JIMP - Jurnal Informatika Merdeka Pasuruan
Vol 5 No 1 Maret 2020
ISSN 2503-1945

Analisis Pemeringkatan Kualitas Klasifier Pada *Dataset* Tidak Seimbang

Choirul Anam [1], Ninanesia Rusdiana2 [2]
[1,2)]Prodi Sistem Informasi, AMIK Taruna Probolinggo
Jl. Raya Leces A3, Leces, Probolinggo 67273
Telp: (0335) 681497, Fax: (0335) 680954
E-mail: **ch.a6.rowi@gmail.com [1]**, rninanesia@gmail.com [2]

Abstract

Classification algorithms C4.5*,* CART*, k-Nearest Neighbors* (k-NN) *and Naive Bayes are included in the "Top 10 algorithms in data mining". The author tests and analyzes the four to get a ranking order according to the quality of performance. A common method to compare the quality of classifier performance for classifying two class labels with a balanced the proportion of the number of classes of datasets is to test the performance of classifier accuracy. For unbalanced datasets such as in this study using this method can be biased, and can even lead to misleading conclusions. By calculating the scores that are a combination of performance parameters "accuracy", "precision", "recall" and "AUC" where the highest value of each parameter is the best will result in a more representative rating of the classifier's performance indicating the quality of the classifier. Two test methods are used, namely 10-fold Cross Validation and Discrete Testing to ensure the results of a representative performance evaluation of each classifier. The implementation of the testing of the four classification algorithms above and the comparative analysis of the performance results in the ranking of the best quality performance ratings, namely:* 1. k-NN, 2. C4.5, 3. CART, 4. *Naive Bayes.*

Keywords: C4.5, CART, k-NN, *Naive Bayes, score*

Abstrak

Algoritma klasifikasi C4.5, CART, *k-Nearest Neighbours* (k-NN) dan *Naive Bayes* termasuk dalam "*Top 10 algorithms in data mining"*. Penulis melakukan pengujian dan analisis pada keempatnya untuk mendapatkan urutan peringkat menurut kualitas kinerjanya. Metode yang umum dan cukup memadai untuk membandingkan kualitas kinerja klasifier untuk klasifikasi dua label kelas dengan proporsi jumlah kelas dari *dataset* yang seimbang adalah dengan menguji kinerja *accuracy* klasifier. Untuk *dataset* yang tidak seimbang seperti dalam penelitian ini menggunakan metode ini bisa bias, bahkan bisa menghasilkan kesimpulan yang menyesatkan. Dengan menghitung skor nilai yang merupakan gabungan dari parameter kinerja "*accuracy*", "*precision*", "*recall*" dan "AUC" dimana nilai tertinggi dari masing-masing parameter adalah yang terbaik akan menghasilkan penilaian kinerja klasifier yang lebih representatif menunjukkan kualitas klasifier. Dilakukan dua metode pengujian yaitu *10-fold Cros Validation* dan Pengujian Secara Diskrit untuk memastikan hasil penilaian kinerja yang representatif dari masing-masing klasifier. Penerapan pengujian terhadap empat algoritma klasifikasi diatas dan analisis perbandingan kinerja menghasilkan urutan peringkat kualitas kinerja terbaik yaitu: 1. k-NN, 2. C4.5, 3. CART, 4. *Naive Bayes*.

Kata Kunci: C4.5, CART, k-NN, *Naive Bayes*, skor

I. PENDAHULUAN

Algoritma klasifikasi C4.5, CART, k-NN dan *Naive Bayes* termasuk dalam "*Top 10 algorithms in data mining"* [1]. Penting untuk mengetahui algoritma mana yang memiliki kualitas terbaik dari keempatnya. Metode yang paling umum untuk membandingkan kualitas klasifier adalah dengan menguji kinerja *accuracy* masing-masing klasifier. Ini pilihan yang cukup memadai untuk klasifikasi dua kelas label dengan proporsi jumlah kelas dari data sampel (*dataset*) yang seimbang [2]. Namun dalam realitas jarang terjadi *dataset* dengan proporsi jumlah kelas yang seimbang. Kondisi ini disebut sebagai *imbalanced problem* [3, 4]. Bukti empiris menunjukkan bahwa ukuran kinerja *accuracy* adalah bias terhadap *dataset*

tidak seimbang [2]. Dalam kondisi ini penggunaan ukuran kinerja *accuracy* saja tidaklah cukup, bahkan bisa menghasilkan kesimpulan yang menyesatkan [2]. Pengukuran kinerja klasifier dengan menambahkan parameter ukuran kinerja "*precision*", "*recall*" dan "AUC" dimana nilai tertinggi dari masing-masing parameter adalah yang terbaik akan menghasilkan penilaian kinerja klasifier yang lebih representatif.

