Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 12(2) 2019: 1-12
http://ejournal.undip.ac.id/index.php/media_statistika

ANALISIS PERBANDINGAN KINERJA CART KONVENSIONAL, *BAGGING* DAN *RANDOM FOREST* PADA KLASIFIKASI OBJEK: HASIL DARI DUA SIMULASI

Yogo Aryo Jatmiko[1], Septiadi Padmadisastra[2], Anna Chadidjah[2]
[1] Badan Pusat Statistik
[2] Departemen Statistika, Universitas Padjajaran

e-mail: yj29289@gmail.com

DOI: 10.14710/medstat.12.1.1-12

Article Info:
Received: 19 March 2018
Accepted: 21 July 2019
Available Online: 24 July 2019

Keywords:
CART, Bagging, Random Forest, Classification

Abstract: The conventional CART method is a nonparametric classification method built on categorical response data. Bagging is one of the popular ensemble methods whereas, Random Forests (RF) is one of the relatively new ensemble methods in the decision tree that is the development of the Bagging method. Unlike Bagging, Random Forest was developed with the idea of adding layers to the random resampling process in bagging. Therefore, not only randomly sampled sample data to form a classification tree, but also independent variables are randomly selected and newly selected as the best divider when determining the sorting of trees, which is expected to produce more accurate predictions. Based on the above, the authors are interested to study the three methods by comparing the accuracy of classification on binary and non-binary simulation data to understand the effect of the number of sample sizes, the correlation between independent variables, the presence or absence of certain distribution patterns to the accuracy generated classification method. Results of the research on simulation data show that the Random Forest ensemble method can improve the accuracy of classification.

1. PENDAHULUAN

Dalam literatur statistika terdapat beberapa teknik pengklasifikasian pada peubah respon biner antara lain regresi logistik (Agresti, 2007) atau fungsi diskriminan (Johnson and Wichern, 2002) yang termasuk dalam metode parametrik. Metode parametrik memiliki kelebihan dalam hal interpretasi, membaca rata-rata ataupun penyimpangan suatu data akan lebih mudah dipahami dibandingkan membaca ranking dari data itu sendiri. Namun, metode-metode tersebut memiliki keterbatasan dalam menyelesaikan masalah pengklasifikasian, Di antaranya: (1) jumlah variabel independen yang banyak menyebabkan kesulitan dalam seleksi variabel terpenting, (2) seringkali distribusi variabel

independen tidak berdistribusi normal, khususnya dalam analisis diskriminan, (3) sesama variabel independen sering terdapat interaksi, dan (4) penerapan model yang dihasilkan cenderung sulit (Lewis, 2000).

Penentuan seseorang termasuk pekerja sektor formal ataupun informal merupakan salah satu contoh klasifikasi. Naibaho dalam Fajariyanto (2017) mengatakan bahwa klasifikasi digunakan ketika suatu objek yang telah ditetapkan berdasarkan atribut objek tersebut perlu diklasifikasikan ke dalam kelas/kelompok/grup. Pekerja sektor formal ataupun informal merupakan data nominal atau data dikotomus dengan skala 1 dan 2 untuk membedakannya. Penggunaan parameter-parameter statistik seperti rata-rata dan standar deviasi menjadi tidak relevan karena rata-rata data hanya menyebar di antara angka 1 dan 2. Keterbatasan tersebut di atas dapat diatasi dengan metode yang tidak terlalu terikat dengan beberapa asumsi, yaitu dengan pendekatan nonparametrik. Beberapa metode nonparametrik yang sering digunakan dalam pengklasifikasian objek diantaranya: *Multivariate Adaptive Regression Spline* (MARS), *Classification and Regression Tree* (CART) dan *Neural Network* (NN).

Metode CART konvensional merupakan metode klasifikasi nonparametrik yang dibangun pada data respon kategorik yang memiliki beberapa kelebihan di antaranya kemampuan bekerja pada struktur data kompleks dan dimensi data yang besar, tidak terikat oleh asumsi kenormalan maupun kehomogenan varians, dapat mengetahui interaksi antar variabel independen dan hasil klasifikasi yang lebih mudah dipahami (R. J. Lewis, 2000). CART juga merupakan metode yang dapat diterapkan untuk data dalam jumlah yang besar, variabel yang sangat banyak dan melalui prosedur pemilah biner (Pratiwi, 2014). Namun, metode CART masih memiliki beberapa kelemahan yaitu menghasilkan pohon yang kurang stabil dimana ketika data *training* mengalami perubahan kecil dapat memberikan perubahan yang signifikan pada pohon yang dihasilkan (Sutton, 2005). Salah satu metode yang dapat digunakan untuk mengatasi kelemahan dari metode CART adalah metode *Ensemble*. Metode *ensemble* merupakan sebuah metode dengan gagasan melakukan berbagai macam gabungan kombinasi dari banyak pemilah tunggal (*classifier*) menjadi sebuah prediksi akhir berdasarkan proses voting mayoritas.

