Perbandingan Algoritma C4.5 dan Classification and Regression Tree (CART) Dalam Menyeleksi Calon Karyawan

Ng Poi Wong[1], Florida N. S. Damanik[2], Christine[3], Edward Surya Jaya[4], Ryan Rajaya[5]
Program Studi Teknik Informatika, STMIK Mikroskil
[1]poiwong@mikroskil.ac.id, [2]florida@mikroskil.ac.id, [3]141110400@students.mikroskil.ac.id,
[4]141111235@students.mikroskil.ac.id, [5]141113279@students.mikroskil.ac.id

Abstrak

Penelitian ini membandingkan keakuratan dari algoritma C4.5 dan Classification and Regression Tree (CART) dalam menyeleksi calon karyawan pada perusahaan. Penelitian ini menggunakan dataset dengan kriteria berupa umur, lama pengalaman kerja, pendidikan terakhir, status pernikahan, jumlah kemampuan yang dimiliki, serta nilai tes seleksi masuk. Pengujian menggunakan 200 data seleksi calon karyawan secara manual dari suatu perusahaan. Pengujian dilakukan dengan menggunakan K-Fold Cross Validation dan menghitung keakuratan algoritma dengan Confusion Matrix. Algoritma C4.5 memiliki tingkat akurasi, tingkat keberhasilan sistem, dan tingkat ketepatan hasil keputusan sebesar 52,83%, 41,48%, dan 43,98%, sedangkan algoritma CART sebesar 53,33%, 44,06%, dan 42,81%.

Kata Kunci : *Akurasi, C4.5, Classification and Regression Tree (CART)*

Abstract

This research compares the accuracy of the C4.5 algorithm and Classification and Regression Tree (CART) for prospective employees selection in companies. This research using dataset with criteria like age, working experience, recent education, marital status, number of abilities possessed, and the result of admission selection test. Testing uses 200 prospective employee selection data manually from a company. Algorithm testing using K-Fold Cross Validation and the accuracy calculation of the algorithm using Confusion Matrix. C4.5 algorithm has a level of accuracy, the success rate of the system, and the level of accuracy of the decision results of 52,83%, 41,48% and 43,98%, and CART algorithm is 53,33%, 44,06%, and 42,81%.

Keyword : *Accuracy, C4.5, Classification and Regression Tree (CART)*

1. PENDAHULUAN

Data mining merupakan suatu proses ekstraksi informasi spesifik dari data dan menyajikan informasi yang relevan dan berguna sehingga dapat dimanfaatkan untuk memecahkan berbagai masalah [1], salah satunya adalah analisis klasifikasi. Analisis klasifikasi digunakan untuk mengambil informasi penting dan relevan tentang data, dan metadata, salah satu algoritma klasifikasi adalah pohon keputusan. Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule* [2].

Terdapat beragam jenis algoritma untuk membuat pohon keputusan, salah satunya adalah algoritma ID3 yang merupakan algoritma dasar dalam membentuk pohon keputusan, dan pengembangan dari algoritma ID3 yang paling sering diterapkan seperti algoritma C4.5 dan algoritma CART (*Classification and Regression Tree*) [3]. Algoritma C4.5 merupakan

pengembangan dari konsep dasar algoritma ID3, dimana pembentukan pohon dilakukan dengan menghitung nilai entropi dan *information gain*, serta nilai *gain ratio* sebagai patokan dalam membentuk akar dan simpul daun pada *decision tree* [4]. Algoritma CART mendefinisikan calon cabang kiri dan calon cabang kanan yang diikuti dengan perhitungan probabilitas setiap calon cabang serta nilai *goodness* yang menjadi patokan dalam membentuk simpul akar maupun simpul daun pada *decision tree* [5]. Proses perhitungan maupun penentuan simpul akar dan cabang dari kedua algoritma tersebut berbeda, namun keduanya menghasilkan tingkat akurasi yang cukup tinggi [6].