Penelitian-penelitian dalam 5 tahun terakhir terkait perbandingan kinerja klasifier tersebut diantaranya adalah: [5] membandingkan akurasi algoritma C4.5, *Neural Network* dan *Naïve Bayes* dalam penerapan otentikasi uang kertas (*banknote authentication*) mendapatkan hasil tingkat akurasi klasifikasi algoritma C4.5 98.5%, *Neural Network* 95% dan *Naïve Bayes* 85%. Penelitian bisa dikembangkan untuk penelitian dengan metode *classification* yang lain untuk mendapatkan akurasi klasifikasi terbaik. [6] membandingkan algoritma C4.5 dan *Naive Bayes* dalam pengujian untuk mendapatkan tingkat akurasi terbaik untuk diterapkan dalam kasus pemilihan konsentrasi keahlian mahasiswa. Tingkat akurasi algoritma C4.5 sebesar 84,43% dan setelah dioptimasi meningkat menjadi 84,98%, tingkat akurasi algoritma *Naive Bayes* sebesar 78,47% dan setelah dioptimasi meningkat menjadi 82,01%. [7] melakukan penelitian komparasi penerapan algoritma C4.5, k-NN dan *Neural Network* dalam proses kelayakan penerimaan kredit kendaraan bermotor dengan pengujian kinerja akurasl/AUC menunjukkan hasil dari algoritma C4.5 sebesar 92.89%/0.958 lebih tinggi dari algoritma k-NN sebesar 77.78%/0.905. [8] dalam penelitian yang mengkomparasi metode-metode *Naive Bayes*, C4.5, dan *Neural Network* yang digunakan untuk prediksi klinis dalam sistem penunjang keputusan dalam bidang kesehatan untuk menghindari resiko-resiko yang terjadi pada proses persalinan, hasil penelitian menunjukkan akurasi prediksi proses persalinan dengan metode *Naïve Bayes* = 94%, metode C4.5 = 90% dan metode *Neural Network* = 93%. [9] dalam penelitian Perbandingan Kinerja Algoritma CART dan *Naive Bayes* Untuk Mendiagnosa Penyakit Diabetes Melitus dengan parameter kinerja *accuracy*, *precision*, *recall* dan *f-measure* menunjukkan hasil algoritma CART memiliki nilai 76.9337%, 0.764, 0.769 dan 0.765 yang lebih baik dari algoritma *Naive Bayes* yang memiliki nilai 73.7569%, 0.732, 0.738 dan 0.734. [10] membandingkan akurasi algoritma C4.5 dan CART dalam memprediksi kategori indeks prestasi mahasiswa dengan data latih tidak seimbang dan berbagai sekanario pengujian menghasilkan rata-rata akurasi dari algoritma CART sebesar 50.32%, lebih tinggi dari rata-rata akurasi algoritma C4.5 sebesar 47.23%. [11] dalam Perbandingan Algoritma C4.5, k-NN dan *Naive Bayes* untuk Penentuan Model Klasifikasi Penanggung Jawab BSI Entrepreneur Center menyimpulkan metode *Naive Bayes* terbaik dengan nilai akurasi sebesar 80% dibanding metode C4.5 dengan nilai akurasi 73.33% dan metode k-NN dengan nilai akurasi 70%. [12] dalam Analisis Performa Algoritma k-NN dan C4.5 pada Klasifikasi Data Penduduk Miskin di Kecamatan Bantul Yogyakarta dengan parameter kinerja *accuracy*, *precisiioan* dan *recall* menyimpulkan algoritma k-NN memiliki performa yang lebih baik dari algoritma C4.5. [13] dalam penelitian Perbandingan Algoritma k-NN dan CART Pada Data Mining Penerimaan Beasiswa dengan menguji *accuracy*, *precision*, *recall* dan *f-measure* menunjukkan algoritma k-NN mendapat nilai 99.2958%, 0.993, 0.993 dan 0.993 lebih baik dari algoritma CART yang mendapat nilai 71.1268%, 0.506, 0.711 dan 0.591.