Bagging (*Bootsrap Aggregating)* dan *Random Forest* merupakan beberapa metode *ensemble*. *Bagging* merupakan salah satu metode *ensemble* yang populer sedangkan, *Random Forests* (RF) adalah salah satu metode *ensemble* yang relatif baru dalam pohon keputusan yang merupakan pengembangan dari metode *Bagging*. Berbeda dengan *Bagging*, *Random Forest* dikembangkan dengan gagasan perlu adanya penambahan *layer* pada proses *resampling* acak pada *bagging*. Oleh karena itu, bukan hanya data sampel yang diambil secara acak untuk membentuk pohon klasifikasi, tetapi juga variabel independen diambil sebagian secara acak dan baru dipilih sebagai pemilah terbaik saat penentuan pemilah pohon, sehingga diharapkan menghasilkan prediksi yang lebih akurat.

Terdapat beberapa penelitian yang menggunakan metode CART, *Bagging* dan *Random Forest* sebagai perbandingan, antara lain penelitian mengenai analisis gesekan yang dilakukan oleh Hayes *et al.* pada tahun 2015 yang menggunakan beberapa metode antara lain *t* test, regresi logistik, CART, *Bagging* dan *Random Forest* menunjukkan bahwa *Pruned* CART dan *Random Forest* paling baik dalam memprediksi model akhir. Sedangkan Shaikhina *et al.* (2017) membandingkan model pohon keputusan CART dan *Random Forest* pada ketidakcocokan *antibody* saat transplantasi ginjal. Hasil penelitian Shaikhina *et al.* menunjukkan bahwa baik CART maupun *Random Forest* memberikan tingkat akurasi yang sama pada sampel yang kecil, namun *Random Forest* memberikan

hasil yang lebih *Robust*. Penelitian lain dilakukan oleh Muttaqin pada tahun 2013 tentang penggunaan metode *ensemble* pada CART untuk perbaikan hasil klasifikasi kemiskinan di kabupaten Jombang menunjukkan bahwa metode *Random Forest* (RF) menghasilkan akurasi klasifikasi yang lebih baik dibandingkan CART dan *Bagging*; serta penelitian yang dilakukan oleh Otok pada tahun 2015 mengenai klasifikasi rumah tangga sangat miskin di kabupaten Jombang menurut paket bantuan yang diharapkan menggunakan metode RF-CART. Hasil penelitian menunjukkan bahwa metode RF-CART mampu memberikan tingkat efisiensi dan efektivitas yang lebih tinggi dalam pengklasifikasian dibandingkan CART.

Berdasarkan hal tersebut diatas, penulis ingin menunjukkan bahwa data non biner dengan metode *Random Forest* memiliki kinerja yang lebih baik dibandingkan dengan data biner. Peneliti juga ingin menunjukkan bahwa metode *Random Forest* data non biner memiliki kinerja lebih baik dibandingkan metode CART dan *Bagging* dengan membandingkan ketepatan klasifikasi ketiga metode tersebut pada data simulasi untuk memahami pengaruh jumlah ukuran sampel, korelasi antar variabel independen, ada tidaknya pola distribusi tertentu terhadap akurasi yang dihasilkan metode klasifikasi. Dalam penelitian ini metode CART yang dimaksud merupakan metode CART konvensional.

2. TINJAUAN PUSTAKA

2.1. *Classification And Regression Trees* (CART)

Classification And Regression Trees (CART) adalah suatu metode nonparametrik dari salah satu teknik eksplorasi data yang dikenal sebagai teknik pohon keputusan (*decision tree*). Pohon keputusan dibentuk secara biner dengan algoritma penyekatan rekursif (*binary recursive partitioning*) (Lewis, 2000). Metode CART dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, dan Charles J. Stone pada tahun 1980-an. Menurut Roger dan Lewis (2000), CART merupakan suatu cara pemilahan sekelompok data dalam suatu ruang yang disebut simpul (node) menjadi dua simpul anak dan setiap simpul anak dapat dipilah lagi menjadi dua simpul anak berikutnya, begitu seterusnya dan berhenti jika telah mendapatkan sekelompok observasi yang relatif homogen. CART menggambarkan hubungan antara variabel dependen dengan satu atau lebih variabel independennya. CART bertujuan untuk mendapatkan pengklasifikasian suatu kelompok data yang akurat (Breiman *et al*, 1984).