Pada penelitian ini dilakukan perbandingan algoritma C4.5 dan CART dengan mengambil kasus proses seleksi calon karyawan. Algoritma akan meninjau atribut-atribut yang diperhatikan dalam menyeleksi calon karyawan, kemudian atribut-atribut tersebut disusun menjadi sebuah *decision tree* yang memberikan *rule* dalam menentukan hasil seleksi dari calon karyawan. Perbandingan kedua algoritma ini diharapkan dapat memberikan solusi pada beberapa masalah perusahaan dalam seleksi calon karyawan, seperti standar kualifikasi, kemampuan yang dimiliki, perilaku, dan sebagainya.

2. METODE PENELITIAN

2.1 Seleksi Calon Karyawan

Seleksi calon karyawan merupakan proses seleksi pelamar atau pencarian calon karyawan yang dimulai ketika para pelamar dicari dan berakhir jika lamaran tersebut diterima oleh perusahaan. Hasilnya berupa sekumpulan data para pencari kerja yang siap untuk diseleksi. Adapun proses seleksi merupakan serangkaian kegiatan yang dilakukan untuk memutuskan apakah pelamar diterima atau tidak [7]. Berbagai prosedur seleksi calon karyawan dapat ditemui pada perusahaan-perusahaan, misalnya pengisian formulir lamaran, tes penerimaan, wawancara, pemeriksaan kesehatan, keputusan penerimaan, induksi dan orientasi. Aktivitas tersebut dijalankan secara berkesinambungan hingga diperoleh sumber daya manusia yang dibutuhkan dan siap menjalankan tugas serta tanggung jawab yang diberikan oleh manajemen perusahaan [8]. Sejumlah standar kualifikasi ditetapkan oleh perusahaan dalam mencari calon karyawan, misalnya umur, pengalaman kerja, pendidikan terakhir, dan sebagainya. Selain itu juga dilakukan sejumlah tes penerimaan sesuai dengan standar yang diinginkan perusahaan, misalnya tes kesehatan, psikologi, dan sebagainya.

Pada penelitian ini, jumlah data seleksi calon karyawan yang dijadikan dataset sebanyak 200 data dummy yang diambil secara acak dari suatu perusahaan, dengan standar kualifikasi terdiri dari umur, pengalaman kerja, pendidikan terakhir, status pernikahan, kemampuan yang dimiliki, hasil nilai tes ujian online dan tatap muka, nilai sikap, dan hasil keputusan seleksi calon karyawan secara manual. Tabel 1 menunjukkan rekapitulasi jumlah dataset dari masing-masing standar yang digunakan untuk pengujian.

Tabel 1. Jumlah dataset untuk pengujian

Standar	Kriteria	Jumlah Data
Umur	< 25 Tahun	52
	25 - 35 Tahun	78
	> 35 Tahun	70
	0 Tahun	69

