan 3 kategori IP, yaitu A yang berarti IP tersebut rendah, B yang berarti IP tersebut sedang dan C yang berarti IP tersebut tinggi. Tabel 5 menunjukkan tabel kategori IP sebanyak 4 kategori, yaitu A yang berarti IP tersebut sangat rendah, B yang berarti IP tersebut rendah, C yang berarti IP tersebut tinggi dan D yang berarti IP tersebut sangat tinggi. Tabel 6 menunjukkan tabel kategori IP sebanyak 5 kategori, yaitu A yang berarti IP tersebut sangat rendah, B yang berarti IP tersebut rendah, C yang berarti IP tersebut sedang, D yang berarti IP tersebut tinggi dan E yang berarti IP tersebut sangat tinggi.

Setelah proses *preprocessing*, program membuat *decision tree* berdasarkan data latih tersebut. *Decision tree* tersebut digunakan untuk melakukan pengujian terhadap data uji. Data uji harus memiliki atribut-atribut yang sama dengan data latih.

D. Proses Pengujian

Data uji yang akan digunakan adalah data mahasiswa angkatan 2015. Tabel 7 menunjukkan 24 skenario pengujian dalam penelitian. Skenario pengujian dibagi berdasarkan jalur penerimaan mahasiswa, yaitu jalur prestasi sebanyak 8 skenario dan jalur non-prestasi sebanyak 16 skenario. Skenario jalur prestasi dibagi berdasarkan keseimbangan data, yaitu jalur prestasi data seimbang sebanyak 4 skenario dan jalur prestasi data tidak seimbang sebanyak 4 skenario.

Skenario jalur non-prestasi dibagi berdasarkan keseimbangan data dan kategorisasi atribut numerik. Pembagian tersebut adalah data seimbang dengan atribut numerik yang dilakukan *threshold* sebanyak 4 skenario, data seimbang dengan atribut numerik yang dilakukan *binning* sebanyak 4 skenario, data tidak

Tabel 2. Kategori atribut numerik sebanyak 5 kategori

Nilai	Kategori	Keterangan
0 – 40	A	Sangat Rendah
41 – 80	B	Rendah
81 – 120	C	Sedang
121 – 160	D	Tinggi
161 – 200	E	Sangat Tinggi

Tabel 3. Kategori indeks prestasi sebanyak 2 kategori

Indeks Prestasi	Kategori IP	Keterangan
0 – 2	A	Rendah
2,1 – 4	B	Tinggi

Tabel 4. Kategori indeks prestasi sebanyak 3 kategori

Indeks Prestasi	Kategori IP	Keterangan
0 – 1.33	A	Rendah
1.34 – 2.67	B	Sedang
2.68 – 4	C	Tinggi

Tabel 5. Kategori indeks prestasi sebanyak 4 kategori

Indeks Prestasi	Kategori IP	Keterangan
0 – 1	A	Sangat Rendah
1.1 – 2	B	Rendah
2.1 – 3	C	Tinggi
3.1 – 4	D	Sangat Tinggi

Tabel 6. Kategori indeks prestasi sebanyak 5 kategori

Indeks Prestasi	Kategori IP	Keterangan
0 – 0.8	A	Sangat Rendah
0.9 – 1.6	B	Rendah
1.7 – 2.4	C	Sedang
2.5 – 3.2	D	Tinggi
3.3 – 4	E	Sangat Tinggi

Tabel 7. Rangkuman pengujian

Nama Skenario	Binning	Metode		Data Latih		Kategori IP
		C4.5	CART	Seimbang	Tidak Seimbang	
Skenario 1	Tidak	V	V	V		2
Skenario 2	Tidak	V	V		V	2
Skenario 3	Tidak	V	V	V		3
Skenario 4	Tidak	V	V		V	3
Skenario 5	Tidak	V	V	V		4
Skenario 6	Tidak	V	V		V	4
Skenario 7	Tidak	V	V	V		5
Skenario 8	Tidak	V	V		V	5
Skenario 9	Tidak	V	V	V		2
Skenario 10	Tidak	V	V		V	2
Skenario 11	Tidak	V	V	V		3
Skenario 12	Tidak	V	V		V	3
Skenario 13	Tidak	V	V	V		4
Skenario 14	Tidak	V	V		V	4
Skenario 15	Tidak	V	V	V		5
Skenario 16	Tidak	V	V		V	5
Skenario 17	Ya	V	V	V		2
Skenario 18	Ya	V	V		V	2
Skenario 19	Ya	V	V	V		3
Skenario 20	Ya	V	V		V	3
Skenario 21	Ya	V	V	V		4
Skenrio 22	Ya	V	V		V	4
Skenario 23	Ya	V	V	V		5
Skenario 24	Ya	V	V		V	5