2.2. *Bagging*

Bagging adalah singkatan dari *bootstrap aggregating*, yaitu salah satu metode *ensemble* yang diusulkan oleh Leo Breiman yang digunakan untuk mereduksi variansi variabel independen sehingga diharapkan dapat memperbaiki kualitas prediksi suatu klasifikasi dari pohon klasifikasi tunggal. Ide dasar metode *Bagging* adalah menggunakan *resampling* acak dengan pengembalian pada *dataset* awal sehingga diperoleh suatu *dataset* baru. *Dataset* baru D_k berukuran sama dengan data *training* $|D|$ dengan mengambil sampel acak berukuran *N* dengan pengembalian dari data *training* (sampel bootstrap S_k dari D_k). *Dataset* baru kemudian digunakan untuk membangkitkan pohon klasifikasi dengan banyak versi. Pohon klasifikasi dari setiap versi kemudian digabungkan untuk memperoleh prediksi akhir.

2.3. *Random Forest*

Random Forests (RF) merupakan salah satu metode *ensemble* untuk meningkatkan akurasi suatu klasifikasi data dari sebuah pemilah tunggal yang tidak stabil melalui kombinasi banyak pemilah dari suatu metode yang sama dengan proses *voting* untuk memperoleh prediksi klasifikasi akhir (Wezel dan Potharst, 2007). Istilah *Random Forests* diusulkan pertama kali oleh Tin Kam Ho dari Bell Labs pada tahun 1995. Pengembangan *Random Forests* dilakukan oleh Leo Breiman pada tahun 2001 dari proses *bootsrap aggregating* atau yang lebih popular dengan sebutan *bagging*.

Pada proses *bagging* digunakan *resampling bootstrap* untuk membangkitkan pohon klasifikasi yaitu suatu teknik bangkitan dengan banyak versi yang kemudian mengkombinasikannya untuk memperoleh prediksi akhir. Sedangkan dalam metode *Random Forests*, proses pengacakan tidak hanya dilakukan pada data sampel saja melainkan juga pada pengambilan variabel independen sehingga pohon klasifikasi yang dibangkitkan akan memiliki ukuran dan bentuk yang berbeda-beda (Liaw dan Weiner, 2002).

Baik metode *bagging* maupun *Random Forest* tidak menjamin 100 persen memberikan akurasi yang lebih baik dibandingkan metode CART. Penggunaan *resampling bootsrap* pada *bagging* dan *Random Forest* memiliki kelemahan, yaitu keduanya berisiko kehilangan informasi, namun kelebihan dari kedua metode tersebut dibandingkan CART adalah diharapkan lebih merepresentasikan populasi.

Algoritma *Random Forest* yang digunakan adalah sebagai berikut:

Langkah 1

Inisialisasi parameter data *training D*

- $\mathcal{D} = \emptyset$, *ensemble*.
- *K*, jumlah pemilah untuk data *training*.

Langkah 2

Untuk setiap $k = 1, \dots, K,$ dibentuk sebuah dataset baru D_k dengan mengambil sampel acak berukuran *N* dengan pengembalian dari data *training* (sampel *bootstrap* S_k dari D_k) (beberapa data dapat terpilih berulangkali dan beberapa data mungkin tidak terpilih sama sekali).

Langkah 3

Untuk setiap $k = 1, \dots, K,$ ambil sampel acak tanpa pengembalian pada variabel independen, yaitu dengan memilih *m* variabel dari *p* variabel. Jumlah variabel yang dapat diambil secara acak (*m*) ditentukan melalui perhitungan: Hastie *et al.*(2009)

i. untuk Klasifikasi, nilai *default* untuk *m* adalah $\sqrt{p}$ dan ukuran simpul terkecil adalah 1.
ii. untuk Regresi, nilai *default* untuk *m* adalah $p/3$ dan ukuran simpul terkecil adalah 5.

Langkah 4

untuk setiap $k = 1, \dots, K,$ dari setiap dataset hasil *resampling bootstrap* disusun sebuah pohon klasifikasi T_k tanpa melalui proses pemangkasan dengan mengulangi secara rekursif setiap tahapan pada setiap *terminal node*, hingga didapatkan simpul berukuran minimum n_{min}.

Langkah 5

Untuk setiap $k = 1, \dots, K$, susun pohon dengan pemilah terbaik $H_k : D_k \rightarrow R$ menggunakan algoritma CART berdasarkan dataset training D_k. pemilihan pemilah terbaik bisa menggunakan cara yang sudah dijelaskan pada subbbab 3.3.3

Langkah 6

Untuk setiap $k = 1, \dots, K$, tambahkan pemilah terbaik tersebut ke *current ensemble*, $\mathcal{D} = \mathcal{D} \cup H_k$.