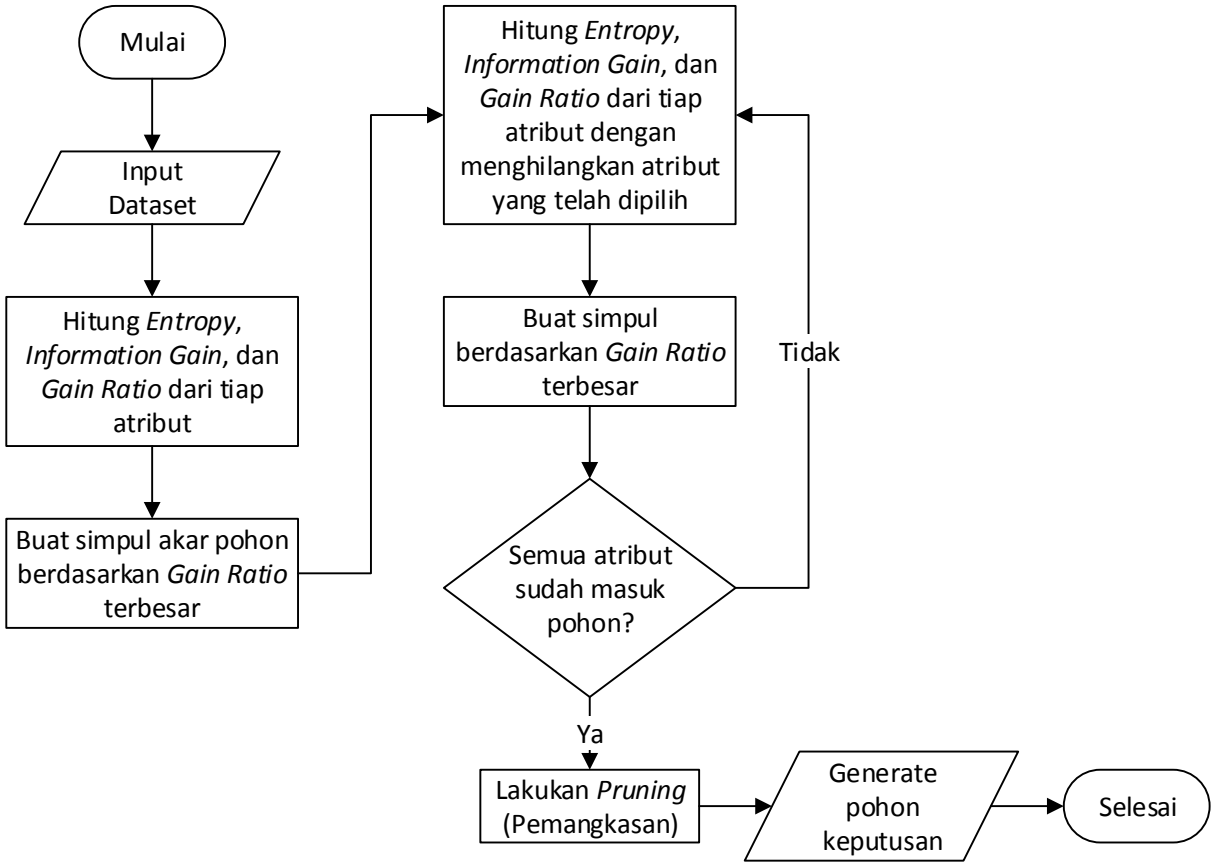
Standar	Kriteria	Jumlah Data
Kemampuan yang dimiliki	5 - 7	86
	8 - 10	114
Nilai Ujian Online	70 - 79	81
	80 - 89	54

Pengalaman Kerja	1 - 2 Tahun	57
	> 2 Tahun	74
Pendidikan Terakhir	SMA	50
	Akademi	44
	Sarjana	53
	Pasca Sarjana	53
Status Pernikahan	Lajang	113
	Menikah	87

Nilai Ujian Tatap Muka	90 - 100	65
	70 - 79	70
	80 - 89	68
	90 - 100	62
Nilai Sikap	Cukup Baik	59
	Baik	59
	Sangat Baik	82

2.2 Algoritma C4.5

Algoritma C4.5 merupakan algoritma untuk membentuk pohon keputusan menggunakan *gain ratio* untuk menyeleksi fitur [9]. Gambar 1 menggambarkan langkah kerja dari algoritma C4.5 dalam membentuk pohon keputusan untuk seleksi data calon karyawan.



Gambar 1. Langkah kerja dari Algoritma C4.5

Algoritma C4.5 menggunakan kriteria *split* yang telah dimodifikasi yang dinamakan *gain ratio* dalam proses pemilihan *split* atribut, dengan rumus :

