Jurnal Statistika dan Aplikasinya Volume 5 Issue 1, June 2021 e-ISSN: 2620-8369

Received: 2 Februari 2021
Revised: 21 Juni 2021
Accepted: 23 Juni 2021
Published: 30 Juni 2021

Analisis Regresi Logistik Binomial dan Algoritma Random Forest pada Proses Pengklasifikasian Penyakit Ginjal Kronis

Abraham Raja Swara Darwanto[1, a)], Taza Luzia Viarindita[1, b)] , Yekti Widyaningsih [1, c)]

[1] *Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Indonesia , Depok*

E-mail: [a)]abraham.raja@sci.ui.ac.id, [b)]taza.luzia@sci.ui.ac.id, [c)*]yekti@sci.ui.ac.id (Penulis Koresponden)

Abstract

Chronic Kidney Disease is a health problem that affects many people around the world and has been considered as an important cause of death with serious number. Chronic kidney disease is a disease that needs to be checked regularly. The authors hope that this disease can be detected with as few tests as possible and at a low cost using binomial logistic regression and classification methods Random Forest. So, the aim of this study is to compare the accuracy of both methods to see which model is the most effective one in predicting CKD. The results showed that the analysis using the binomial logistic regression method has higher accuracy level than the Random Forest classification method with an accuracy value of 97.5%, where both methods used the same factors, namely Specific Gravity, Albumin, Serum Creatinine, Hemoglobin, and Packed Cell Volume.

Keywords: Random Forest, Binomial Logistic Regression, CKD, Classification

Abstrak

Penyakit Ginjal Kronis (*Chronic Kidney Disease*) merupakan masalah kesehatan yang banyak diderita oleh masyarakat di seluruh dunia dan telah diangkat sebagai penyebab penting kematian dengan angka yang besar. Penyakit ginjal kronis termasuk penyakit yang perlu diperiksa secara rutin. Penulis berharap penyakit ini dapat dideteksi dengan tes sesedikit mungkin dan biaya yang rendah dengan menggunakan regresi logistik binomial dan metode klasifikasi *Random Forest*. Sehingga tujuan dari penelitian ini adalah ingin membandingkan akurasi dari kedua metode untuk melihat model yang paling efektif dalam memprediksi CKD. Hasil penelitian menunjukkan bahwa analisis menggunakan metode regresi logistik binomial memiliki keakuratan yang lebih tinggi dibandingkan dengan metode klasifikasi *random forest* yaitu sebesar 97.5%, di mana kedua metode menggunakan faktor-faktor yang sama yaitu *Specific Gravity*, *Albumin*, *Serum Creatinine*, *Hemoglobin*, dan kadar *Packed Cell Volume*.

Kata kunci: Random Forest, Regresi Logistik Binomial, CKD, Klasifikasi

PENDAHULUAN

Penyakit Ginjal Kronis (*Chronic Kidney Disease*) merupakan masalah kesehatan yang banyak diderita oleh masyarakat di seluruh dunia, khususnya pada negara-negara yang memiliki tingkat penghasilan rendah ataupun menengah. Penyakit ginjal kronis adalah keadaan terjadinya penurunan fungsi ginjal yang menyebabkan ginjal tidak dapat membuang racun dan produk sisa dari darah, yang ditandai adanya protein dalam urin serta penurunan laju filtrasi glomerulus. Penyakit ini bersifat progresif dan umumnya tidak dapat pulih kembali (*irreversible*). Gejala penyakit ini umumnya adalah tidak ada nafsu makan, mual, muntah, pusing, sesak nafas, rasa lelah, edema pada kaki dan tangan, serta uremia (Almatsier, 2006).

Menurut *World Health Organization* (WHO), penyakit gagal ginjal kronis berkontribusi pada beban penyakit dunia dengan angka kematian sebesar 850.000 jiwa per tahun (Pongsibidang, 2016). Menurut studi *Global Burden Disease* 2010 yang dilakukan oleh *International Society of Nephrology*, penyakit ginjal kronis telah diangkat sebagai penyebab penting kematian di seluruh dunia dengan jumlah kematian meningkat sebesar 82,3% dalam dua dekade terakhir (J. Radhakrishnan et al, 2014; R. Lozano et al, 2010). Hal-hal tersebut menunjukkan bahwa penyakit ginjal kronis memerlukan perhatian lebih, salah satunya dengan penanganan cepat melalui sistem prediksi yang akurat.

Dikatakan bahwa setiap tahun, seseorang yang memiliki salah satu faktor risiko CKD, seperti riwayat keluarga gagal ginjal, hipertensi, atau diabetes harus diperiksa. Semakin cepat mereka mengetahui tentang penyakit ini, maka semakin cepat pula mereka bisa mendapatkan pengobatan. Dalam mendeteksi penyakit ginjal, terdapat pemeriksaan fungsi ginjal yang dilakukan secara rutin ataupun sifatnya hanya pemeriksaan tambahan. Untuk kasus tes secara rutin/berkala, penulis berharap penyakit ini dapat dideteksi dengan tes sesedikit mungkin dan dengan biaya yang rendah. Sehingga tujuan dari penelitian ini adalah untuk memberikan model yang efektif untuk memprediksi CKD dengan jumlah prediktor yang paling sedikit.