Langkah 7

Ulangi tahapan dari langkah 2 s.d. langkah 6 hingga diperoleh sejumlah *K* pohon klasifikasi yang diinginkan sehingga pada akhirnya akan didapatkan pohon sebanyak $\{T_k\}_1^K$. Secara *default*, pada `package randomForest` dalam software `R` jumlah pohon yang dibentuk sebanyak 500 pohon (Liam dan Weiner, 2006).

Langkah 8

Melakukan prediksi klasifikasi data sampel akhir dengan mengkombinasikan hasil prediksi sejumlah *K* pohon klasifikasi berdasarkan aturan *majority vote*. Pemilah gabungan *H* dibuat sebagai aggregasi dari setiap pemilah $H_k, k = 1, \dots, K$ dan contoh d_i diklasifikasikan ke kelas c_j sesuai dengan jumlah suara yang diperoleh dari pemilah tertentu H_k.

$$H(d_i, c_j) = sign\left(\sum_{k=1}^{K} H_k(d_i, c_j)\right) \quad (1)$$

Probabilitas nilai $H(d_i, c_j)$ berkisar antara 0 dan 1. Nilai *cut off* yang digunakan sebesar 0,5. Jika $H(d_i, c_j)$ < 0,5 maka akan diklasifikasikan sebagai kategori 0. Misalkan $\widehat{H}_k(d_i, c_j)$ merupakan kelas prediksi dari pohon *Random Forest* ke-*k*, maka $\widehat{H}_{rf}^K(d_i, c_j) = majority\ vote\ \{\widehat{H}_k(d_i, c_j)\}_1^K$ (Machova *et al*, 2006).

3. METODE PENELITIAN

3.1. Evaluasi Ketepatan Klasifikasi

Data simulasi dibagi menjadi dua berdasarkan *pareto law* yang membagi data dengan perbandingan 80:20, yaitu data *training* sebesar 80 persen dan data *testing* sebesar 20 persen. Sebenarnya tidak ada batasan berapa besaran proporsi yang akan digunakan sebagai data *training* ataupun data *testing* (Breiman *et al.*, 1984). Pengukuran ketepatan klasifikasi diukur melalui *total accuracy rate* (1-APER) yang dihitung berdasarkan Tabel Klasifikasi (Hosmer dan Lemeshow, 2000). 1-APER menunjukkan akurasi keseluruhan suatu klasifikasi. Bentuk umum Tabel Klasifikasi disajikan pada Tabel 1.

Tabel 1 Tabel Klasifikasi

Aktual	Prediksi		Total
	0	1	
0	n_{11}	n_{12}	$n_{1.}$
1	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Dengan,

n_{11} = jumlah observasi dari kelas 1 yang tepat diprediksi sebagai kelas 1
n_{22} = jumlah observasi dari kelas 2 yang tepat diprediksi sebagai kelas 2
n_{12} = jumlah observasi dari kelas 1 yang salah diprediksi sebagai kelas 2
n_{21} = jumlah observasi dari kelas 2 yang salah diprediksi sebagai kelas 1
$n_{1.}$ = jumlah observasi dari kelas 1
$n_{2.}$ = jumlah observasi dari kelas 2
N = jumlah observasi

Berdasarkan Tabel 1, persamaan untuk menghitung ketepatan klasifikasi berupa sensitivitas, spesifikasi, *G-means*, dan *total accuracy rate* (1-APER) adalah sebagai berikut:

$$total\ accuracy\ rate\ (1-\text{APER})\ (\text{dalam}\ \%) = \frac{n_{11}+n_{22}}{N} \times 100\% \qquad (2)$$

Kemudian pengukuran ketepatan klasifikasi berdasarkan *total accuracy rate* (1-APER) diterapkan pada klasifikasi menggunakan algoritma CART, *bagging* dan *Random Forest*.

3.2. Simulasi

Untuk memahami pengaruh jumlah ukuran sampel, korelasi antar variabel independen, ada tidaknya pola distribusi tertentu terhadap akurasi yang dihasilkan metode klasifikasi maka perlu dilakukan simulasi. Data simulasi dibangkitkan pada satu model, yaitu model yang terdiri dari 1 variabel dependen dengan 3 variabel independen $\{g(x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\}$. Simulasi dilakukan pada model dengan variabel dependen dengan respon biner serta variabel independen bertipe biner dan non biner menggunakan *Software R*.