$$gain\ ratio(a) = \frac{gain(a)}{split(a)} \quad (1)$$

dimana :
a : atribut
gain(a) : information gain pada atribut a

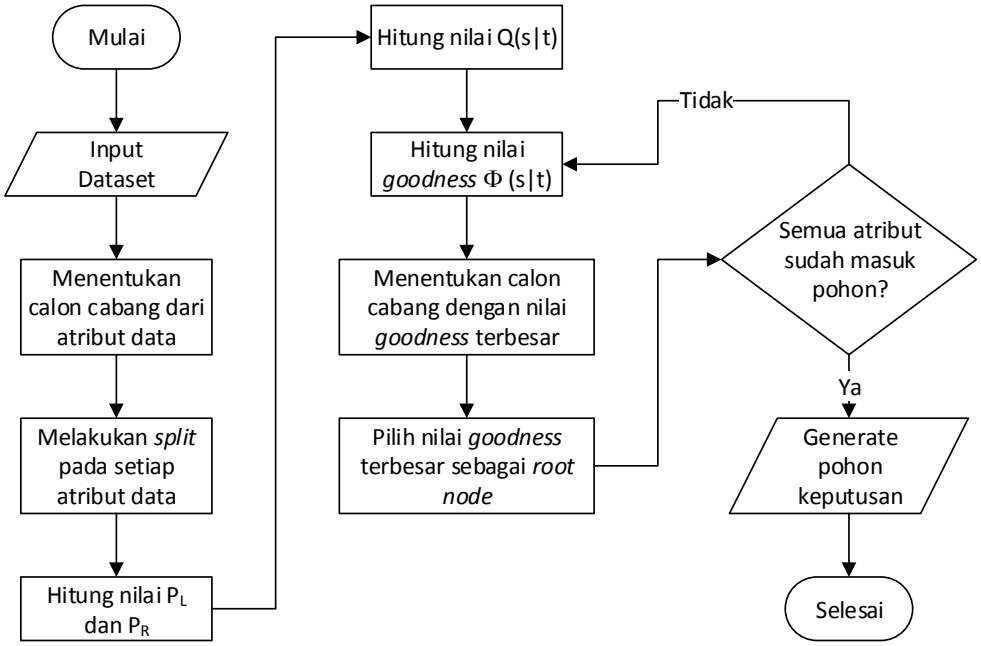
split(a) : split information pada atribut a
Atribut dengan nilai *gain ratio* tertinggi dipilih sebagai atribut test untuk simpul, dengan *gain* adalah *information gain*. Pendekatan ini menerapkan normalisasi pada *information gain* dengan menggunakan *split information*. *SplitInfo* menyatakan *entropy* atau informasi potensial dengan rumus :

$$SplitInfo(S,A) = -\sum_{i=1}^{n} \frac{Si}{S} \log_2 \frac{Si}{S} \quad (2)$$

dimana :
S : ruang (data) sampel yang digunakan.
A : atribut.
Si : jumlah sampel untuk atribut i

2.3 Algoritma CART

Algoritma CART merupakan algoritma untuk membentuk pohon keputusan dengan pembentukkan pohon klasifikasi, dimana setiap *parent node* akan mengalami pemecahan tepat menjadi dua *child node* dan setiap *child node* memiliki siklus berulang untuk menjadi *parent node* kembali. Siklus ini akan terus berulang hingga tidak ada lagi kesempatan untuk melakukan pemecahan berikutnya [10]. Gambar 2 menggambarkan langkah kerja dari algoritma CART dalam membentuk pohon keputusan untuk seleksi data calon karyawan.



Gambar 2. Langkah kerja dari Algoritma CART

Nilai *goodness* ϕ(s|t) dari calon cabang s pada noktah keputusan t, didefiniskan dengan rumus :

$$\Phi(s|t) = 2P_L P_R Q(s|t) \quad (3)$$

$$Q(s|t) = \sum_{j=1}^{jumlah\ kategori} |P(j|t_L) - P(j|t_R)| \quad (4)$$

Dimana :
tL : calon cabang kiri dari noktah keputusan t
tR : calon cabang kanan dari noktah keputusan t

$$P_L = \frac{jumlah\ catatan\ pada\ calon\ cabang\ kiri\ t_L}{jumlah\ catatan\ pada\ data\ latihan} \quad (5)$$

$$P_R = \frac{jumlah\ catatan\ pada\ calon\ cabang\ kanan\ t_R}{jumlah\ catatan\ pada\ data\ latihan} \quad (6)$$

$$P(j|t_L) = \frac{jumlah\ catatan\ berkategori\ j\ pada\ calon\ cabang\ kiri\ t_L}{jumlah\ catatan\ pada\ noktah\ keputusan\ t} \quad (7)$$

$$P(j|t_R) = \frac{jumlah\ catatan\ berkategori\ j\ pada\ calon\ cabang\ kanan\ t_R}{jumlah\ catatan\ pada\ noktah\ keputusan\ t} \quad (8)$$