METODOLOGI

Data

Dataset yang digunakan diperoleh dari website UCI Machine Learning Repository. Dataset ini berisi kumpulan pasien CKD dari Rumah Sakit Apollo, India pada tahun 2015 yang diambil selama periode dua bulan. Data-data yang terdiri dari 400 observasi ini memiliki *missing* dan noisy values di dalamnya. Data tersebut meliputi 250 rekam medis penderita CKD dan 150 rekam medis tanpa CKD. Sehingga dapat dihitung persentase masing-masing kelas adalah 62,5% dengan CKD dan 37,5% tanpa CKD. Usia pengamatan ini bervariasi dari 2 hingga 90 tahun. Deskripsi dari dataset CKD dijelaskan pada TABEL 1.

TABEL 1. Deskripsi Variabel-variabel

Nama	*Deskripsi*	*Tipe*
Age (age)	Umur pasien	Numerik : tahun
Blood pressure (bp)	Tekanan darah pasien	Numerik : mm/Hg
Specific gravity (sg)	Rasio kepadatan urin	Nominal : 1.005, 1.010, 1.015, 1.020,1.025
Albumin (al)	Tingkat albumin dalam darah	Nominal: 0,1,2,3,4,5
Sugar (su)	Tingkat gula pasien	Nominal: 0,1,2,3,4,5
Red blood cells (rbc)	Jumlah sel darah merah pasien	Nominal:normal,abnormal
Pus cell (pc)	Jumlah sel nanah pasien	Nominal:normal,abnormal
Pus cell clumps (pcc)	Adanya gumpalan sel nanah dalam darah	Nominal: ada, tidak ada
Bacteria (ba)	Adanya bakteri dalam darah	Nominal: ada, tidak ada
Blood glucose (bgr)	Jumlah kandungan glukosa dalam darah	Numeric: mgs/dl
Blood urea (bu)	Tingkat urea darah pasien	Numeric: mgs/dl
Serum creatinine (sc)	Tingkat kreatinin serum dalam darah	Numeric: mgs/dl
Sodium (sod)	Tingkat sodium dalam darah	Numeric: mEq/L
Potassium (pot)	Tingkat kalium dalam darah	Numeric: mEq/L
Hemoglobin (hemo)	Tingkat kalium dalam darah	Numeric: gms
Packed cell volume (pcv)	Kadar *packed cell volume*	Numeric
White blood cell count (wc)	Jumlah sel darah putih pasien	Numeric: cells/cumm
Red blood cell count (rc)	Jumlah sel darah merah pasien	Numeric millions/cmm
Hypertension (htn)	Apakah pasien memiliki hipertensi/ tidak	Nominal: ya, tidak
Diabetes mellitus (dm)	Apakah pasien memiliki diabetes/ tidak	Nominal: ya, tidak
Coronary artery disease (cad)	Apakah pasien memiliki penyakit arteri koroner/ tidak	Nominal: ya, tidak
Appetite (appet)	Nafsu makan pasien	Nominal: bagus, buruk
Pedal Edema (pe)	Apakah pasien memiliki pedal edema/ tidak	Nominal: ya, tidak
Anemia (ane)	Apakah pasien memiliki anemia/ tidak	Nominal: ya, tidak
Class	Apakah pasien memiliki penyakit ginjal atau tidak	Nominal:CKD, tidak CKD

Metode Penelitian

Imputasi Data

Permasalahan yang sering muncul pada suatu data adalah adanya ketidaklengkapan data pada suatu variabel atau sering disebut dengan *missing data*. Adanya *missing data* menyebabkan kesulitan dalam melakukan analisis terhadap data tersebut karena analisis statistik hanya dapat diterapkan pada data yang lengkap. Untuk menangani *missing data*, salah satu metode yang dapat digunakan adalah dengan mengestimasi nilai *missing* dengan suatu nilai tertentu yang dianggap sesuai, atau sering disebut dengan imputasi.

Terdapat beberapa metode yang dapat digunakan untuk mengestimasi nilai dari *missing data* tersebut yang dikelompokkan menjadi dua, yaitu metode tradisional dan metode modern. Metode modern muncul karena keterbatasan dari metode tradisional. Salah satu dari metode modern, yaitu metode *Multivariate Imputation by Chained Equations* (MICE). MICE dikenal juga dengan "*Fully Conditional Specification*" atau "*Sequential Regression Multiple Imputation*" (Azur, 2011). MICE mengubah masalah imputasi menjadi serangkaian estimasi di mana setiap variabel mendapat giliran untuk diregresikan pada variabel lain.

MICE berjalan melalui proses iteratif: Pada iterasi pertama, model imputasi untuk variabel dengan *missing values* yang paling sedikit diperkirakan hanya menggunakan data lengkap. Selanjutnya, variabel dengan *missing values* kedua tersedikit dihitung menggunakan data lengkap dan nilai yang diperhitungkan dari iterasi terakhir. Setelah setiap variabel melalui proses ini, siklus diulangi menggunakan data dari iterasi terakhir. Biasanya, sepuluh iterasi dilakukan di mana nilai yang diperhitungkan setelah iterasi ke-10 dan terakhir merupakan satu set data yang diperhitungkan (Stuart, E.A. et al, 2009).

Seperti disebutkan, MICE memiliki kemampuan penting untuk menangani jenis variabel yang berbeda karena setiap variabel diperhitungkan menggunakan model imputasinya sendiri (Bartlett et al. 2014). Proses ini memberikan fleksibilitas dan memungkinkan untuk menghubungkan kumpulan data yang mencakup ratusan variabel.

