HYBRID METODE BOOSTRAP DAN TEKNIK IMPUTASI PADA METODE C4-5 UNTUK PREDIKSI PENYAKIT GINJAL KRONIS

Ahmad Ilham [1]
[1] Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Semarang
Alamat e-mail : ahmadilham@unimus.ac.id

ABSTRACT

Missing values is a serious problem that most often found in real data today. The C4.5 method is a popular classification predictive modeling used because of its ease of implementation. However, C4.5 is still weak when testing data that contains large missing. In this study we used a hybrid approach the bootstrap method and k-NN imputation to overcome missing values. The proposed method tested using Chronic Kidney Disease (CKD) data, and evaluated using accuracy and AUC. The results showed that the proposed method was superior in overcoming missing values in CKD. It can be concluded that the proposed method is able to overcome missing values for chronic kidney disease prediction.

Keywords : *Missing Values, Bootstrap, K-NN, Chronic Kidney Disease Prediction*

PENDAHULUAN

Penyakit ginjal adalah kelainan yang mempengaruhi fungsi ginjal. Selama tahap akhir, penyakit ginjal dapat menyebabkan gagal ginjal. Menurut data terbaru yang dirilis oleh Pusat Pencangkokan Organ Saudi, 10.203 pasien yang didiagnosis menderita penyakit ginjal menerima hemodialysis [1]. Diabetes, tekanan darah tinggi, dan gaya hidup yang tidak sehat telah menyebabkan peningkatan jumlah pasien dengan CKD [2]. Pasien dengan CKD menderita berbagai efek samping. Komplikasi ini termasuk kerusakan pada sistem saraf dan kekebalan tubuh yang mengganggu aktivitas sehari-hari. Untuk membantu dalam pencegahan CKD, teknik pembelajaran mesin dapat digunakan untuk mendiagnosis CKD pada tahap awal.

Dalam berbagai literatur [3] sejumlah model prediksi seperti Radial Basis Function (RBF), Learning Vector Quantization (LVQ), C4.5, Clasification and Regression Tree (CART), Bayesian Tree, Random Forest (RF), Artificial Neural Network (ANN)+Fuzzy Neural Network (FNN), Hybrid Prediction Model (HPM), Sim+F2, Real code Genetic Algorithm (GA), dan Fuzzy Min Max (FMM), FMM- CART-RF untuk mendukung diagnosis penyakit ginajal.

Decision Tree adalah algoritma yang telah banyak diterapkan diberbagai bidang, seperti bidang pengobatan [4], bidang bisnis [5] dan deteksi gagal ginjal [6]. Di bidang kesehatan contohnya penerapan Decision Tree untuk memprediksi pasien kanker payudara [7]. Algoritma C4.5 Decision Tree merupakan pengembangan dari metode Iterative Dichotomiser 3 (ID3) yang dapat bekerja pada variabel kontinyu dan missing value.

Dalam berbagai literature [8], [9] missing value sering terjadi karena adanya nilai-nilai yang hilang di atribut, kesalahan sering terjadi dalam prosedur entri data secara manual, kesalahan peralatan atau pengukuran yang salah.

Data yang hilang lebih dikenal dengan sebutan missing value dalam data mining dapat menyebabkan terjadinya hasil atau keputusan yang bias disebabkan oleh missing value pada data dan data yang lengkap [10][11].

Ada dua pendekatan dalam mengatasi missing value dalam data, yaitu, pendekatan teknik toleransi [12] dan pendekatan teknik imputasi [13]. Pendekatan teknik toleransi seperti metode klaster dan seleksi fitur [12][14][15] dapat digunakan untuk mengatasi missing value pada data. Fokus dalam penelitian ini adalah pendekatan kedua yaitu teknik imputasi adalah cara mengisi missing value dengan nilai yang adil dari metode imputasi yang digunakan sebelum menjadi data lengkap yang siap modelkan dan dianalisis. Keunggulan dari teknik imputasi adalah kegiatan imputasi data dalam menangani missing value tidak bergantung pada pemilihan metode prediksi dan klasifikasi akan tetapi dapat memilih algoritma pembelajaran yang sesuai setelah imputasi [13].

Dalam penelitian ini metode C4.5 akan diusulkan sebagai model pengklasifikasi. Missing value pada data di tahap pre-processing diatasi menggunakan ukuran pemusatan data yaitu mengganti missing value dengan nilai rata-rata (mean), nilai yang sering muncul (mode) dan imputasi metode k-NN. Tujuan penelitian ini adalah menganalisis peningkatan kinerja metode C4.5 dalam prediksi penyakit ginjal kronis.