Pembangkitan data dengan tiga variabel independen biner

Langkah-langkah pembangkitan data dengan tiga variabel independen biner (X_1, X_2 dan X_3) dapat diuraikan dalam algoritma simulasi sebagai berikut:

Langkah 1

Menetapkan set.seed(1) untuk menjamin bahwa hasil ulang simulasi memberikan hasil yang sama, kemudian menetapkan beberapa nilai korelasi antar variabel independen yang akan digunakan dalam simulasi, yaitu : 0; 0,1; 0,3; dan 0,6. Penentuan beberapa nilai korelasi tersebut berdasarkan pendapat Sugiyono dalam Fajariyanto (2017) di mana korelasi bernilai 0 dan 0,1 sebagai korelasi tingkat sangat rendah, korelasi bernilai 0,3 sebagai korelasi tingkat rendah, serta nilai korelasi 0,6 sebagai korelasi tingkat kuat.

Langkah 2

Berdasarkan nilai korelasi pada langkah 1, dengan menggunakan *package R bindata*, ketiga variabel X_1, X_2 dan X_3 dibangkitkan dari distribusi *multivariate binary*. Variabel X_1 dimisalkan sebagai variabel tingkat pendidikan (rendah, tinggi), variabel X_2 dimisalkan sebagai status perkawinan (belum kawin, kawin), dan variabel X_3 dimisalkan sebagai variabel umur (15-64, selain 15-64). Kasus khusus

dari distribusi diskrit di mana variabel hanya memiliki dua nilai, baik 0 dan 1 ataupun ya dan tidak disebut distribusi biner (Leisch *et al*., 1998). Ketiga variabel X_1, X_2 dan X_3 dibangkitkan pada beberapa ukuran sampel, yaitu dengan n = 100, 500, dan 1000.

Langkah 3

Membentuk variabel Y yang dibangkitkan dari distribusi Bernoulli dengan probabilitas ditunjukkan Persamaan (3) dengan menetapkan nilai parameter β_0 = -0,43034, β_1 = -0,88075, β_2 = 0,92194, dan β_3 = 0,06612. Penetapan nilai parameter tersebut berdasarkan hasil penelitian Fajariyanto (2017).

$$\pi(x_i) = \frac{\exp(-0{,}43034 - 0{,}88075x_1 + 0{,}92194x_2 + 0{,}06612x_3)}{1 + \exp(-0{,}43034 - 0{,}88075x_1 + 0{,}92194x_2 + 0{,}06612x_3)} \quad (3)$$

Distribusi Bernoulli adalah distribusi yang menghasilkan dua kemungkinan, yaitu sukses atau gagal. Fungsi kepadatan peluang percobaan Bernoulli adalah:

$$f(p) = \begin{cases} \pi^p(1-\pi)^{1-p}, p = 0,1 \\ 0, p \text{ lainnya} \end{cases} \quad (4)$$

Langkah 4

Membentuk data *frame* yang terdiri dari variabel Y, X_1, X_2 dan X_3 dimana setiap variabel baik variabel dependen maupun variabel independen dibuat menjadi variabel kategorik.

Langkah 5

Melakukan klasifikasi menggunakan metode klasifikasi CART. Selanjutnya ketepatan akurasi data dihitung baik pada data *training* maupun data *testing*. Kemudian dalam simulasi *bagging* dan *Random Forest* maka pada langkah ke-5 ini klasifikasi menggunakan kedua metode tersebut.

Langkah 6

Mengulangi langkah 2 sampai dengan 5 sebanyak 100 kali iterasi, kemudian menghitung nilai rata-rata akurasi yang didapatkan.

Pembangkitan data dengan tiga variabel independen non biner

Langkah-langkah pembangkitan data dengan tiga variabel independen non biner (X_1, X_2 dan X_3) dapat diuraikan dalam algoritma simulasi sebagai berikut:

Langkah 1

Menetapkan set.seed(1) untuk menjamin bahwa hasil ulang simulasi memberikan hasil yang sama, kemudian menetapkan beberapa nilai korelasi antar variabel independen yang akan digunakan dalam simulasi, yaitu : 0, 0,1, 0,3, dan 0,6. Penentuan beberapa nilai korelasi tersebut berdasarkan pendapat Sugiyono dalam Fajariyanto (2017) di mana korelasi bernilai 0 dan 0,1 sebagai korelasi tingkat sangat rendah, korelasi bernilai 0,3 sebagai korelasi tingkat rendah, serta nilai korelasi 0,6 sebagai korelasi tingkat kuat.