Embedded Feature Selection

Feature selection merupakan bagian penting untuk mengoptimalisasi kinerja dari pengklasifikasian (Wang et al., 2011). Apabila menggunakan *dataset* yang cukup besar, eliminasi atribut yang kurang relevan menggunakan algoritma *feature selection* yang tepat dapat meningkatkan akurasi dan mempercepat proses *learning* (Dash et al., 1997). Algoritma *feature selection* dapat dibedakan menjadi dua tipe, yaitu *filter* dan *wrapper* (Yuanning Liu et al., 2011). Metode *wrapper* mengukur "kebergunaan" fitur/variabel berdasarkan performa pengklasifikasian. Sebaliknya, metode *filter* menggunakan sifat-sifat intrinsik dari variabel (relevansi fitur) yang diukur menggunakan statistik univariat, bukan berdasarkan performa. Artinya, metode *wrapper* pada dasarnya mengoptimalkan kinerja pengklasifikasi), tetapi secara komputasi juga akan menjadi lebih mahal dibandingkan dengan metode *filter* karena langkah-langkah proses *learning* yang berulang (Guyon et al., 2003). Metode *embedded feature selection* menggabungkan kualitas dari kedua metode tersebut, yang diimplementasikan oleh algoritma yang mempunyai metode *feature selection* mereka sendiri. Salah satu *feature selection* oleh algoritma *Random Forest*, metode tersebut cocok untuk proses pengklasifikasian biner dengan variabel bebas kategorik maupun numerik. Dalam metode *feature selection* ini, dilakukan perbandingan terhadap besar *Gini Mean Decrease*. Dalam algoritma pengklasifikasian pohon, *Gini Mean Decrease* adalah rata-rata penurunan total sebuah variabel pada *node impurity*-nya, dibobot dengan proporsi sampel yang mencapai node tersebut di setiap individu *decision tree* pada *Random forest*-nya. Ini secara efektif mengukur seberapa penting variabel untuk memperkirakan nilai dari variabel target di semua pohon. *Gini Mean Decrease* yang lebih tinggi menunjukkan kepentingan variabel yang lebih tinggi.

Model Regresi Logistik

Menurut Hosmer dan Lemeshow (2000), regresi logistik adalah suatu metode yang dapat digunakan untuk mencari hubungan antara variabel respon yang bersifat dichotomus (skala nominal/ordinal dengan dua kategori) dengan satu atau lebih variabel prediktor berskala kategori atau kontinu. Fungsi kepekatan peluang (fkp) bagi variabel random yang terdistribusi logistik adalah:

$$f(x) = \frac{exp\left\{\frac{x-\mu}{\tau}\right\}}{\tau\left[1+exp\left\{\frac{x-\mu}{\tau}\right\}\right]^2}, -\infty \leq x \leq \infty; \tau > 0$$

dengan mean μ dan variansi $\sigma^2 = \frac{\pi^2\tau^2}{3}$

Model regresi logistik terdiri dari regresi logistik dengan respon biner, ordinal, dan multinomial. Regresi logistik biner adalah suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (*y*) yang bersifat biner (*dichotomous*) dengan variabel prediktor (*x*) yang bersifat kategorik atau kontinu. Hasil respon variabel dichotomus memiliki dua kriteria, yaitu

$y = 1$ mewakili kemungkinan sukses dengan probabilitas $p(x)$;
$y = 0$ mewakili kemungkinan gagal dengan probability $1 - p(x)$,

dimana variabel respon (y) mengikuti distribusi Bernoulli untuk setiap observasi tunggal.

Pada regresi logistik dapat disusun model yang terdiri dari banyak variabel prediktor, dikenal sebagai model multivariabel. Rata-rata bersyarat dari y jika diberikan nilai x adalah $\mathrm{p}(x) = E(y \mid x)$. Model regresi logistik multivariabel dengan p variabel prediktor adalah sebagai berikut.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

dimana, $p =$ banyaknya variabel prediktor

Dengan menggunakan transformasi logit dari p (x) untuk mempermudah pendugaan parameter regresi yang dirumuskan sebagai berikut.

$$\pi(x)\{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}\} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

$$\pi(x) + \{\pi(x)\, e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}\} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

$$\pi(x) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} - \pi(x)\, e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

$$\pi(x) = \{1-\pi(x)\}e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$$

$$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = ln\,[e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}]$$

$$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

Sehingga diperoleh persamaan sebagai berikut.

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p \qquad (2.1)$$

$g(x)$ disebut dengan fungsi logit model regresi logistik biner dengan p variabel prediktor. Model regresi logistik pada Persamaan (2.1) dapat dituliskan dalam bentuk:

$$\pi(x) = \frac{exp\,(g(x))}{1+exp\,(g(x))}$$

Penaksiran Parameter Model Regresi Logistik

Menurut Hosmer dan Lemeshow (2000) serta McCullagh dan Nelder 20 (1989) metode estimasi yang digunakan adalah metode *maximum likelihood* (MLE). Nilai parameter γ dapat diperoleh dengan memaksimumkan fungsi likelihood-nya. Hal tersebut dilakukan dengan metode turunan pertama fungsi likelihood-nya terhadap setiap parameter yang disamakan dengan nol. Terkadang sulit menemukan turunan dari fungsi likelihood-nya sehingga yang dilakukan adalah menemukan nilai maksimum dari logaritma natural fungsi likelihood tersebut atau fungsi log-likelihood.