Penelitian ini disusun sebagai berikut. Pada bagian 2, ada penjelasan tentang tinjauan pustaka. Di bagian 3, dijelaskan presentasi dari metode yang diusulkan. Hasil percobaan membandingkan metode yang diusulkan dengan yang lain diurai di bagian 4. Bagian terakhir dikhususkan untuk menyimpulkan karya artikel ini.

A. Missing Value

Missing value terjadi mungkin adanyak alat (fisik) pengukuran yang digunakan rusak, perubahan desain eksperimental selama pengumpulan data dan pengumpulan beberapa dataset yang serupa tetapi tidak identic. Pada fitur numerik, *missing value* ditandai dengan masukan *out of range* dengan nilai -1 yang seharusnya bernilai positif yaitu > 0. Sedangkan untuk atribut nominal, *missing value* ditandai dengan nilai kosong atau tanda hubung dan mungkin bilangan bulat negatif (misalnya -1, -2, dll) [16]. Bisa juga karena adanya responden dalam sebuah survey menolak menjawab beberapa pertanyaan tertentu seperti usia dan pendapatan, dsb.

Umumnya, keberadaan missing value dapat mengurangi kinerja pemodelan dalam data mining [17], tidak heran bila kasus ini menjadi masalah serius sehingga banyak diteliti oleh sebagian kalangan peneliti data mining.

Menurut [18] ada tiga masalah yang dapat terjadi pada hasil model prediksi data mining, di antaranya : 1) hilangnya efisiensi, 2) komplikasi dalam penanganan ekstraksi dan analisis data, dan 3) dapat membuat hasil pemodelan menjadi bias disebabkan adanya perbedaan antara missing value dengan data yang lengkap.

Missing value dapat diatasi melalui tiga cara yang berbeda [19]:

Cara 1. membuang contoh yang mengandung missing value dari atribut;

Cara 2. menggunakan prosedur maximum likelihood, di mana diperkirakan parameter model untuk data lengkap dan kemudian digunakan untuk imputasi menggunakan nilai means;

Cara 3. menggunakan metode imputasi missing value, bertujuan untuk mengisi missing value dengan nilai yang telah diperkirakan;

Cara 4. Dalam kebanyakan kasus, atribut dataset tidak saling bergantungan antara satu dengan yang lain. Oleh karena itu missing value dapat ditentukan dengan identifikasi hubungan antar atribut.

B. Metode Imputasi

1) Most Common Imputation

Imputasi dengan nilai Mean dan Mode atau dikenal dengan Most Common Imputation (MC) adalah metode imputasi dari metode statistic sederhana di mana missing value digantikan dengan nilai perkiraan yang masuk akal (satu perkiraan per missing value) sebelum di masukkan ke dalam data keseluruhan [19]. Metode ini memiliki kelemahan yaitu mengurangi varian pada variabel karena nilai yang diisikan adalah sama pada setiap variabel [20] (lihat Formula (1).

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (1)$$

Dimana:

$\bar{X}$ = nilai rata-rata

X_i = nilai sampel ke-i

n = jumlah sampel

2) k-NN

K-NN merupakan metode imputasi populer dalam menangani missing value [21]. Keunggulannya adalah: 1) dapat digunakan untuk memprediksi dua tipe data yaitu data diskret dan kontinyu. Imputasi data diskret menggunakan nilai modus dan pada data kontinyu menggunakan nilai mean. 2) pada setiap item yang mengalami missing value tidak diperlukan adanya pembentukan model prediksi [22]. Kelemahan dari imputasi k-NN adalah ketika melakukan pengamatan untuk mencari nilai yang paling sesuai terhadap missing value, algoritma imputasi k-NN akan melakukan pencarian di semua dataset sehingga membutuhkan waktu yang lama jika dataset-nya besar. Akan tetapi metode imputasi k-NN tetap merupakan metode yang baik dalam menangani missing value [23] (lihat Formula 2)a.

$$d(X_a, X_b) = \sqrt{\sum_{i=1}^{n} (X_{ai}, X_{bi})^2} \quad (2)$$

Dimana: $d(X_a, X_b)$ merupakan jarak antara target observasi X_a dan observasi X_b. X_{ai} adalah nilai variabel ke-i pada target observasi $X_a, i = 1, 2, \dots, n$. X_{bi} adalah nilai variabel ke-i pada target observasi lainnya $X_b, i = 1, 2, \dots, n$.