Tabel 8. Tabel tabulasi silang

	B_1	B_2	Total
A_1	P_{11}	P_{12}	P_{1+}
A_2	P_{21}	P_{22}	P_{2+}
Total	P_{+1}	P_{+2}	n

Keterangan:
A dan B = Variabel $P_{1+} = P_{11} + P_{12}$
$P_{+1} = P_{11} + P_{21}$ $P_{2+} = P_{21} + P_{22}$
$P_{+2} = P_{12} + P_{22}$ $n = P_{+1} + P_{+2}$ atau $P_{1+} + P_{2+}$

seimbang dengan atribut numerik yang dilakukan *threshold* sebanyak 4 skenario dan data tidak seimbang dengan atribut numerik dilakukan *binning* sebanyak 4 kategori. Sebelum dilakukan pengujian, dilakukan proses *preprocessing* terhadap data uji dengan langkah yang sama dengan proses *preprocessing* data latih.

E. Pengukuran Akurasi Algoritma

Metode yang digunakan untuk mengukur keakuratan hasil prediksi adalah tabel tabulasi silang (*crosstab*). Tabulasi silang digunakan untuk menghitung akurasi dari algoritma C4.5 dan algoritma CART, seperti ditunjukkan dalam Tabel 8. Nilai akurasi hasil prediksi dinyatakan dalam Persamaan 1.

$$Kecocokan(\%) = \frac{P_{11} + P_{22}}{n} \quad (1)$$

III. Hasil dan Pembahasan

Hasil proses *preprocessing* dari semua data latih dan data uji pada tiap skenario ditunjukkan dalam Tabel 9. Skenario nomor ganjil merupakan skenario dengan data seimbang. Data seimbang cenderung berjumlah kecil karena terdapat selisih yang besar antara satu kategori indeks prestasi dengan kategori indeks prestasi lainnya.

Rangkuman hasil pengujian skenario 1 sampai skenario 24 dan parameternya ditunjukkan dalam Tabel 10. Akurasi rata-rata untuk algoritma C4.5 adalah 41,48% dan algoritma CART adalah 42,65%. Dari tabel 10, diketahui bahwa kedua algoritma, C4.5 dan CART, dapat mencapai akurasi sebesar 86,86% (skenario 2). Akurasi tersebut dihasilkan dari data jalur prestasi yang tidak seimbang dengan kategori indeks prestasi sebanyak 2 kategori. Untuk data jalur non-prestasi, akurasi terbaik untuk algoritma C4.5 sebesar 61,54% (skenario 18) dimana data latih tidak seimbang dan atribut numerik dikategorikan, sedangkan akurasi terbaik untuk algoritma CART sebesar 63,16% (skenario 10) dimana data latih tidak seimbang dan atribut numerik tidak dikategorikan.

Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur prestasi tanpa *binning* ditunjukkan dalam Tabel 11. Akurasi rata-rata algoritma C4.5 adalah 52,84% dan algoritma CART adalah 53,52%. Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur non-prestasi ditunjukkan dalam Tabel 12. Akurasi rata-rata untuk algoritma C4.5 adalah 35,35% dan algoritma CART adalah 37,38%.

Tabel 9. Total data latih dan data uji sebelum dan setelah *preprocessing*

Nama Skenario	Total Data Latih		Total Data Uji	
	Sebelum	Setelah	Sebelum	Setelah
Skenario 1	158	158	193	137
Skenario 2	613	554	193	137
Skenario 3	120	120	193	137
Skenario 4	613	554	193	137
Skenario 5	128	128	193	137
Skenario 6	613	554	193	137
Skenario 7	125	125	193	137
Skenario 8	613	554	193	137
Skenario 9	558	558	61	57
Skenario 10	1150	1089	61	57
Skenario 11	465	465	61	57
Skenario 12	1150	1089	61	57
Skenario 13	420	420	61	57
Skenario 14	1150	1089	61	57
Skenario 15	395	395	61	57
Skenario 16	1150	1089	61	57
Skenario 17	588	588	61	57
Skenario 18	1150	1089	61	57
Skenario 19	465	465	61	57
Skenario 20	1150	1089	61	57
Skenario 21	420	420	61	57
Skenario 22	1150	1089	61	57
Skenario 23	395	395	61	57
Skenario 24	1150	1089	61	57