Fungsi log-likelihood adalah bentuk logaritma dari fungsi likelihood, yang dituliskan dalam bentuk:

$$ln\,L(\,\pi(x)|y_1,\ldots,y_n) = \sum_{i=1}^{n} ln\,(\pi(x)^{y_i}(1-\pi(x)^{1-y_i})$$

Berikut langkah-langkah dalam menentukan penduga parameter dengan metode *Maximum Likelihood Estimation*:

1. Menentukan fungsi *likelihood*

$$L(\pi(x)|y_1, \dots, y_n) = \prod_{i=1}^{n} \pi(x)^{y_i}(1-\pi(x)^{1-y_i}$$

2. Menentukan fungsi *log-likelihood*

$$ln\, L(\,\pi(x)|y_1, \dots, y_n) = \sum_{i=1}^{n} ln\, (\pi(x)^{y_i}(1-\pi(x)^{1-y_i})$$

3. Memaksimumkan fungsi *log-likelihood*

   Memaksimumkan fungsi *log-likelihood* untuk memperoleh nilai $\hat{\gamma}$ dapat dilakukan dengan langkah-langkah berikut:

   a) Nilai $\hat{\gamma}$ diperoleh dari turunan pertama dengan disamadengankan nol

$$\left|\frac{\partial l}{\partial \beta}\right|_{\gamma=\hat{\gamma}} = 0$$

   b) Nilai $\hat{\gamma}$ dikatakan memaksimumkan $L(\gamma)$ jika

$$\left|\frac{\partial^2 l}{\partial^2 \gamma}\right|_{\gamma=\hat{\gamma}} < 0$$

4. Menyelesaikan fungsi log-likelihood yang diperoleh pada langkah 2 atau 3 dan mendapatkan $\hat{\gamma}$ sebagai estimator *Maximum Likelihood Estimation*.

Likelihood Ratio Test

Uji Rasio Likelihood diperoleh dengan cara membandingkan fungsi *log-likelihood* dari seluruh variabel bebas dengan fungsi *log-likelihood* tanpa variabel bebas. Uji Rasio Likelihood digunakan untuk menguji kelayakan model yang diperoleh dari estimasi parameter, bertujuan untuk mengetahui apakah variabel prediktor yang terdapat dalam model berpengaruh nyata atau tidak secara keseluruhan.

Goodness of Fit Test

Hosmer and Lemeshow's Goodness of Fit Test digunakan untuk menguji kelayakan model regresi. *Hosmer and Lemeshow's Goodness of Fit Test* menguji data empiris cocok atau sesuai dengan model (tidak ada perbedaan antara model dengan data sehingga model dapat dikatakan *fit)*. Pengujiannya dapat melihat nilai dari *Hosmer and Lemeshow's Goodness of Fit Test.*

Model Klasifikasi Random Forest

Metode *random forest* menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection*. Dalam *random forest*, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas *n* pengamatan dan *p* peubah penjelas, *random forest* dilakukan dengan cara (Breiman, 2001; Breiman & Cutler, 2003):

1. Lakukan penarikan contoh acak berukuran *n* dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan *bootstrap.*
2. Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih *m* peubah penjelas secara acak, dimana $m < p$. Pemilah terbaik dipilih dari *m* peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.
3. Ulangi langkah 1 dan 2 sebanyak *k* kali, sehingga terbentuk sebuah hutan yang terdiri atas *k* pohon.

Respons suatu amatan diprediksi dengan menggabungkan (*aggregating*) hasil prediksi *k* pohon. Pada masalah klasifikasi dilakukan berdasarkan *majority vote* (suara terbanyak). *Error* klasifikasi *random forest* diduga melalui *error* OOB yang diperoleh dengan cara (Breiman 2001; Breiman & Cutler 2003; Liaw & Wiener, 2002):

1. Lakukan prediksi terhadap setiap data OOB pada pohon yang bersesuaian. Data OOB (*Out of Bag)* adalah data yang tidak termuat dalam contoh *bootstrap.*
2. Secara rata-rata, setiap amatan gugus data asli akan menjadi data OOB sebanyak sekitar 36% dari banyak pohon. Oleh karena itu, pada langkah 1, masing-masing amatan gugus data asli mengalami prediksi sebanyak sekitar sepertiga kali dari banyaknya pohon. Jika *a* adalah sebuah amatan dari gugus data asli, maka hasil prediksi *random forest* terhadap *a* adalah gabungan dari hasil prediksi setiap kali *a* menjadi data OOB.
3. *Error* OOB dihitung dari proporsi misklasifikasi hasil prediksi *random forest* dari seluruh amatan gugus data asli.

Breiman dan Cutler (2003) menyarankan untuk mengamati *error* OOB saat dan *k* kecil, lalu memilih *m* yang menghasilkan *error* OOB terkecil. Jika *random forest* dilakukan dengan menghasilkan *variable importance*, disarankan untuk menggunakan banyak pohon, misalnya 1000 pohon atau lebih. Jika peubah penjelas yang dianalisis sangat banyak, nilai tersebut dapat lebih besar agar *variable importance* yang dihasilkan semakin stabil.

HASIL DAN PEMBAHASAN

Pra-pemrosesan Data

Langkah pertama yang harus dilakukan dalam proses pengklasifikasian adalah dengan mengeksplor dan memahami dataset yang digunakan hingga data tersebut siap untuk dilanjutkan pada proses pemodelan, hal ini disebut pra-pemrosesan data.