Tujuan dari penelitian ini adalah melakukan pengujian dan analisis komprehensip terhadap algoritma klasifikasi C4.5, CART, k-NN dan *Naive Bayes* dengan metode pengujian gabungan dan skoring gabungan parameter kinerja untuk mendapatkan nilai kinerja yang representatif dari masing-masing klasifier untuk kondisi *dataset* tidak seimbang, untuk selanjutnya dilakukan pemeringkatan klasifier berdasarkan skor nilai kinerja.

II. MASALAH

Penelitian-penelitian yang telah begitu banyak dilakukan untuk membandingkan kinerja klasifier banyak menyandarkan pada nilai ukuran kinerja *accuracy* saja. Dari hasil penelitian-penelitian yang lalu menunjukkan bahwa perbandingan tingkat *accuracy* antara dua klasifier bisa berkebalikan untuk obyek klasifikasi yang berbeda. Permasalahan yang menjadi dasar dilakukannya penelitian ini adalah:

1. Untuk *dataset* tidak seimbang bagaimana cara pengujian secara komprehensip tentang kualitas dari klasifier-klasifier populer di atas ?
2. Bagaimana cara pengukuran kinerja klasifier secara komprehensip sehingga menunjukkan nilai yang reperesentatif menunjukkan kualitas klasifier ?

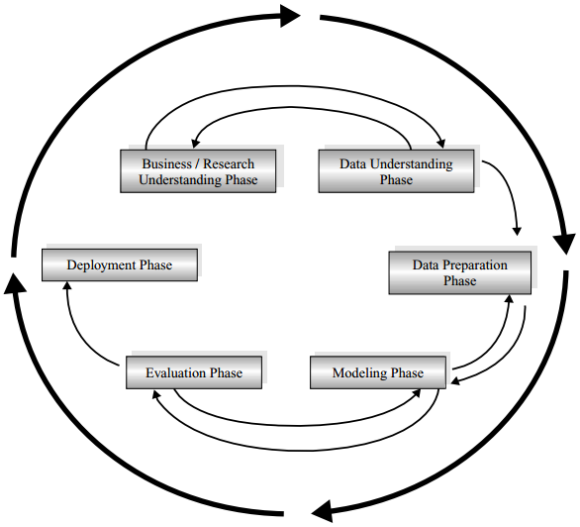
III. METODE PELAKSANAAN

- Metode Pengumpulan Data

Data diambil dari data sekunder institusi yang berupa data proses penentuan mahasiswa penerima beasiswa PPA, yang akan digunakan sebagai *dataset* eksperimen penelitian. Data yang didapatkan masih dalam bentuk *raw data* sehingga memerlukan proses *preparation*.

- 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 1 menjelaskan tentang siklus hidup pengembangan *data mining* yang telah ditetapkan dalam CRISP-DM.



Gambar 1. Tahapan proses CRISP dalam *data mining* [14]

Penjelasan dari proses CRISP:

1. Business/Research Understanding Phase

Memahami sistem dan permasalahan yang akan digunakan dalam pengujian dan evaluasi kinerja klasifier.

2. Data Understanding Phase

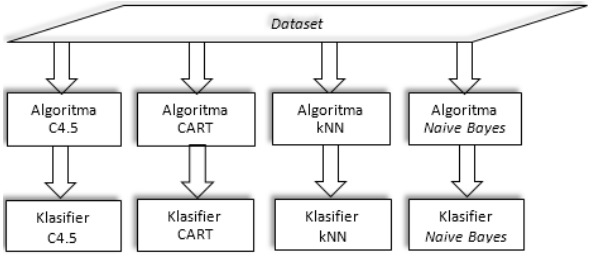
Data yang digunakan adalah data sekunder yang berupa rekaman proses penentuan mahasiswa penerima beasiswa PPA dengan distribusi kelas yang tidak seimbang, dimana jumlah mahasiswa yang lolos (penerima beasiswa) jauh lebih kecil dari jumlah mahasiswa yang tidak lolos.