2.4 K-Fold Cross Validation

Cross Validation merupakan metode statistik yang dapat digunakan untuk mengevaluasi kinerja model / algoritma dimana data dipisahkan menjadi dua subset yaitu data proses pembelajaran dan data validasi / evaluasi. Model / algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi, selanjutnya pemilihan jenis *cross-validation* dapat didasarkan pada ukuran dataset. *10-Fold Cross Validation* adalah salah satu dari *cross validation* yang direkomendasikan untuk pemilihan model terbaik karena cenderung memberikan estimasi akurasi yang lebih baik dalam pengklasifikasian Dalam *10-Fold Cross Validation*, data dibagi menjadi 10 *fold* yang berukuran sama, sehingga akan memiliki 10 subset data untuk mengevaluasi kinerja model / algoritma [11]. Pada penelitian ini, hasil seleksi calon karyawan dengan algoritma C4.5 dan CART akan diuji dengan menggunakan *10-Fold Cross Validation*.

2.5 Confusion Matrix

Confusion Matrix merupakan suatu metode yang digunakan untuk menghitung akurasi atau kinerja dari *data mining*. Terdapat 4 istilah dalam representasi hasil proses klasifikasi, yakni *True Positive* (TP) menyatakan observasi bernilai positif, dan diprediksi positif, *False Negative* (FN) menyatakan observasi bernilai positif, tetapi diprediksi negatif, *True Negative* (TN) menyatakan observasi bernilai negatif, dan diprediksi negatif, dan *False Positive* (FP) menyatakan observasi bernilai negatif, tetapi diprediksi positif. Tabel 2 menggambarkan *Confusion Matrix* dengan ke-4 istilah (TP, FN, TN, dan FP).

Tabel 2. *Confusion Matrix*

		Nilai Prediksi	
		Positive	**Negative**
Nilai Observasi	**Positive**	*True Positive* (TP)	*False Negative* (FN)
	Negative	*False Positive* (FP)	*True Negative* (TN)

Dengan ke-4 istilah tersebut, dilakukan perhitungan *Accuracy* untuk menentukan tingkat kedekatan antara nilai prediksi dengan nilai observasi, perhitungan *Recall* untuk menentukan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi, dan perhitungan *Precision* utnuk menentukan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem [12] dengan rumus :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

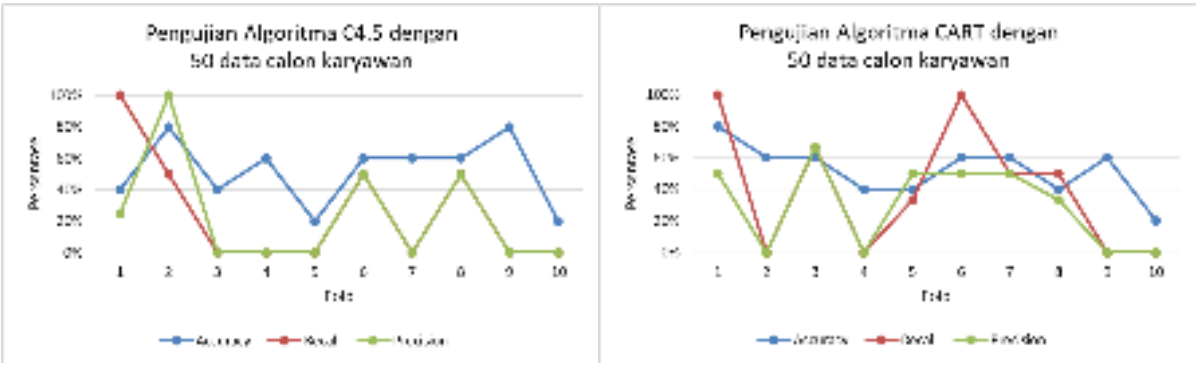
$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Pada penelitian ini, hasil pengujian data seleksi calon karyawan dengan *10-Fold Cross Validation* akan dilanjutkan menghitung tingkat akurasi dengan *Confusion Matrix*.