Pada dataset *chronic kidney disease*, terdapat permasalahan terkait *missing values*, yakni terdapat 10% dari seluruh nilai yang meliputi 24 variabel bebas pada dataset tersebut yang tidak tersedia, TABEL 2 menunjukkan bahwa 24 dari 25 variabel memiliki setidaknya sebuah *missing* value. Karena itu perlu dilakukan imputasi terhadap nilai-nilai tersebut, pada penelitian ini digunakan metode *Multivariate Imputation by Chained Equations* (MICE), adapun *software* yang digunakan pada penelitian ini adalah Rstudio 4.0.3. Pada GAMBAR 1 ditunjukkan 30 observasi pertama dari dataset yang telah diimputasi, yang akan digunakan untuk proses selanjutnya.

TABEL 2. Persentase *Missing Values* pada Variabel-variabel Independen

Nama	Persentase *Missing Values*	Nama	Persentase *Missing Values*
Age (age)	2.3%	Sodium (sod)	22.0%
Blood pressure (bp)	3.0%	Potassium (pot)	22.5%
Specific gravity (sg)	11.8%	Hemoglobin (hemo)	13.0%
Albumin (al)	11.5%	Packed cell volume (pcv)	17.8%
Sugar (su)	12.3%	White blood cell count (wc)	26.5%
Red blood cells (rbc)	38.0%	Red blood cell count (rc)	32.8%
Pus cell (pc)	16.3%	Hypertension (htn)	0.5%
Pus cell clumps (pcc)	1.0%	Diabetes mellitus (dm)	0.5%
Bacteria (ba)	1.0%	Coronary artery disease (cad)	0.5%
Blood glucose (bgr)	11.0%	Appetite (appet)	0.3%
Blood urea (bu)	4.8%	Pedal Edema (pe)	0.3%
Serum creatinine (sc)	4.5%	Anemia (ane)	0.3%

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
1	48	80	1.02	1	0	normal	normal	notpreser	notpreser	121	36	1.2	138	4.1	15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
2	7	50	1.02	4	0	normal	normal	notpreser	notpreser	261	18	0.8	132	4.4	11.3	38	6000	4.5	no	no	no	good	no	no	ckd
3	62	80	1.01	2	3	normal	normal	notpreser	notpreser	423	53	1.8	124	4.2	9.6	31	7500	4.5	no	yes	no	poor	no	yes	ckd
4	48	70	1.005	4	0	normal	abnormal	present	notpreser	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
5	51	80	1.01	2	0	normal	normal	notpreser	notpreser	106	26	1.4	135	3.9	11.6	35	7300	4.6	no	no	no	good	no	no	ckd
6	60	90	1.015	3	0	abnormal	abnormal	notpreser	notpreser	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
7	68	70	1.01	0	0	abnormal	normal	notpreser	notpreser	100	54	24	104	4	12.4	36	8500	4.7	no	no	no	good	no	no	ckd
8	24	80	1.015	2	4	normal	abnormal	notpreser	notpreser	410	31	1.1	135	4.5	12.4	44	6900	5	no	yes	no	good	yes	no	ckd
9	52	100	1.015	3	0	normal	abnormal	present	notpreser	138	60	1.9	142	2.7	10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
10	53	90	1.02	2	0	abnormal	abnormal	present	notpreser	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
11	50	60	1.01	2	4	normal	abnormal	present	notpreser	490	55	4	135	4.4	9.4	28	9000	3.6	yes	yes	no	good	no	yes	ckd
12	63	70	1.01	3	0	abnormal	abnormal	present	notpreser	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
13	68	70	1.015	3	1	normal	normal	present	notpreser	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
14	68	70	1.02	3	0	normal	normal	notpreser	notpreser	98	86	4.6	135	3.4	9.8	27	6200	3	yes	yes	yes	poor	yes	no	ckd
15	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
16	40	80	1.015	3	0	normal	normal	notpreser	notpreser	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
17	47	70	1.015	2	0	abnormal	normal	notpreser	notpreser	99	46	2.2	138	4.1	12.6	37	6900	5.4	no	no	no	good	no	no	ckd
18	47	80	1.005	2	0	normal	normal	notpreser	notpreser	114	87	5.2	139	3.7	12.1	37	9800	4.3	yes	no	no	poor	no	no	ckd
19	60	100	1.025	0	3	abnormal	normal	notpreser	notpreser	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
20	62	60	1.015	1	0	normal	abnormal	present	notpreser	100	31	1.6	138	4.9	10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
21	61	80	1.015	2	0	abnormal	abnormal	notpreser	notpreser	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
22	60	90	1.01	3	5	abnormal	normal	notpreser	notpreser	264	180	76	4.5	3.5	10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd
23	48	80	1.025	4	0	normal	abnormal	notpreser	notpreser	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
24	21	70	1.01	0	0	abnormal	normal	notpreser	notpreser	100	18	6	134	3.7	8.1	28	9600	4	no	no	no	poor	no	yes	ckd
25	42	100	1.015	4	0	normal	abnormal	notpreser	present	172	50	1.4	129	4	11.1	39	8300	4.6	yes	no	no	poor	no	no	ckd
26	61	60	1.025	0	0	normal	normal	notpreser	notpreser	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes	no	good	no	yes	ckd
27	75	80	1.015	0	0	abnormal	normal	notpreser	notpreser	156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes	no	poor	no	no	ckd
28	69	70	1.01	3	4	normal	abnormal	notpreser	notpreser	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes	yes	good	yes	no	ckd
29	75	70	1.01	1	3	abnormal	normal	notpreser	notpreser	123	31	1.4	138	4.2	12	43	7500	5.4	no	yes	no	good	no	no	ckd
30	68	70	1.005	1	0	abnormal	abnormal	present	notpreser	99	28	1.4	142	4.7	12.9	38	7900	4.2	no	no	yes	good	no	no	ckd