Langkah 2

Berdasarkan nilai korelasi pada langkah 1, dengan menggunakan *package R GenOrd*, ketiga variabel X_1, X_2 dan X_3 dibangkitkan dari distribusi *multivariate multinomial*. Variabel X_1 dimisalkan sebagai variabel tingkat pendidikan (rendah,

sedang, dan tinggi), variabel X_2 dimisalkan sebagai status perkawinan (belum kawin, kawin, cerai hidup, dan cerai mati), dan variabel X_3 dimisalkan sebagai variabel umur (15-64 tahun). Ketiga variabel X_1, X_2 dan X_3 dibangkitkan pada beberapa ukuran sampel, yaitu dengan n = 100, 500, dan 1000.

Langkah 3

Membentuk variabel Y yang dibangkitkan dari distribusi Bernoulli dengan probabilitas ditunjukkan Persamaan (4) dengan menetapkan nilai parameter β_0 = -0,43034, β_1 = -0,88075, β_2 = 0,92194, dan β_3 = 0,06612. Penetapan nilai parameter tersebut berdasarkan hasil penelitian Fajariyanto (2017).

$$\pi(x_i) = \frac{\exp(-0{,}43034 - 0{,}88075x_1 + 0{,}92194x_2 + 0{,}06612x_3)}{1 + \exp(-0{,}43034 - 0{,}88075x_1 + 0{,}92194x_2 + 0{,}06612x_3)} \quad (3)$$

Langkah 4

Membentuk data *frame* yang terdiri dari variabel Y, X_1, X_2 dan X_3 dimana setiap variabel baik variabel dependen maupun variabel independen dibuat menjadi variabel kategorik.

Langkah 5

Melakukan klasifikasi menggunakan metode klasifikasi CART. Selanjutnya ketepatan akurasi data dihitung baik pada data *training* maupun data *testing*. Kemudian dalam simulasi *Bagging* dan *Random Forest* maka pada langkah ke-5 ini klasifikasi menggunakan kedua metode tersebut.

Langkah 6

Mengulangi langkah 2 sampai dengan 5 sebanyak 100 kali iterasi, kemudian menghitung nilai rata-rata akurasi yang didapatkan.Distribusi B

4. HASIL DAN PEMBAHASAN

4.1. Simulasi Menggunakan Tiga Variabel Independen Biner

Pada subbab 4.1 ini akan dilakukan simulasi untuk klasifikasi data menggunakan metode CART, *bagging* dan *Random Forest* dengan mengacu pada persamaan (2) serta langkah-langkah simulasi seperti yang telah dijelaskan pada subbab 3.2.

Simulasi pada data dengan tiga variabel independen biner menggunakan sampel berukuran 100, 500, 1000 dan korelasi antar variabel independen 0, 0,1, 0,3, 0,6 memberikan nilai rata-rata akurasi seperti terlihat pada Tabel 2.

Tabel 2 Nilai Rata-rata Akurasi Klasifikasi pada Variabel Independen Biner Berukuran n = 100

ρ	Data *Training*			Data *Testing*		
	CART	*Bagging*	RF	CART	*Bagging*	RF
0	68,464	69,002	68,990	60,728	60,516	60,688
0,1	66,780	67,362	67,486	60,442	59,978	59,426
0,3	65,848	66,395	66,600	58,827	58,388	58,102
0,6	63,259	64,376	64,573	57,935	57,869	58,026

Tabel 2. memperlihatkan bahwa metode *bagging* memiliki akurasi *training* yang tinggi pada saat tidak terjadi korelasi yaitu sebesar 69,002 persen. Sedangkan *Random*

Forest memiliki akurasi *training* yang lebih tinggi dibandingkan CART dan *Bagging* ketika korelasi bernilai 0,1, 0,3 dan 0,6. Namun berdasarkan data *testing*, CART menghasilkan akurasi yang lebih baik dibandingkan *bagging* dan *Random Forest* pada saat korelasi bernilai 0, 0,1 dan 0,3 yaitu sebesar 60,728 persen, 60,442 persen dan 58,827 persen. Pada saat korelasi bernilai 0,6 *Random Forest* menghasilkan akurasi *testing* yang lebih baik yaitu sebesar 58,026 persen dibandingkan metode lainnya. Dalam hal ini, CART dapat menangani klasifikasi data berukuran sampel n = 100 dengan 3 variabel independen bertipe biner secara lebih baik meskipun korelasi antar variabel independen meningkat. Namun, secara umum ketiga metode menghasilkan akurasi klasifikasi yang kurang baik karena nilai akurasi masih berada di bawah 70 persen.

Dengan cara yang sama maka untuk ukuran sampel n = 500 dan n=1000 akan menghasilkan nilai rata-rata akurasi seperti tercantum pada Tabel 3 dan Tabel 4.