3. Data Preparation Phase

Melakukan pengolahan pada *raw data* dengan menghilangkan atribut-atribut data yang tidak berpengaruh terhadap variabel keputusan, menangani data kosong dan melakukan transformasi data sehingga menghasilkan *dataset* yang lebih ringkas dan informatif terhadap proses klasifikasi.

4. Modeling Phase

Membangun model klasifikasi (klasifier) dengan tool "*Rapidminer software for data mining"* untuk proses klasifikasi menggunakan algoritma C4.5, CART, k-NN dan *Naive Bayes* seperti diperlihatkan pada Gambar 2.



Gambar 2. Pembangunan model klasifikasi

5. Evaluation Phase

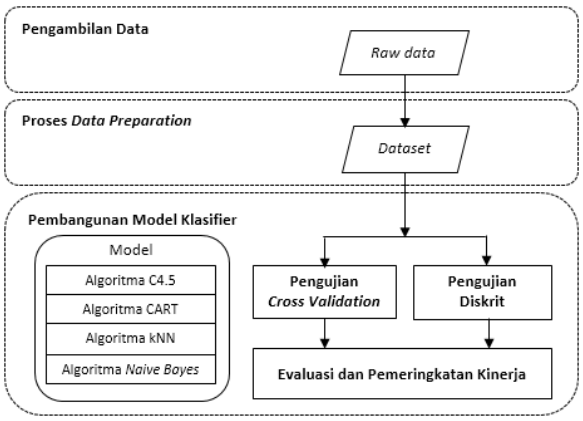
Melakukan pengujian dan evaluasi terhadap masing-masing klsifier, mendapatkan nilai kinerja masing-masing model, mengkomparasi dan menyusun peringkat kinerja klasifier.

6. Deployment Phase

Penelitian berhenti di *evaluation phase* setelah diperoleh urutan peringkat kinerja klasifier.

- Alur Penelitian

Tahapan penelitian yaitu dimulai pengambilan data (*raw data*), lalu proses *data preparation* menjadi *dataset*, kemudian pembangunan klasifier, terus dilakukan pengujian klasifier dan pemeringkatan kinerja klasifier. Tahapan penelitian seperti ditunjukkan pada Gambar 3.



Gambar 3. Alur Penelitian

IV. HASIL DAN PEMBAHASAN

1. Pembentukan *dataset*

Raw data dalam format *excel* sejumlah 383 data tersusun dari 25 kolom (*field*), terdapat 10 data yang dikeluarkan dari analisis karena isinya tidak lengkap. 7 kolom dalam *raw data* yang nilainya berpengaruh terhadap keputusan sebagi atribut, yaitu semester, beasiswa lain, IPK, prestasi ko/ekstra kurikuler, penghasilan orang tua, biaya listrik dan jumlah tanggungan.

Untuk tujuan proses klasifikasi berjalan efisien maka pada nilai atribut-atribut dilakukan konversi [15] ke nilai kategoris sehingga atribut dan nilai pada *dataset* seperti pada Tabel 1.

Tabel 1. Atribut *data set*

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

Terbentuk *dataset* yang terdiri dari 373 data yang digunakan untuk eksperimen dalam penelitian ini. Gambar 4 menunjukkan potongan *dataset* yang dimaksud.

NO	Semester	Beasiswa lain	IPK	Prestasi ko/ekstra	Penghasilan orang tua	Biaya listrik	Jumlah tanggungan	Keputusan
1	Atas	Tidak	Cukup	Tidak	Sedang	Sedang	Sedang	**Tidak**
2	Atas	Tidak	Cukup	Tidak	Sedang	Sedang	Sedang	**Tidak**
3	Atas	Tidak	Tinggi	Tidak	Rendah	Rendah	Rendah	**Ya**
4	Atas	Tidak	Tinggi	Tidak	Sedang	Rendah	Sedang	**Ya**
5	Atas	Tidak	Cukup	Tidak	Sedang	Rendah	Rendah	**Tidak**
369	Bawah	Tidak	Cukup	Tidak	Rendah	Rendah	Sedang	**Ya**
370	Bawah	Tidak	Cukup	Tidak	Rendah	Rendah	Sedang	**Ya**
371	Bawah	Tidak	Tinggi	Tidak	Rendah	Sedang	Sedang	**Ya**
372	Bawah	Tidak	Tinggi	Tidak	Sedang	Rendah	Sedang	**Ya**
373	Bawah	Tidak	Cukup	Tidak	Sedang	Sedang	Sedang	**Ya**