GAMBAR 1. Dataset dengan Imputasi pada *Missing Values*-nya

Embedded Feature Selection

Proses feature selection dilakukan untuk memilih variabel-variabel prediktor yang paling penting. Pada penelitian ini digunakan metode embedded feature selection yang didasari oleh metode pengklasifikasian *Random Forest* dalam prosesnya. Dari 24 variabel bebas, terpilih 5 variabel bebas yang memiliki nilai *Gini Mean Decrease* tertinggi yang dapat dilihat pada TABEL 3, yakni *Specific Gravity* (sg), *Albumin* (al), *Serum Creatinine* (sc), *Hemoglobin* (hemo), dan *Packed Cell Volume* (pcv).

TABEL 3. Perbandingan Variable Importance Berdasarkan Mean Gini Decrease

Variabel	*Mean Gini Decrease*	Variabel	*Mean Gini Decrease*
age	0.969091351	sod	1.315447216
bp	1.523597331	pot	0.78949915
sg	17.05708495	hemo	33.31265551
al	12.98506356	pcv	21.67107805
su	1.686800653	wbcc	1.008336634
rbc	7.019265968	rbcc	8.44805915
pc	0.726965387	htn	4.076832581
pcc	0.014266275	dm	4.545965148
ba	0.031380309	cad	0.011407527
bgr	6.01135625	appet	0.919093742
bu	4.947979447	pe	1.648597494
sc	18.38906104	ane	0.108562962

Pemodelan Klasifikasi dengan Regresi Logistik

Model Regresi Logistik

Model regresi logistik yang digunakan berbentuk $y_i = E(y_i) + \varepsilon_i$ untuk $i = 1,2, \dots, n$, di mana y_i adalah variabel dependen biner yang menyatakan apakah pasien pada observasi ke-i terkena penyakit ginjal kronis atau tidak, dan dengan definisi $E(y_i)$ adalah sebagai berikut,

$$E(y_i) = \hat{y} = \hat{\pi}_i = \frac{exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})}{1 + exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})}$$

$$= \frac{1}{1 + exp\,(-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}))}$$

Sehingga model regresi logistik untuk pemodelan variabel dependen biner yang menyatakan apakah pasien pada observasi ke-i terkena penyakit ginjal kronis atau tidak adalah sebagai berikut:

$$E(y_i) = \frac{1}{1 + exp\,(-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}))}$$

dimana,

x_1 = *Specific Gravity* ; x_2 = *Albumin* ; x_3 = *Serum Creatinine* ; x_4 = *Hemoglobin;* x_5 = *Packed Cell Volume*

Model taksiran regresi logistik untuk pemodelan penyakit ginjal kronis ini dapat diperoleh dengan penaksiran parameter $\beta_1, \beta_2, \beta_3, \beta_4$, dan β_5. TABEL 4 merupakan ringkasan taksiran statistik yang diperoleh menggunakan *software* Rstudio.

TABEL 4. Ringkasan Model Regresi Logistik untuk Pengklasifikasian Penderita Penyakit Ginjal Kronis

Parameter	Taksiran	*Standard Error*	Nilai *z*	Signifikansi
Konstanta	-577.094	153.4983	-3.76	Signifikan
Specific Gravity	548.8269	149.1406	3.68	Signifikan
Albumin	-15.445	191.1233	-0.08	Tidak Signifikan
Serum Creatinine	-3.9636	1.1448	-3.462	Signifikan
Hemoglobin	1.2518	0.4668	2.682	Signifikan
Packed Cell Volume	0.1259	0.1252	1.006	Tidak Signifikan

Dari lima variabel independen yang digunakan pada model, terdapat dua variabel independen yang tidak signifikan, karena itu peneliti mempertimbangkan untuk tidak mengikutsertakan dua variabel tersebut yaitu *Albumin* dan *Packed Cell Volume*, karena itu pada bagian selanjutnya akan dilakukan pengujian *Likelihood Ratio* untuk membandingkan model awal dengan model tanpa variabel independen *Albumin* dan *Packed Cell Volume*.

Uji Likelihood Ratio

Pada pengujian ini, akan diuji apakah model tereduksi lebih baik daripada model lengkap yang didefinisikan, dalam hal ini model lengkap didefinisikan sebagai berikut:

$$\widehat{y_i} = \frac{1}{1 + exp\,(-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}))}$$

Di mana y_i = *Chronic Kidney Disease* ; x_1 = *Specific Gravity* ; x_2 = *Albumin* ; x_3 = *Serum Creatinine* ; x_4 = *Hemoglobin* ; x_5 = *Packed Cell Volume*

Dan model reduksi adalah sebagai berikut:

$$\widehat{y_i} = \frac{1}{1 + exp\,(-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}))}$$

Di mana y_i = *Chronic Kidney Disease* ; x_1 = *Specific Gravity* ; x_2 = *Serum Creatinine* ; x_3 = *Hemoglobin*

Prosedur yang akan dilakukan untuk melakukan pengujian tersebut adalah sebagai berikut:

a) Hipotesis
   H_0: *reduced model* lebih baik daripada *complete model*
   H_1: *complete model* lebih baik daripada *reduced model*
b) Tingkat Signifikansi
   $\alpha = 0.05$
c) Statistik Uji
   $\Delta G^2 = -2\,log\,L\,complete\,model - (-2\,log\,L\,reduced\,model)$
   $\Delta G^2 \sim \chi_k^2$

Model	Nilai $logLikelihood$	ΔG^2	p-value
Model Tereduksi	-24.877	11.874	0.00264
Model Lengkap	-18.940		

d) Aturan Keputusan
H_0 ditolak jika p-value $< \alpha = 0.05$
Tolak H_0 karena p-value $= 0.00264 < \alpha$

e) Kesimpulan
Model lengkap lebih baik daripada model yang direduksi.