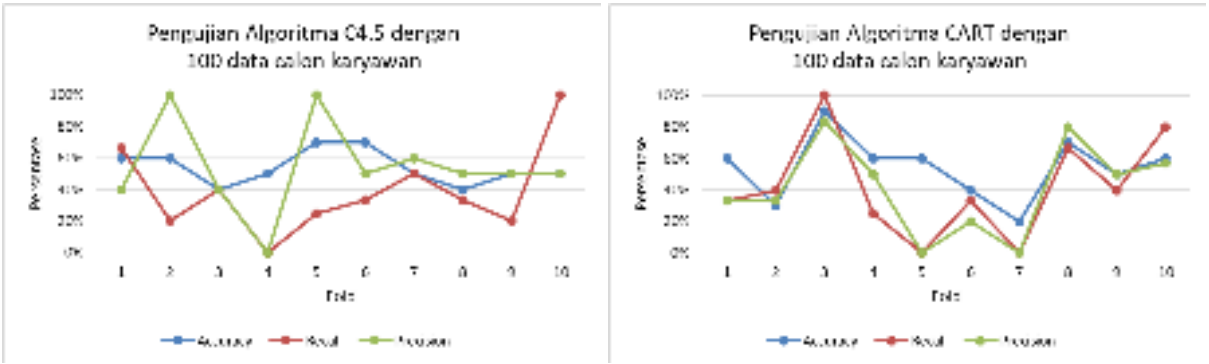
3. HASIL DAN PEMBAHASAN

Hasil seleksi calon karyawan dengan algoritma C4.5 dan CART akan diuji dengan menggunakan *K-Fold Cross Validation* dengan nilai k = 10 (*10-Fold Cross Validation*), dimana akan dilakukan pengujian sebanyak 3 kali, yakni terhadap 50, 100, dan 200 data calon karyawan. Pada pengujian terhadap 50, 100, dan 200 data calon karyawan akan terdiri dari 5, 10, dan 20 data calon karyawan untuk masing-masing *fold*.

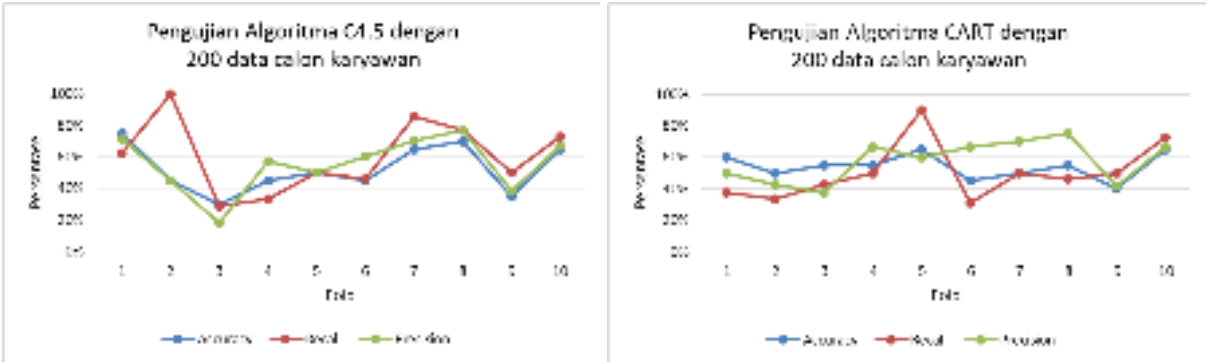
Gambar 3 menggambarkan hasil pengujian seleksi data calon karyawan dengan algoritma C4.5 dan CART terhadap 50 data calon karyawan sebanyak 10 *fold*, dengan masing-masing *fold* terdiri dari 5 data calon karyawan, Gambar 4 menggambarkan hasil pengujian seleksi data calon karyawan dengan algoritma C4.5 dan CART terhadap 100 data calon karyawan sebanyak 10 *fold*, dengan masing-masing *fold* terdiri dari 10 data calon karyawan, dan Gambar 5 menggambarkan hasil pengujian seleksi data calon karyawan dengan algoritma C4.5 dan CART terhadap 200 data calon karyawan sebanyak 10 *fold*, dengan masing-masing *fold* terdiri dari 20 data calon karyawan.