Tabel 10. Akurasi algoritma C4.5 dan CART untuk semua skenario

Nama Skenario	Atrr Numerik	Akurasi (%)		Data Latih		Kategori IP
		C4.5	CART	Seimbang	Tidak Seimbang	
Skenario 1	-	57,66	57,66	V		2
Skenario 2	-	86,86	86,86		V	2
Skenario 3	-	40,15	37,23	V		3
Skenario 4	-	63,50	62,77		V	3
Skenario 5	-	43,07	40,88	V		4
Skenario 6	-	66,42	68,61		V	4
Skenario 7	-	20,44	18,98	V		5
Skenario 8	-	51,82	52,55		V	5
Skenario 9	Threshold	45,61	56,14	V		2
Skenario 10	Threshold	52,63	63,16		V	2
Skenario 11	Threshold	33,33	26,32	V		3
Skenario 12	Threshold	40,35	49,12		V	3
Skenario 13	Threshold	31,58	31,58	V		4
Skenario 14	Threshold	24,56	28,07		V	4
Skenario 15	Threshold	22,81	26,32	V		5
Skenario 16	Threshold	28,07	36,60		V	5
Skenario 17	Binning	57,41	56,14	V		2
Skenario 18	Binning	61,54	57,89		V	2
Skenario 19	Binning	31,37	29,82	V		3
Skenario 20	Binning	28,30	42,11		V	3
Skenario 21	Binning	26,42	24,56	V		4
Skenario 22	Binning	35,29	22,81		V	4
Skenario 23	Binning	18,87	14,04	V		5
Skenario 24	Binning	27,45	33,33		V	5
	Rata-rata	**41,48**	**42,65**			

Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur prestasi dan non-prestasi dengan kategori indeks prestasi sebanyak 2 kategori dinyatakan dalam Tabel 13. Akurasi rata-rata untuk algoritma C4.5 adalah 60,29% dan algoritma CART adalah 62,98%.

Tabel 11. Akurasi algoritma C4.5 dan CART pada kasus penerimaan mahasiswa jalur prestasi

Nama Skenario	Data Latih		Akurasi (%)		Kategori IP
	Seimbang	Tidak Seimbang	C4.5	CART	
Skenario 1	V		57,66	57,66	2
Skenario 2		V	86,86	86,86	2
Skenario 3	V		40,15	37,23	3
Skenario 4		V	66,42	68,61	3
Skenario 5	V		43,07	40,88	4
Skenario 6		V	63,50	62,77	4
Skenario 7	V		20,44	18,98	5
Skenario 8		V	51,82	52,55	5
		Rata-rata	**52,84**	**53,52**	

Tabel 12. Akurasi algoritma C4.5 dan CART pada kasus penerimaan mahasiswa jalur non-prestasi.

Nama Skenario	Data Latih		Akurasi (%)		Kategori IP
	Seimbang	Tidak Seimbang	C4.5	CART	
Skenario 9	V		45,61	56,14	2
Skenario 10		V	52,63	63,16	2
Skenario 11	V		33,33	26,32	3
Skenario 12		V	40,35	49,12	3
Skenario 13	V		31,58	31,58	4
Skenario 14		V	24,56	28,07	4
Skenario 15	V		22,81	26,32	5
Skenario 16		V	28,07	36,60	5
Skenario 17	V		57,41	56,14	2
Skenario 18		V	61,54	57,89	2
Skenario 19	V		31,37	29,82	3
Skenario 20		V	28,30	42,11	3
Skenario 21	V		26,42	24,56	4
Skenario 22		V	35,29	22,81	4
Skenario 23	V		18,87	14,04	5
Skenario 24		V	27,45	33,33	5
		Rata-rata	**35,35**	**37,38**	

Tabel 13. Akurasi algoritma C4.5 dan CART dengan kategori indeks prestasi sebanyak 2 kategori

Nama Skenario	Jalur Penerimaan	Atribut Numerik	Akurasi (%)		Data Latih Seimbang
			C4.5	CART	
Skenario 1	Jalur Prestasi	-	57,66	57,66	V
Skenario 2	Jalur Prestasi	-	86,86	86,86	X
Skenario 9	Jalur Non-Prestasi	Threshold	45,61	56,14	V
Skenario 10	Jalur Non-Prestasi	Threshold	52,63	63,16	X
Skenario 17	Jalur Non-Prestasi	Binning	57,41	56,14	V
Skenario 18	Jalur Non-Prestasi	Binning	61,54	57,89	X
		Rata-rata	**60,29**	**62,98**	

Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur prestasi dan non-prestasi dengan indeks prestasi sebanyak 3 kategori dinyatakan dalam Tabel 14. Akurasi rata-rata untuk algoritma C4.5 adalah 39,99% dan algoritma CART adalah 42,20%.

Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur prestasi dan non-prestasi dengan indeks prestasi sebanyak 4 kategori dinyatakan dalam Tabel 15. Akurasi rata-rata untuk algoritma C4.5 adalah 37,40% dan algoritma CART adalah 35,11%.

Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur prestasi dan non-prestasi dengan indeks prestasi sebanyak 5 kategori dinyatakan dalam Tabel 16. Akurasi rata-rata untuk algoritma C4.5 adalah 28,24% dan algoritma CART adalah 30,30%.