Dari hasil tersebut, model regresi logistik yang digunakan dalam pengklasifikasian adalah model lengkap, di mana model taksirannya diperoleh berdasarkan TABEL 3 sebagai berikut:

$$\widehat{y_i} = \frac{1}{1 + exp\,(-(577.094 + 548.827x_{1i} - 15.445x_{2i} - 3.9636x_{3i} + 1.2518x_{4i} + 0.126x_{5i}))}$$

Uji Goodness of Fit

Uji *Hosmer and Lemeshow's Goodness of Fit* akan dilakukan untuk menguji apakah model regresi logistik cocok untuk digunakan. Prosedur yang akan dilakukan untuk melakukan pengujian tersebut adalah sebagai berikut.

a) Hipotesis
H_0: Tidak ada perbedaan signifikan antara model dengan nilai observasi (model cocok)
H_1: Tidak demikian

b) Tingkat Signifikansi
$\alpha = 0.05$

c) Statistik Uji

$$X_{HL}^2 = \sum_{i=1}^{g} \frac{(O_i - N_i\bar{\pi}_i)^2}{N_i\bar{\pi}_i(1 - \bar{\pi}_i)}$$

Uji *Hosmer and Lemeshow goodness of fit (GOF)*

X_{HL}^2	p-value
1.1047	0.9975

d) Aturan Keputusan
H_0 ditolak jika p-value $< \alpha = 0.05$
Tidak menolak H_0 karena p-value $= 0.9975 > \alpha$

e) Kesimpulan
Tidak ada perbedaan signifikan antara model dengan nilai observasi (model cocok).

Hasil Klasifikasi Regresi Logistik

Dengan menggunakan metode *Holdout Cross Validation 80:20,* peluang ketepatan klasifikasi regresi logistik penderita penyakit ginjal kronis pada data *training* dapat dilihat pada TABEL 5, dan peluang ketepatan klasifikasi penderita penyakit ginjal kronis pada data *testing* dapat dilihat pada TABEL 6.

TABEL 5. Hasil Klasifikasi Model Regresi Logistik Binomial pada Data *Training*

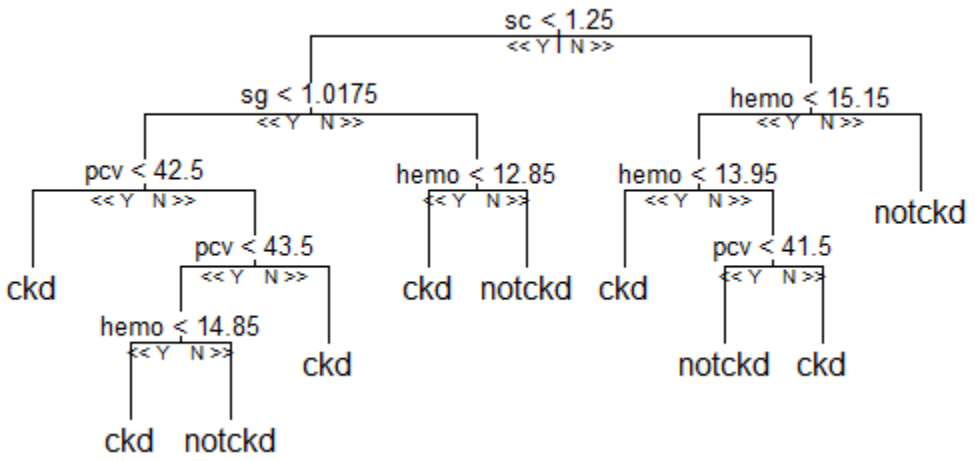
Observasi	Prediksi		Persentase Kebenaran Klasifikasi
	Menderita CKD	Tidak Menderita CKD	
Menderita CKD	194	5	97.49%
Tidak menderita CKD	6	115	95%
Total Persentase	97%	95.83%	96.56%

TABEL 6. Hasil Klasifikasi Model Regresi Logistik Binomial pada Data *Testing*

Observasi	Prediksi		Persentase Kebenaran Klasifikasi
	Menderita CKD	Tidak Menderita CKD	
Menderita CKD	50	2	96.15%
Tidak menderita CKD	0	28	100%
Total Persentase	100%	93.33%	97.50%

Pemodelan Klasifikasi dengan *Random Forest*

Pada tahapan pemodelan dengan metode *random forest*, 200 kemungkinan jumlah pohon dicoba untuk menentukan model yang optimal dengan variabel *Specific Gravity* (sg), *Albumin* (al), *Serum Creatinine* (sc), *Hemoglobin* (hemo), dan *Packed Cell Volume* (pcv), dan variabel CKD (class) sebagai respon. Bentuk dari pohon klasifikasi yang diperoleh menggunakan metode *random forest* dalam pengklasifikasian penderita penyakit ginjal kronis ditunjukkan pada GAMBAR 2, dengan estimasi OOB (*Out of Bag) error rate* sebesar 1.56%.