Gambar 4. Potongan *dataset*

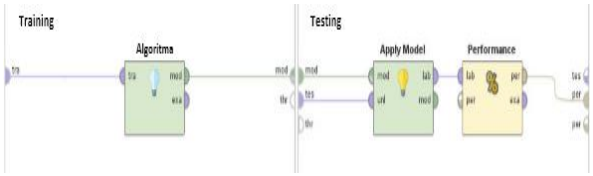
Dari statistik *dataset* didapat distribusi kelas Keputusan:"Ya" adalah 121 (32.44%) dan kelas Keputusan:"Tidak" adalah 252 (67.56%) maka termasuk distribusi kelas yang tidak seimbang.

2. Pembangunan Model dan Pengujian

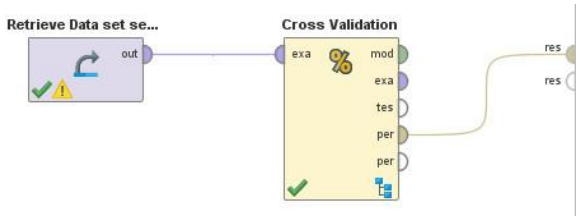
Pembangunan model klasifikasi untuk dilakukan pengujian metode *10-fold cross validation* dan pengujian secara diskrit.

a. Pengujian dengan *10-fold Cross Validation*

Dengan menggunakan tool *Rapidminer Studio* pembangunan model klasifikasi untuk ketiga algoritma pada pengujian ini seperti pada Gambar 5 dan proses validasi dari ketiga model menggunakan konfigurasi yang sama seperti Gambar 6.



Gambar 5. Konfigurasi model klasifikasi



Gambar 6. Konfigurasi *10-fold Cross Validation*

Hasil pengujian merupakan nilai-nilai yang terangkum dalam *PerformanceVector* dan informasi *execution time* dari masing-masing klasifier ditunjukkan pada Tabel 2.

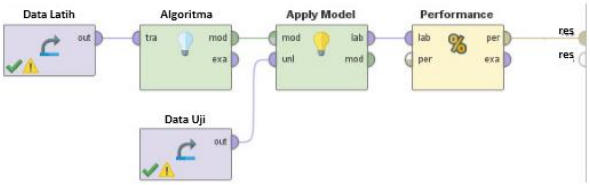
Tabel 2. Data kinerja hasil pengujian klasifier

Parameter Kinerja	Model C4.5	Model CART	Model kNN	Model *Naive Bayes*
Accuracy	0.9733	0.9651	0.9813	0.9410
Precision	0.9295	0.9546	0.9702	0.9550
Recall	0.9750	0.9679	0.9750	0.8608
AUC	0.980	0.991	0.980	0.985
execution time	0 s	0 s	0 s	0 s

Pada *dataset* dengan distribusi kelas yang tidak seimbang, pengukuran kinerja model klasifikasi dengan menggunakan parameter *accuracy* saja tidaklah cukup [2], pengukuran kinerja model juga melibatkan parameter "*precision*", "*recall*" dan "AUC" dimana nilai tertinggi adalah yang terbaik. Parameter kinerja "*execution time*" ketiga model menunjukkan hasil 0 s (nol detik) yang mengindikasikan jumlah *dataset* yang ada tidak membutuhkan waktu proses yang dapat diperbandingkan.

b. Pengujian secara Diskrit

Hasil pengujian sebelumnya belum memastikan salah satu klasifier secara *real* dan konsisten mempunyai kinerja lebih baik dari yang lain. Untuk melengkapi data hasil uji dilakukan pengujian secara diskrit pada ketiga model klasifikasi dimana data-uji bukan merupakan bagian dari data-latih. Pembangunan model klasifikasi untuk ketiga algoritma pada pengujian ini seperti pada Gambar 7.