Tabel 14. Akurasi algoritma C4.5 dan CART dengan kategori indeks prestasi sebanyak 3 kategori

Nama Skenario	Jalur Penerimaan	Atribut Numerik	Akurasi (%)		Data Latih Seimbang
			C4.5	CART	
Skenario 3	Jalur Prestasi	-	40,15	37,23	V
Skenario 4	Jalur Prestasi	-	66,42	68,61	X
Skenario 11	Jalur Non-Prestasi	Threshold	33,33	26,32	V
Skenario 12	Jalur Non-Prestasi	Threshold	40,35	49,12	X
Skenario 19	Jalur Non-Prestasi	Binning	31,37	29,82	V
Skenario 20	Jalur Non-Prestasi	Binning	28,30	42,11	X
		Rata-rata	**39,99**	**42,20**	

Tabel 15. Akurasi algoritma C4.5 dan CART dengan kategori indeks prestasi sebanyak 4 kategori

Nama Skenario	Jalur Penerimaan	Atribut Numerik	Akurasi (%)		Data Latih Seimbang
			C4.5	CART	
Skenario 5	Jalur Prestasi	-	43,07	40,88	V
Skenario 6	Jalur Prestasi	-	63,50	62,77	X
Skenario 13	Jalur Non-Prestasi	Threshold	31,58	31,58	V
Skenario 14	Jalur Non-Prestasi	Threshold	24,56	28,07	X
Skenario 21	Jalur Non-Prestasi	Binning	26,42	24,56	V
Skenario 22	Jalur Non-Prestasi	Binning	35,29	22,81	X
		Rata-rata	**37,40**	**35,11**	

Tabel 16. Akurasi algoritma C4.5 dan CART dengan kategori indeks prestasi sebanyak 5 kategori

Nama Skenario	Jalur Penerimaan	Atribut Numerik	Akurasi (%)		Data Latih Seimbang
			C4.5	CART	
Skenario 7	Jalur Prestasi	-	20,44	18,98	V
Skenario 8	Jalur Prestasi	-	51,82	52,55	X
Skenario 15	Jalur Non-Prestasi	Threshold	22,81	26,32	V
Skenario 16	Jalur Non-Prestasi	Threshold	28,07	36,60	X
Skenario 23	Jalur Non-Prestasi	Binning	18,87	14,04	V
Skenario 24	Jalur Non-Prestasi	Binning	27,45	33,33	X
		Rata-rata	**28,24**	**30,30**	

Tabel 17. Akurasi algoritma C4.5 dan CART pada kasus penerimaan mahasiswa jalur non-prestasi dengan atribut numerik di-*threshold*

Nama Skenario	Akurasi (%)		Seimbang	Tidak Seimbang	Kategori IP
	C4.5	CART			
Skenario 9	45,61	56,14	V		2
Skenario 10	52,63	63,16		V	2
Skenario 11	33,33	26,32	V		3
Skenario 12	40,35	49,12		V	3
Skenario 13	31,58	31,58	V		4
Skenario 14	24,56	28,07		V	4
Skenario 15	22,81	26,32	V		5
Skenario 16	28,07	36,60		V	5
Rata-rata	**34,87**	**39,66**			

Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur non-prestasi dengan atribut numerik di-*threshold* dinyatakan dalam Tabel 17. Akurasi rata-rata untuk algoritma C4.5 adalah 37,37% dan algoritma CART adalah 41,00%. Rangkuman hasil pengujian untuk kasus penerimaan mahasiswa melalui jalur non-prestasi dengan atribut numerik di-*binning* dinyatakan dalam Tabel 18. Akurasi rata-rata untuk algoritma C4.5 adalah 35,83% dan algoritma CART sebesar 35,09%.

Tabel 18. Akurasi algoritma C4.5 dan CART pada kasus penerimaan mahasiswa jalur non-prestasi dengan atribut numerik di-*binning*

Nama Skenario	Akurasi (%)		Seimbang	Tidak Seimbang	Kategori IP
	C4.5	CART			
Skenario 17	57,41	56,14	V		2
Skenario 18	61,54	57,89		V	2
Skenario 19	31,37	29,82	V		3
Skenario 20	28,30	42,11		V	3
Skenario 21	26,42	24,56	V		4
Skenario 22	35,29	22,81		V	4
Skenario 23	18,87	14,04	V		5
Skenario 24	27,45	33,33		V	5
Rata-rata	**35,83**	**35,09**			