GAMBAR 2. Pohon Klasifikasi Model *Random Forest* pada Pengklasifikasian Penderita Penyakit Ginjal Kronis

Model di atas dibentuk oleh data *training* yang diperoleh menggunakan *holdout cross validation* dengan rasio 80:20. Kemudian pada data *testing* dilakukan klasifikasi pada model *random forest* yang telah diperoleh. Hasil klasifikasi metode *random forest* terhadap penderita penyakit ginjal kronis pada data *training* dapat dilihat pada TABEL 7, dan peluang ketepatan klasifikasi penderita penyakit ginjal kronis pada data *testing* dapat dilihat pada TABEL 8.

TABEL 7. Hasil Klasifikasi Model *Random Forest* pada Data *Training*

Observasi	Prediksi		Persentase Kebenaran Klasifikasi
	Menderita CKD	Tidak Menderita CKD	
Menderita CKD	198	2	99%
Tidak menderita CKD	3	117	97.5%
Total Persentase	98.51%	98.32%	98.48%

TABEL 8. Hasil Klasifikasi Model *Random Forest* pada Data *Testing*

Observasi	Prediksi		Persentase Kebenaran Klasifikasi
	Menderita CKD	Tidak Menderita CKD	
Menderita CKD	49	2	96.08%
Tidak menderita CKD	1	28	96.55%
Total Persentase	98%	93.33%	96.25%

Perbandingan Hasil Klasifikasi

Regresi logistik binomial dan metode *random forest* merupakan dua metode yang dapat digunakan sebagai metode pengklasifikasian, sehingga dapat dilakukan perbandingan pada ketepatan hasil klasifikasi kedua metode tersebut untuk bisa menyimpulkan mana metode yang lebih baik. Dapat dilihat pada TABEL 9 bahwa baik metode regresi logistik binomial maupun metode random forest menghasilkan akurasi yang sangat baik yakni lebih besar dari 95%.

TABEL 9. Perbandingan Hasil Klasifikasi Model Regresi Logistik dan *Random Forest*

Observasi	Persentase Ketepatan Klasifikasi			
	Regresi Logistik Binomial		*Random Forest*	
	Data *Training*	Data *Testing*	Data *Training*	Data *Testing*
Menderita CKD	97.49%	96.15%	99%	96.08%
Tidak Menderita CKD	95%	100%	97.5%	96.55%
Akurasi Keseluruhan	96.56%	97.5%	98.48%	96.25%

Hasil yang diperoleh pada TABEL 9 menunjukkan bahwa pada data *training*, hasil klasifikasi *random forest* lebih akurat sebesar 1.92%, namun sebaliknya pada data *testing* diperoleh bahwa akurasi metode regresi logistik sedikit lebih tinggi dibanding random forest dalam pengklasifikasian penderita penyakit ginjal kronis yakni 97.5%.

KESIMPULAN

Penelitian ini menunjukkan bahwa analisis regresi logistik binomial maupun algoritma *Random Forest* dapat digunakan dalam pengklasifikasian penderita penyakit ginjal kronis dengan keakuratan yang lebih tinggi diperoleh dengan metode regresi logistik binomial yaitu 97.5%, di mana kedua metode menggunakan faktor-faktor yang sama yaitu *Specific Gravity, Albumin, Serum Creatinine, Hemoglobin*, dan kadar *Packed Cell Volume*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada University of California dan Rumah Sakit Apollo India yang telah menyediakan data untuk penelitian ini, juga kepada Departemen Matematika Universitas Indonesia sebagai tempat kami memperoleh setiap ilmu pengetahuan yang kami terapkan pada penelitian ini. Selain itu penulis juga berterima kasih kepada pihak-pihak lain yang telah ikut menyumbangkan ide, saran, dan dukungan sehingga penelitian ini dapat diselesaikan.

REFERENSI

Almatsier, S. 2006. *Prinsip Dasar Ilmu Gizi*. Gramedia Pustaka Utama: Jakarta

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple Imputation by Chained Equation: What is it and how does it work? *International Journal of Methods in Psychiatric Research,* 40-49.

Bartlett, J.W. et al., 2014. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical methods in medical research,* 24(4),pp.462 487.

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

Breiman L, Cutler A. 2003. Manual on Setting Up, Using, and Understanding Random Forest V4.0.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(3), 131-156.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research,* 3(Mar), 1157-1182.

Hosmer, D.W., & Lemeshow (2000). *Applied Logistic Regression*, John Wiley and Sons. USA.

J. Radhakrishnan et al, "Taming the chronic kidney disease epidemic: a global view of surveillance efforts," *Kidney Int.*2014, vol. 86, (2), pp. 246-250, 2014.

Liaw A, Wiener M. Des 2002. Classification and Regression by randomForest. *RNews Vol.* 2/3:1822.

Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An Improved Particle Swarm Optimization for Feature Selection. *Journal of Bionic Engineering*, 8(2), 191 200.

Pongsibidang , G. S. (2016). Resiko Hipertensi, Diabetes Militus Dan Mengkonsumsi Obat Herbal pada Kejadian Gagagl Ginjal Kronik Di RSUP DR Wahidin Sudiro Husodo Makasar Tahun 2015. *Journal Wiyata.*3(2) 162 - 167.

R. Lozano et al, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, (9859), pp. 2095- 2128, 2012.

Stuart, E.A. et al., 2009. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *American journal of epidemiology*, 169(9), pp.1133–9.

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications,* 38(7), 8696–8702.