Gambar 7. Konfigurasi pengujian secara diskrit

Dilakukan 10 kali proses uji menggunakan data-latih dan data-uji dari *dataset* yang diatur dimana setiap proses uji menggunakan data-uji berupa bagian data 1 tahun dari *dataset* dan bagian data 9 tahun sisanya dari *dataset* digunakan sebagai data-latih seperti pada Tabel 3.

Tabel 3. Pengaturan data-latih dan data-uji

Uji	Data-Uji			Data-Latih		
	Tahun (Jumlah data)	Keputusan		Jumlah data	Keputusan	
		Ya	Tidak		Ya	Tidak
1	2008 (31)	11	21	342	110	232
2	2009 (43)	14	29	330	107	223
3	2010 (40)	15	25	329	106	223
4	2011 (53)	16	37	320	105	215
5	2012 (42)	14	28	331	107	224
6	2013 (31)	15	16	342	106	236
7	2014 (37)	7	30	336	114	222
8	2015 (33)	7	26	340	114	226
9	2016 (30)	7	23	343	114	229
10	2017 (33)	15	18	340	106	234

Rata-rata nilai parameter kinerja pada *PerformanceVector* hasil uji masing-masing klasifier dirangkum dalam sebuah tabel seperti diperlihatkan pada Tabel 4.

Tabel 4. Rata-rata nilai kinerja klasifier yang diuji

Parameter Kinerja	Model C4.5	Model CART	Model k-NN	Model *Naive Bayes*
Accuracy	0.9398	0.9385	0.9748	0.9344
Precision	0.8945	0.9053	0.9586	0.9303
Recall	0.9457	0.9248	0.9590	0.8717
AUC	0.9626	0.9726	0.9707	0.9863

c. Evaluasi Hasil Pengujian

Penilaian skor kinerja klasifier yaitu dilakukan dengan menjumlahkan nilai semua parameter kinerja dengan mempertimbangkan bobot yang sama untuk masing-masing parameter kinerja. Penilaianan skor kinerja hasil pengujian adalah sebagai berikut:

- Hasil penilaian skor kinerja klasifier dari hasil pengujian *10-fold Cross Validation* ditunjukkan pada Tabel 5.

Tabel 5. Perhitungan skor kinerja klasifier

Parameter Kinerja		Model C4.5	Model CART	Model k-NN	Model *Naive Bayes*
Nama	Nilai Kinerja				
Accuracy	1.00	0.9733	0.9651	0.9813	0.9410
Precision	1.00	0.9295	0.9348	0.9702	0.9550
Recall	1.00	0.9750	0.9667	0.9750	0.8608
AUC	1.00	0.980	0.991	0.980	0.985
Skor		**3.8578**	**3.8576**	**3.9065**	**3.7418**

Dari Tabel 5 menunjukkan urutan peringkat dari nilai skor kinerja tertinggi adalah: model k-NN, model C4.5, model CART, model *Naive Bayes*.

- Hasil penilaian skor kinerja klasifier dari hasil pengujian secara Diskrit ditunjukkan pada Tabel 6.

Tabel 6. Penilaian skor kinerja klasifier hasil uji secara Diskrit

Parameter Kinerja		Model C4.5	Model CART	Model k-NN	Model *Naive Bayes*
Nama	Nilai Terbaik				
Accuracy	1.00	0.9398	0.9385	0.9748	0.9344
Precision	1.00	0.8945	0.9053	0.9586	0.9303
Recall	1.00	0.9457	0.9248	0.9590	0.8717
AUC	1.00	0.9626	0.9726	0.9707	0.9863
Skor		**3.7426**	**3.7411**	**3.8631**	**3.7227**

Dari Tabel 6 menunjukkan urutan peringkat dari nilai skor kinerja tertinggi adalah: model k-NN, model C4.5, model CART, model *Naive Bayes*.

Dari dua metode pengujian menunjukkan hasil urutan yang sama dalam hal peringkat skor kinerja, yaitu: peringkat 1 (tertinggi) model k-NN, peringkat 2 model C4.5, peringkat 3 model CART dan peringkat 4 (terendah) model *Naive Bayes*.