Gambar 3. Pengujian Algoritma C4.5 dan CART dengan 50 data calon karyawan



Gambar 4. Pengujian Algoritma C4.5 dan CART dengan 100 data calon karyawan



Gambar 5. Pengujian Algoritma C4.5 dan CART dengan 200 data calon karyawan

Tabel 3. Perbandingan Algoritma C4.5 dan CART

Algoritma	Jumlah Data Pengujian	*Accuracy*	*Recall*	*Precision*
C4.5	50 data	52 %	25 %	22,5 %
	100 data	54 %	38,83 %	54 %
	200 data	52,5 %	60,59 %	55,44 %
	Rata-rata	**52,83 %**	**41,48 %**	**43,98 %**
CART	50 data	52 %	40 %	30 %
	100 data	54 %	41,83 %	40,71 %
	200 data	54 %	50,33 %	57,70 %
	Rata-rata	**53,33 %**	**44,06 %**	**42,81 %**

Tabel 3 menunjukkan hasil perbandingan algoritma C4.5 dan CART berdasarkan semua pengujian yang telah dilakukan terhadap 50, 100, dan 200 data calon karyawan. Dari hasil perbandingan menunjukkan algoritma CART memiliki tingkat akurasi (*Accuracy*) dalam menyeleksi data calon karyawan sebesar 53,33%, dimana lebih tinggi dibandingkan algoritma C4.5 yakni sebesar 52,83%. Untuk tingkat keberhasilan sistem (*Recall*) dalam menyeleksi data calon karyawan, algoritma CART memiliki tingkat keberhasilan sistem sebesar 44,06%, dimana lebih tinggi dibandingkan algoritma C4.5 yakni sebesar 41,48%. Sedangkan untuk tingkat ketepatan keputusan (*Precision*) hasil seleksi calon karyawan, algoritma C4.5 memiliki tingkat ketepatan yang lebih tinggi yakni sebesar 43,98% dibandingkan algoritma CART yakni sebesar 42,81%.

Dari tabel 3 juga menunjukkan tingkat akurasi (*Accuracy*) tidak dipengaruhi oleh banyaknya data pengujian, akan tetapi tingkat keberhasilan sistem (*Recall*) dan tingkat ketepatan keputusan (*Precision*) pada algoritma C4.5 akan meningkat cukup signifikan apabila jumlah data pengujian semakin banyak, yakni tingkat *Recall* sebesar 25% (50 data), 38,83% (100 data), dan 60,59% (200 data), sedangkan tingkat *Precision* sebesar 22,5% (50 data), 54% (100 data), dan 55,44% (200 data). Sedangkan pada algoritma CART, tingkat keberhasilan sistem (*Recall*) dan tingkat ketepatan keputusan (*Precision*) hanya terjadi peningkatan minor apabila jumlah data pengujian semakin banyak, yakni tingkat *Recall* sebesar 40% (50 data), 41,83% (100 data), dan 50,33% (200 data), sedangkan tingkat *Precision* sebesar 30% (50 data), 40,71% (100 data), dan 57,7% (200 data).