Tabel 19. Akurasi algoritma C4.5 dan CART dengan data latih yang seimbang

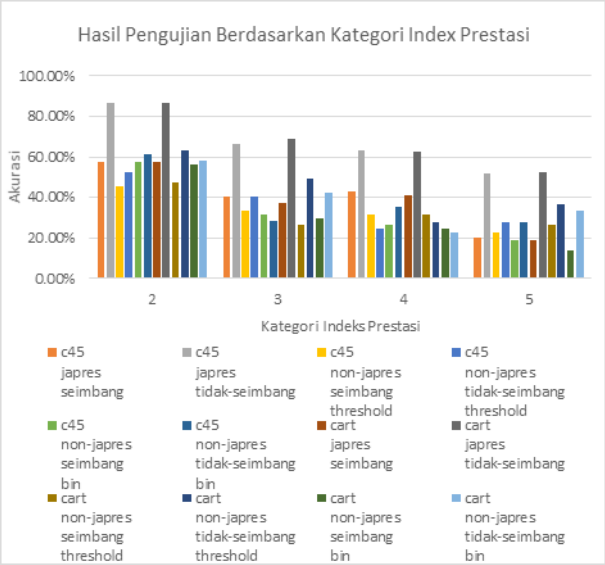
Nama Skenario	Jalur Penerimaan	Atribut Numerik	Akurasi (%)		Kategori IP
			C4.5	CART	
Skenario 1	Japres	-	57,66	57,66	2
Skenario 3	Japres	-	40,15	37,23	3
Skenario 5	Japres	-	43,07	40,88	4
Skenario 7	Japres	-	20,44	18,98	5
Skenario 9	Non-Japres	Threshold	45,61	56,14	2
Skenario 11	Non-Japres	Threshold	33,33	26,32	3
Skenario 13	Non-Japres	Threshold	31,58	31,58	4
Skenario 15	Non-Japres	Threshold	22,81	26,32	5
Skenario 17	Non-Japres	Binning	57,41	56,14	2
Skenario 19	Non-Japres	Binning	31,37	29,82	3
Skenario 21	Non-Japres	Binning	26,42	24,56	4
Skenario 23	Non-Japres	Binning	18,87	14,04	5
		Rata-rata	**35,73**	**34,97**	

Tabel 20. Akurasi algoritma C4.5 dan CART dengan data latih yang tidak seimbang.

Nama Skenario	Jalur Penerimaan	Atribut Numerik	Akurasi (%)		Kategori IP
			C4.5	CART	
Skenario 2	Japres	-	86,86	86,86	2
Skenario 4	Japres	-	66,42	68,61	3
Skenario 6	Japres	-	63,50	62,77	4
Skenario 8	Japres	-	51,82	52,55	5
Skenario 10	Non-Japres	Threshold	52,63	63,16	2
Skenario 12	Non-Japres	Threshold	40,35	49,12	3
Skenario 14	Non-Japres	Threshold	24,56	28,07	4
Skenario 16	Non-Japres	Threshold	28,07	36,60	5
Skenario 18	Non-Japres	Binning	61,54	57,89	2
Skenario 20	Non-Japres	Binning	28,30	42,11	3
Skenario 22	Non-Japres	Binning	35,29	22,81	4
Skenario 24	Non-Japres	Binning	27,45	33,33	5
		Rata-rata	**47,23**	**50,32**	

Rangkuman hasil pengujian dengan data latih yang seimbang dinyatakan dalam Tabel 19. Akurasi rata-rata untuk algoritma C4.5 adalah 35,73% dan algoritma CART adalah 34,97%. Rangkuman hasil pengujian dengan data latih yang tidak seimbang dinyatakan dalam Tabel 20. Akurasi rata-rata untuk algoritma C4.5 adalah 47,23% dan algoritma CART adalah 50,32%.

Berdasarkan kategorisasi atribut numerik untuk data jalur non-prestasi, algoritma C4.5 dengan atribut numerik di-*binning* memiliki akurasi rata-rata yang lebih baik, yaitu sebesar 35,83%, dibandingkan dengan atribut numerik yang di-*threshold*, yaitu sebesar 34,87%. Algoritma CART dengan atribut numerik di-*threshold* memiliki akurasi rata-rata yang lebih baik, yaitu sebesar 39,66% dibandingkan dengan atribut



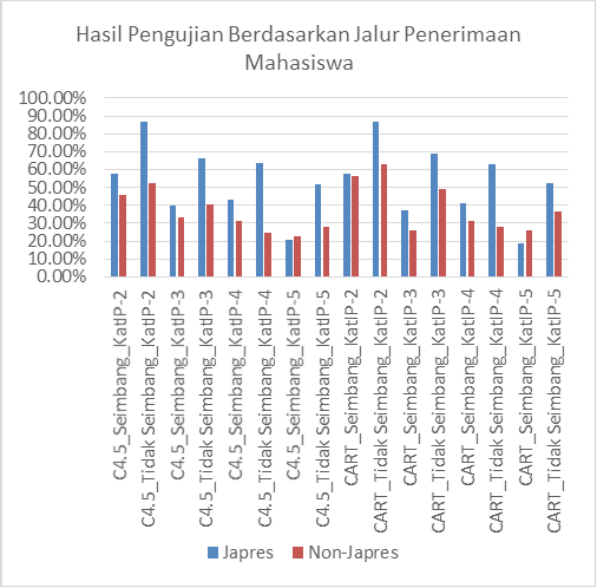
Gambar 2. Hasil pengujian berdasarkan kategori indeks prestasi