Tabel 3 Nilai Rata-rata Akurasi Klasifikasi pada Variabel Independen Biner Berukuran n = 500

ρ	Data *Training*			Data *Testing*		
	CART	*Bagging*	RF	CART	*Bagging*	RF
0	65,142	65,157	65,182	64,139	63,672	64,197
0,1	64,473	64,510	64,500	63,210	62,889	63,269
0,3	63,067	63,097	63,118	62,092	62,176	61,854
0,6	61,902	61,965	61,982	61,746	61,457	61,752

Tabel 4 Nilai Rata-rata Akurasi Klasifikasi pada Variabel Independen Biner Berukuran n = 1000

ρ	Data *Training*			Data *Testing*		
	CART	*Bagging*	RF	CART	*Bagging*	RF
0	64,821	64,821	64,826	64,847	64,680	64,786
0,1	64,045	64,055	64,050	64,258	63,938	64,295
0,3	63,164	63,193	63,189	63,062	62,661	62,884
0,6	61,748	61,822	61,824	61,232	61,264	61,239

4.2. Simulasi Menggunakan Tiga Variabel Independen Non Biner

Pada bagian ini akan dilakukan simulasi untuk klasifikasi data menggunakan metode CART, *Bagging* dan *Random Forest* dengan mengacu pada Persamaan (2) serta langkah-langkah simulasi seperti yang telah dijelaskan pada subbab 3.2.

Simulasi pada data dengan tiga variabel independen non biner dimana variabel X_1 dimisalkan sebagai variabel tingkat pendidikan (rendah, sedang, dan tinggi), variabel X_2 dimisalkan sebagai variabel status perkawinan (belum kawin, kawin, cerai hidup, dan cerai mati), dan variabel X_3 dimisalkan sebagai variabel umur (15-64 tahun). Simulasi menggunakan sampel berukuran 100, 500, 1000 dan korelasi antar variabel independen 0; 0,1; 0,3; dan 0,6 memberikan nilai rata-rata akurasi seperti terlihat pada Tabel 5.

Tabel 5 Nilai Rata-rata Akurasi Klasifikasi pada Variabel Independen Non Biner Berukuran n = 10

ρ	Data *Training*			Data *Testing*		
	CART	*Bagging*	RF	CART	*Bagging*	RF
0	83,930	98,192	89,686	77,594	76,595	80,533
0,1	83,741	98,065	89,338	76,481	75,614	80,222
0,3	83,709	97,564	88,764	78,284	76,294	80,051
0,6	83,854	96,276	87,895	76,933	74,213	78,502

Tabel 5. memperlihatkan bahwa metode *Bagging* memiliki akurasi *training* yang jauh lebih tinggi dibandingkan CART dan *Random Forest* di semua titik korelasi. Akurasi *training* tertinggi terjadi ketika tidak terdapat korelasi dalam data, yaitu sebesar 98,192 persen. Namun berdasarkan data *testing*, *Random Forest* menghasilkan akurasi yang lebih baik dibandingkan *Bagging* dan CART di semua titik korelasi dimana akurasi *testing* tertinggi terjadi ketika korelasi bernilai 0 yaitu sebesar 80,533 persen. Metode *Bagging* justru memiliki akurasi *testing* terendah di semua titik korelasi. Dalam hal ini, *Random Forest* dapat menangani klasifikasi data berukuran sampel n = 100 dengan 3 variabel independen bertipe non biner secara lebih baik meskipun korelasi antar variabel independen meningkat. Secara umum, ketiga metode menghasilkan akurasi klasifikasi yang baik karena nilai akurasi berada di atas 70 persen.

Dengan cara yang sama maka untuk ukuran sampel n = 500 dan n = 1000 akan menghasilkan nilai rata-rata akurasi seperti tercantum pada Tabel 6 dan Tabel 7.

Tabel 6 Nilai Rata-rata Akurasi Klasifikasi pada Variabel Independen Non Biner Berukuran n = 500

ρ	Data *Training*			Data *Testing*		
	CART	*Bagging*	RF	CART	*Bagging*	RF
0	84,113	93,496	84,292	79,833	77,250	80,662
0,1	84,070	93,447	84,064	80,099	77,048	80,976
0,3	83,858	92,603	83,848	79,274	75,902	80,646
0,6	83,247	90,327	83,018	79,176	76,901	79,712

Tabel 7 Nilai Rata-rata Akurasi Klasifikasi pada Variabel Independen Non Biner Berukuran n = 1000

ρ	Data *Training*			Data *Testing*		
	CART	*Bagging*	RF	CART	*Bagging*	RF
0	83,441	90,118	83,198	81,073	77,662	81,412
0,1	83,052	89,978	82,868	80,405	77,384	80,593
0,3	82,881	89,346	82,597	80,128	77,015	80,456
0,6	81,781	87,043	81,425	79,236	76,442	79,625