V. KESIMPULAN

1. Dua metode pengujian kinerja klasifier dan penilaian skor kinerja gabungan dari semua pameter kinerja dilakukan untuk memastikan hasil penilaian kinerja yang representatif untuk menunjukkan kualitas dari klasifier.
2. Hasil dua metode pengujian kinerja klasifier menunjukkan urutan peringkat kinerja yang sama, yaitu (urutan dari peringkat tertinggi): 1. k-NN, 2. C4.5, 3. CART, 4. *Naive Bayes*.
3. Perlunya dilakukan pengujian lebih lanjut dengan jumlah data yang jauh lebih besar sehingga faktor waktu proses melengkapi bobot penilaian kinerja klasifier.

DAFTAR PUSTAKA

[1] Kumar, V.; Wu, X..; Quinlan, J.R.: Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; Zhou, Z.H.; Steinbach, M.; Hand, D.J.; Steinberg, D. .*Top 10 algorithms in data mining*. Knowl Inf Syst (2008) 14:1–37 DOI 10.1007/s10115-007-0114-2. Springer-Verlag London Limited, 2007

[2] Stapor, K. .*Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations*. Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017, Advances in Intelligent Systems and Computing 578, DOI 10.1007/978-3-319-59162-9 2, 2018

[3] Japkowicz, N., Stephen, N.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 40–49, 2002

[4] Sun, Y., et al.: Classification of imbalanced data: a review. Int. J. Pattern Recogn. Artif. Intell. 23(4), 687–719, 2009

[5] Sani, K.; Winarno, W.W.; Fauziati, S. Analisis Perbandingan Algoritma *Classification* Untuk *Authentication* Uang Kertas (Sudi kasus: *Banknote Athentication*). Jurnal Informatika Vol. 10, No. 1, 2015

[6] Supritanti, W.; Kusrini; Amborowati, A. Perbandingan Kinerja Algoritma C4.5 Dan *Naive Bayes* Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa. Jurnal INFORMA Politeknik Indonusa Surakarta ISSN : 2442-7942 Vol. 1 Nomor 3, 2016.

[7] Astuti, Puji. Komparasi Penerapan Algoritma C4.5, k-NN dan Neural Network Dalam Proses Kelayakan Penerimaan Kredit Kendaraan Bermotor. Faktor Exacta 9(1): 87-101, ISSN: 1979-276X, 2016

[8] Amalia, H.; Evicienna. Komparasi Metode *Data Mining* untuk Penentuan Proses Persalinan Ibu Melahirkan. Jurnal Sistem Informasi (Journal of Information Systems). 2/13, 103-109 DOI: http://dx.doi.org/10.21609/jsi.v13i2.545, 2017

[9] Subarkah, Pangkas; Santiko, Irfan; Astuti, Tri. Perbandingan Kinerja Algoritma CART dan *Naive Bayes* Untuk Mendiagnosa Penyakit Diabetes Melitus. CITISEE, ISBN: 978-602-60280-1-3, 2017

[10] Alverina, Dea; Chrismanto, Antonius Rachmat; Santoso, R. Gunawan. Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa. Jurnal Teknologi dan Sistem Komputer, 6(2), 76-83, tersedia di https://jtsiskom.undip.ac.id, 2018

[11] Nurhasan, Fuad; Hikmah, Noer; Utami, Dwi Yuni. Perbandingan Algoritma C4.5, k-NN dan *Naive Bayes* untuk Penentuan Model Klasifikasi Penanggung Jawab BSI Entrepreneur Center. Jurnal PILAR Nusa Mandiri Vol. 14, No. 2. 2018

[12] Astuti, Femi Dwi; Guntara, Mohammad. Analisis Performa Algoritma k-NN dan C4.5

pada Klasifikasi Data Penduduk Miskin di Kecamatan Bantul Yogyakarta. JURTI, Vol.2 No.2, ISSN: 2579-8790, 2018.

[13] Rosyidi, Rahman. Perbandingan Algoritma k-NN dan CART Pada Data Mining Penerimaan Beasiswa. CESS (Journal of Computer Engineering System and Science), Vol. 4 No. 2, p-ISSN :2502-7131, e-ISSN :2502-714x, 2019

[14] Larose, D.T. Discovering Knowledge in Data. An Introduction to Data Mining. John Wiley & Sons, Inc., 2005

[15] Shmueli, G.; Patel, N.R.; Bruce, P.C. Data Mining for Business Intelligence. A John Wiley & Sons, Inc., Publication. 2010