numerik yang di-*binning*, yaitu sebesar 35,09%. Berdasarkan keseimbangan data, data yang tidak seimbang memiliki akurasi rata-rata yang baik, yaitu C4.5 memiliki akurasi rata-rata sebesar 47,23% dan CART memiliki akurasi rata-rata sebesar 50,32%, dibandingkan dengan data yang seimbang.

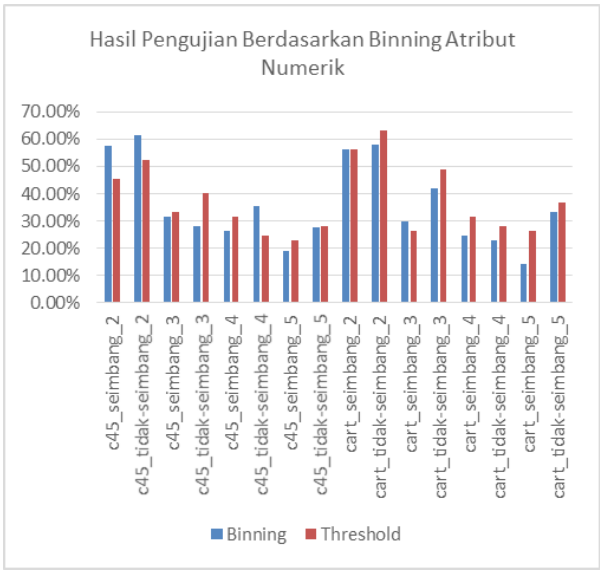
Secara keseluruhan, akurasi rata-rata algoritma CART lebih tinggi daripada akurasi rata-rata algoritma C4.5. Akurasi rata-rata algoritma CART sebesar 42,65% sedangkan akurasi rata-rata algoritma C4.5 sebesar 41,48%. Akurasi rata-rata kedua algoritma tersebut sangat dipengaruhi oleh faktor-faktor seperti jumlah kategori indeks prestasi, keseimbangan data, kategorisasi (*binning*) untuk atribut numerik, dan kualitas data itu sendiri.

Perbandingan akurasi algoritma terhadap jumlah_kategori_ip ditunjukkan dalam Gambar 2. Semakin banyak kategori IP, semakin kecil akurasi algoritma. Grafik tersebut juga menunjukkan bahwa prediksi jalur prestasi (japres) data tidak seimbang di-*threshold* menunjukkan nilai akurasi lebih tinggi dibandingkan dengan data lain.

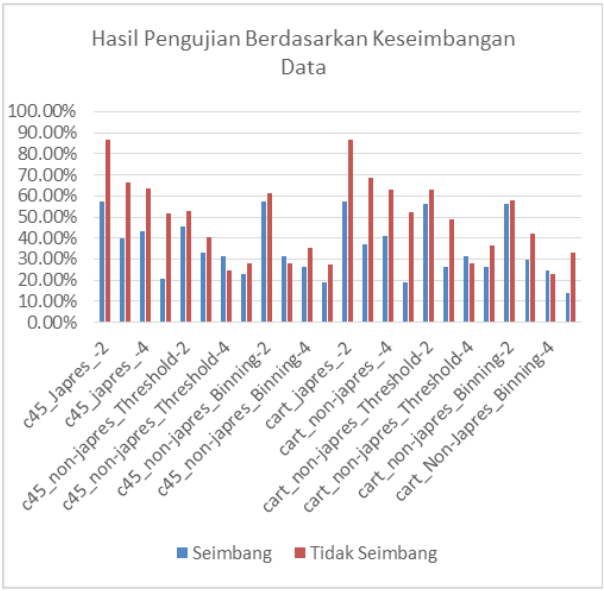
Gambar 3 menunjukkan perbandingan hasil pengujian berdasarkan jalur penerimaan mahasiswa, yaitu jalur prestasi (japres) dan jalur non-prestasi (non-japres) yang di-*threshold*. Pada beberapa skenario, data japres menghasilkan akurasi yang jauh lebih tinggi daripada data non-japres dengan atribut lain yang sama, terutama pada jumlah kategori indeks prestasi yang kecil. Pada beberapa skenario lainnya, data japres menghasilkan akurasi yang tidak jauh lebih tinggi dari data non-japres dan bahkan pada beberapa skenario data non-japres menghasilkan akurasi lebih tinggi daripada data japres. Ada kesamaan pola perbandingan akurasi data japres dan data non-japres pada metode C4.5 dan CART. Hanya pada skenario kategori indeks prestasi 5 dengan data seimbang, akurasi C4.5 untuk data japres lebih tinggi daripada data non-japres, sedangkan akurasi



Gambar 3. Hasil pengujian berdasarkan jalur penerimaan mahasiswa



Gambar 4. Grafik hasil pengujian berdasarkan *binning* atribut numerik



Gambar 5. Grafik hasil pengujian berdasarkan keseimbangan data

CART untuk data japres lebih rendah daripada data non-japres.