4. KESIMPULAN

Algoritma CART memiliki tingkat akurasi dan tingkat keberhasilan yang lebih tinggi dalam menyeleksi calon karyawan jika dibandingkan dengan algoritma C4.5, sedangkan

algoritma C4.5 memiliki tingkat ketepatan keputusan hasil seleksi calon karyawan yang lebih tepat jika dibandingkan dengan algoritma CART. Besaran tingkat akurasi dari algoritma C4.5 dan CART tidak dipengaruhi oleh jumlah data calon karyawan yang akan diseleksi, sedangkan besaran tingkat keberhasilan sistem dan tingkat ketepatan keputusan hasil seleksi calon karyawan dengan algoritma C4.5 dan CART dipengaruhi oleh jumlah data calon karyawan yang diseleksi, dimana semakin banyak jumlah data calon karyawan yang akan diseleksi, maka semakin tinggi tingkat keberhasilan sistem dan tingkat ketepatan hasil keputusan seleksi.

5. SARAN

Disarankan untuk dilakukan penelitian lanjutan untuk membandingkan algoritma dari sisi kinerja waktu yang diperlukan dalam proses klasifikasi, atau membandingkan algoritma data mining yang memanfaatkan *gini index*.

DAFTAR PUSTAKA

[1] Dataaspirant, *Introduction to Data Mining Techniques*, http://dataaspirant.com/2014/09/16/data-mining/, diakses tgl 22 Juli 2018.

[2] Hermawati, F., A., 2013, *Data Mining*, Andi Offset, Yogyakarta.

[3] Cinaroglu, S., 2016, *Comparison of Performance of Decision Tree Algorithms and Random Forest: An Application on OECD Countries Health Expenditures*, https://www.ijcaonline.org/research/volume138/number1/cinaroglu-2016-ijca-908704.pdf, diakses tgl 06 Agustus 2018.

[4] Sari, R., D., I., dan Sindunata, Y., 2014, *Penerapan Data Mining untuk Analisa Pola Perilaku Nasabah dalam Pengkreditan Menggunakan Metode C.45 Studi Kasus pada KSU Insan Kamil Demak*, https://lp2m.asia.ac.id/wp-content/uploads/2015/05/JURNAL-RINA-DEWI.pdf, diakses 20 Juni 2018.

[5] Susanto, S., dan Suryadi, D., 2010, *Pengantar Data Mining*, Andi Publisher, Yogyakarta.

[6] Assiroj, P., 2016, *Kajian Perbandingan Teknik Klasifikasi Algoritma C4.5, Naive Bayes dan CART untuk Prediksi Kelulusan Mahasiswa (Studi Kasus : STMIK Rosma Karawang)*, http://jurnal.likmi.ac.id/Jurnal/7_2016/0716_01_PRIATI.pdf, diakses tgl 20 Mei 2018.

[7] Fahmi, A., Siswanto, A., Farid, M., F., dan Arijulmanan, 2014, *HRD Syariah Teori dan Implementasi*, Gramedia, Jakarta.

[8] Ferdinand, A., 2006, *Metode Penelitian Manajemen: Pedoman Penelitian untuk Penulisan Skripsi, Tesis, dan Disertasi Ilmu Manajemen*, Badan Penerbit Universitas Diponegoro, Semarang.

[9] Sharma, S., Agrawal, J., dan Sharma, S., 2013, *Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies*, https://pdfs.semanticscholar.org/063f/f47785db552ddc49b1df71b4c0497b5d3fe5.pdf, diakses tgl 20 Juni 2018.

[10] Lewis, R., J., 2000, *An Introduction to Classification and Reggression Trees (CART) Analysis*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf, diakses tgl 18 Juni 2018.

[11] Wibowo, A., 2017, 10 Fold-Cross Validation, https://mti.binus.ac.id/2017/11/24/10-fold-cross-validation/, diakses tgl 27 Mei 2018.

[12] Sharma, A., Confusion Matrix in Machine Learning, https://www.geeksforgeeks.org/confusion-matrix-machine-learning/, diakses tgl 30 Mei 2018.