5. KESIMPULAN

Berdasarkan penghitungan akurasi pada data simulasi yang telah dilakukan, maka dapat disimpulkan bahwa ketiga metode pada variabel independen bertipe non biner menghasilkan kinerja yang lebih baik dibandingkan variabel independen bertipe biner karena menghasilkan nilai akurasi yang sangat rendah (di bawah 70 persen) untuk ketiga

metode klasifikasi. Metode *Random Forest* menghasilkan kinerja paling baik dibandingkan CART dan *Bagging* pada saat variabel independen yang digunakan berisi variabel non biner untuk setiap titik korelasi. Dengan kata lain, penggunaan metode *Random Forest* pada variabel bertipe non biner menghasilkan klasifikasi yang lebih baik dibandingkan metode CART dan *Bagging*.

DAFTAR PUSTAKA

Agresti, A. 2007. *An Introduction to Categorical Data Analysis.* Canada: John Wiley and Sons.

Breiman, L. 2001. Random Forests. *Machine Learning*, Vol.45, 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., dan Stone, C. J. 1984. *Classification and Regression trees.* Monterey: Wadsworth Press.

Fajariyanto, E. 2017. *Perbandingan Kinerja Regresi Logistik dan Neural Network dalam Pengklasifikasian Objek (Studi Kasus: Klasifikasi Angkatan Kerja di Kabupaten Kepahiang Provinsi Bengkulu).* Tesis Program Magister Statistika Terapan Fakultas MIPA, Universitas Padjajaran Bandung. Tidak dipublikasikan.

Friedman, J. H. dan Hall, P. 1999. *On Bagging and Nonlinear Estimation.* Stanford Website. Melalui < https://statweb.stanford.edu/~jhf/>. Diakses pada tanggal 15 September 2017.

Hastie, T., Tibshirani, R., dan Friedman, J. 2009. *The Elements of Statistical Learning (2nd Ed).* New York: Springer.

Hayes, T., Usami, S., Jacobucci, R. dan McArdle, J. J. 2015. Using Classification and Regression Trees (CART) and Random Forests to Analyze Attrition: Results From Two Simulations. *Psychol Aging*, 30(4), 911-929.

Hosmer, D.W. dan Lemeshow, S. 2000. *Applied Logistic Regression*. Canada: John Wiley and Sons.

Jatmiko, Y. A., Padmadisastra, S., dan Chadidjah, A. 2018. *Perbandingan Kinerja CART Konvensional, Bagging, dan Random Forest pada Klasifikasi Objek*. Tesis Universitas Padjajaran, Bandung. Tidak dipublikasikan.

Johnson, R. A. dan Winchern, D. W. 2002. *Applied Multivariate Statistical Analysis.* USA: Prentice-Hall.

Leisch, F., Weingessel, A. dan Hornik, K. 1998. *On The Generation of Correlated Artificial Binary Data.* Working Paper No. 13 June 1998.

Lewis, R. J. 2000. *An Introduction to Classification and Regression Tree (CART).* Department of Emergency Medicine Harbor – UCLA Medical Center, Torrance, California.

Liaw, A. dan Wiener, M. 2002. Classification and Regression by Random Forest. *R News*, 2, 18-22.

Machova, K., Barcak, F., dan Bednar, P. 2006. A Bagging Method using Decision Trees in the Role of Base Classifiers. *Acta Polytechnica Hungarica*. 3(2), 121-132.

Muttaqin, M. J. 2013. *Metode Ensemble pada CART untuk Perbaikan Klasifikasi Kemiskinan di Kabupaten Jombang*. Tesis Statistika Institut Teknologi Sepuluh November, Surabaya. Tidak Dipublikasikan.

Otok, B. W. dan Seftiana, D. 2015. *Klasifikasi Rumah Tangga Sangat Miskin di Kabupaten Jombang Menurut Paket Bantuan Rumah Tangga yang Diharapkan dengan Pendekatan Random Forests Classification and Regression Trees (RF-CART).* Digital Library Institut Teknologi Sepuluh November. Melalui http://digilib.its.ac.id/ITS-paper-13121150006884/37939. Diakses pada tanggal 21 Agustus 2017.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., dan Higgins, R. 2019. Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation. *Biomedical Signal Processing and Control*. 52, 456-462

Sutton, C. D. 2005. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics*, 24, 303-329.

Wezel, M.V. dan Potharst, R. 2007. Improved Customer Choice Predictions using Ensemble Methods. *European Journal of Operational Research*, 181, 436-452.