Gambar 4 menunjukkan perbandingan hasil pengujian berdasarkan pengkategorian atribut numerik pada kasus penerimaan mahasiswa pada jalur non-prestasi, yaitu atribut numerik yang di-*binning* dan atribut numerik yang di-*threshold*. Dari grafik tersebut dapat dilihat bahwa tidak ada pola dampak *binning* terhadap akurasi. *Binning* meningkatkan akurasi secara signifikan pada saat jumlah kategori indeks prestasi 2, menurunkan akurasi secara signifikan pada saat jumlah kategori indeks prestasi 3 dan 4, dan tidak berdampak secara signifikan pada saat jumlah kategori indeks prestasi 5. Pola tersebut terlihat pada algoritma C4.5 dan CART.

Gambar 5 menunjukkan grafik perbandingan hasil pengujian berdasarkan keseimbangan data, yaitu data latih yang seimbang dan data latih yang tidak seimbang. Dari grafik tersebut dapat dilihat bahwa 20 dari 24 pasang skenario dengan parameter yang sama diprediksi menggunakan data latih yang sudah diseimbangkan memiliki akurasi yang lebih rendah dibandingkan dengan prediksi menggunakan data latih yang tidak diseimbangkan. Sebanyak 6 dari 20 pasang skenario tersebut memiliki selisih akurasi lebih dari 20%. Sebanyak 4 dari 24 pasang skenario diprediksi menggunakan data latih yang diseimbangkan memiliki akurasi yang sedikit lebih tinggi dibandingkan dengan menggunakan data latih yang tidak diseimbangkan.

Dari hasil penelitian ini, didapatkan bahwa akurasi terbaik algoritma C4.5 dan CART pada data jalur prestasi lebih tinggi dari data jalur non-prestasi. Pada data jalur prestasi, akurasi terbaik kedua algoritma tersebut mencapai 86,86%, sedangkan pada data jalur non-prestasi, akurasi terbaik algoritma C4.5 adalah 61,54% dan CART adalah 63,16%.

Dari hasil pengujian kedua algoritma tersebut dapat dinyatakan bahwa akurasi terbaik algoritma CART pada data jalur non-prestasi (data numerik) adalah 63,16%. Jika dibandingkan dengan [8] yang menyatakan bahwa algoritma CART cocok digunakan untuk data numerik daripada algoritma C4.5, maka dapat dinyatakan juga bahwa algoritma CART memberikan akurasi yang lebih baik daripada algoritma C4.5 pada data numerik.

Ketika data yang digunakan adalah data jalur prestasi (data non-numerik), dapat dinyatakan bahwa algoritma C4.5 dan CART memiliki akurasi yang sama, yaitu 86,86%. Jika dibandingkan dengan [8] yang menyatakan bahwa algoritma C4.5 memberikan akurasi yang lebih baik daripada algoritma CART pada data non-numerik dan [9] yang menyatakan bahwa akurasi algoritma CART lebih baik daripada algoritma C4.5 pada data non-numerik, maka dapat dinyatakan bahwa hal tersebut bertentangan dengan hasil penelitian yang

Tabel 21. Perbandingan akurasi algoritma C4.5, CART, regresi logistik, K-Nearest Neighbor, dan Näive Bayes Classifier.

Algoritma/Metode	Jalur Penerimaan	
	Jalur Prestasi	Jalur Non-Prestasi
C4.5	86.86%	61.54%
CART	86.86%	63.16%
Regresi Logistik [1], [5]	72.00%	60.70%
K-Nearest Neighbor [6]	61.00%	52.00%
Naïve Bayes Classifier [6]	55.00%	41.00%

menyatakan bahwa algoritma C4.5 dan CART memiliki akurasi yang sama pada data non-numerik.

Perbandingan algoritma C4.5 dan CART dengan algoritma lain ditunjukkan dalam Tabel 21, yaitu regresi logistik untuk jalur prestasi [1] dan non-prestasi [5] serta metode K-Nearest Neighbor (k=2) dan Naive Bayes Classifier [6]. Algoritma yang paling baik untuk memprediksi kategori indeks prestasi mahasiswa untuk jalur prestasi adalah algoritma C4.5 dan CART dengan akurasi sebesar 86,86% dan algoritma yang paling baik untuk memprediksi kategori indeks prestasi mahasiswa untuk jalur non-prestasi adalah algoritma CART dengan akurasi sebesar 63,16%. Metode K-Nearest Neighbor menggunakan nilai k = 5 dan k = 7. Rasio data uji dan data latih yang digunakan menentukan akurasi prediksi, seperti jika dibandingkan dengan akurasi terbaik 92,79% dalam Untari [3] yang menggunakan C4.5 dan rasio data latih berbeda-beda (10%-90%) dan dengan akurasi 80,87% dalam Aprilia dkk. [7] yang menggunakan CART dan rasio data latih 84%. Penelitian ini menggunakan rasio data latih tetap, yaitu sebesar 76,05% dan akurasi terbaik diukur menggunakan data tidak seimbang.

Permasalahan pada penelitian ini adalah adanya *missing value* pada saat pengujian yang diatasi dengan cara *discard*. Penggunaan cara lain untuk mengatasi *missing value* dapat dilakukan dengan *imputation*, C4.5 *strategy*, *null strategy*, dan *lazy decision tree* seperti yang disajikan dalam [11]. Prediksi indeks prestasi (IP) mahasiswa semester satu yang dilakukan secara dini ini dapat digunakan untuk menanggulangi masalah-masalah yang mungkin akan dialami oleh mahasiswa di kemudian hari. Fakultas bisa melacak mahasiswa-mahasiswa yang berpotensi memiliki IP semester satu yang rendah untuk dilakukan pembimbingan lebih lanjut terhadap mahasiswa tersebut. Selain itu, prediksi juga bisa dimanfaatkan untuk melakukan penyaringan calon mahasiswa baru.

IV. KESIMPULAN

Algoritma C4.5 dan CART memiliki akurasi yang sama untuk memprediksi kategori IP mahasiswa baru pada jalur prestasi (data non-numerik), yaitu sebesar 86,86%. Untuk memprediksi kategori IP mahasiswa baru pada jalur non-prestasi (data numerik), algoritma CART memberikan akurasi lebih baik daripada C4.5, yaitu 63,16% berbanding 61,54%.

DAFTAR PUSTAKA

[1] R. G. Santosa and A. R. Chrismanto, "Logistic Regression Model for Predicting First Semester Students Gpa Category Based on High School Academic Achievement," *Journal of Arts, Science & Commerce*, vol. VIII, no. 2(1), pp. 58–66, April 2017.

[2] D. Indriana, A. I. Widowati, and S. Surjawati,"Faktor-Faktor yang Mempengaruhi Prestasi Akademik: Studi Kasus pada Mahasiswa Program Studi Akuntansi Universitas Semarang," *Jurnal Dinamika Sosial Budaya*, vol. 18, no. 1, pp. 39-48, Juni 2016.

[3] D. Untari, "Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5," Skripsi, Universitas Dian Nuswantoro, Semarang, 2014 [Online]. Available: http://eprints.dinus.ac.id/13181/

[4] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *Ultimatics,* vol. VI, no. 1, pp. 15–20, Juni 2014.

[5] R. G. Santosa and A. R. Chrismanto, "Perbandingan Akurasi Model Regresi Logistik Untuk Prediksi Kategori IP Mahasiswa Jalur Prestasi dengan Non Jalur Prestasi," *Jurnal Teknik dan Ilmu Komputer*, vol. 7, no. 25, pp. 107–121, Januari 2018.

[6] V. H. A. Sari, R. G. Santosa, and A. Rachmat, "Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier dalam Memprediksi Kategori Indeks Prestasi Mahasiswa," Laporan tidak dipublikasi, 2017.

[7] T. Aprilia, N. Gusriani, and K. Parmikanti, "Klasifikasi Ketepatan Masa Studi Mahasiswa FMIPA Unpad Angkatan 2001-2006 dengna Menggunakan Metode Classification and Regression Trees (CART)," *Jurnal Matematika Integratif,* vol. 11, no. 1, pp. 7-14, April 2015

[8] I. Rahmayuni, "Perbandingan Performansi Algoritma C4.5 dan CART Dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang," *Jurnal TEKNOIF*, vol. 2, no. 1, pp. 40–46, April 2014.

[9] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," *International Journal of Modern Education and Computer Science*, vol. 5, no. 5, pp. 18–27, 2013.

[10] C. Brooks, *Entreprise NoSQL For Dummies*. John Wiley & Sons, Inc, 2014.

[11] S. Gavankar and S. Sawarkar, "Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility," in *Proc. of 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS 2015)*, Malaysia, Dec. 2-4, 2015, pp. 122–126.