3) Metode bootstrapping

Metode bootstrap adalah teknik resampling yang digunakan untuk memperkirakan statistik pada suatu populasi dengan mengambil sampel dataset dengan penggantian [24]. Teknik penarikan sampel metode bootstrap adalah dengan pengembalian dari sebuah sampel asli. Sampel asli merupakan sampel yang diperoleh dari hasil observasi yang diperlakukan seolah-olah sebagai populasi.

Tujuan utama metode bootstrap adalah menyiapkan data sebelum pemasangan model atau penyetelan hyperparameter model harus terjadi dalam for-loop pada sampel data. Hal ini untuk menghindari kebocoran data di mana pengetahuan tentang dataset uji digunakan untuk meningkatkan model. Pada gilirannya, dapat menghasilkan estimasi optimis dari model yang dibentuk. Berikut adalah formula dasar dari boostrapping:

$$1 - (1 - \frac{1}{N})^N \quad (3)$$

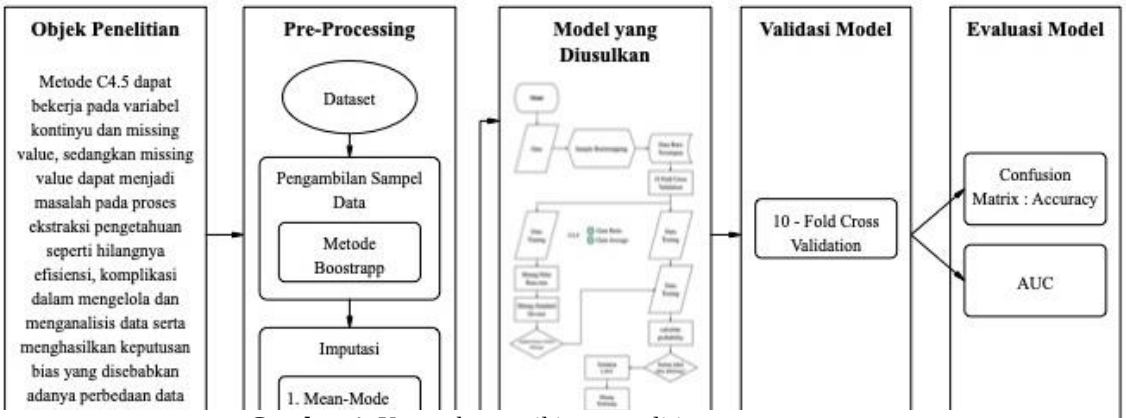
$$acc_{boot} = \frac{1}{m} \sum_{i=1}^{m} (0{,}632\ x\ e_i + 0{,}368\ c\ acc_s) \quad (4)$$

$$\hat{\theta} = \sum_{m=1}^{M} \hat{\theta}^{*m} . \frac{1}{M} \quad (5)$$

$$SE = \frac{\sigma}{\sqrt{n}} \quad (6)$$

C. Evaluasi Model

Evaluasi model yang digunakan dalam penelitian ini adalah confusion



Gambar 1. Kerangka pemikiran penelitian.

matrix. Dari confusion matrix maka dapat diketahui jumlah data dari masing-masing kelas yang diprediksi secara benar dan salah. Dengan mengetahui jumlah tersebut maka akan diketahui akurasi prediksi (lihat Formula 7).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Analisis Receiver Operating Characteristic (ROC) adalah metode standar untuk mengevaluasi kinerja model prediksi [25], juga akan digunakan. Kurva ROC dibagi menjadi 2 dimensi, yaitu tingkat *TP* diplot pada sumbu *Y* dan *FP* diplot pada sumbu *X*. Untuk menggambarkan grafis yang menentukan klasifikasi mana yang lebih baik, digunakan metode untuk menghitung luas daerah dibawah kurva ROC yang disebut AUC yang diartikan sebagai probabilitas. AUC mengukur kinerja diskriminatif dengan memperkirakan probabilitas output dari sampel yang dipilih secara acak dari populasi positif atau negatif, semakin besar AUC, semakin kuat klasifikasi yang digunakan. Karena AUC adalah bagian dari daerah unit persegi, nilainya akan selalu antara 0.0 dan 1.0. Formula perhitungan nilai AUC adalah sebagai berikut:

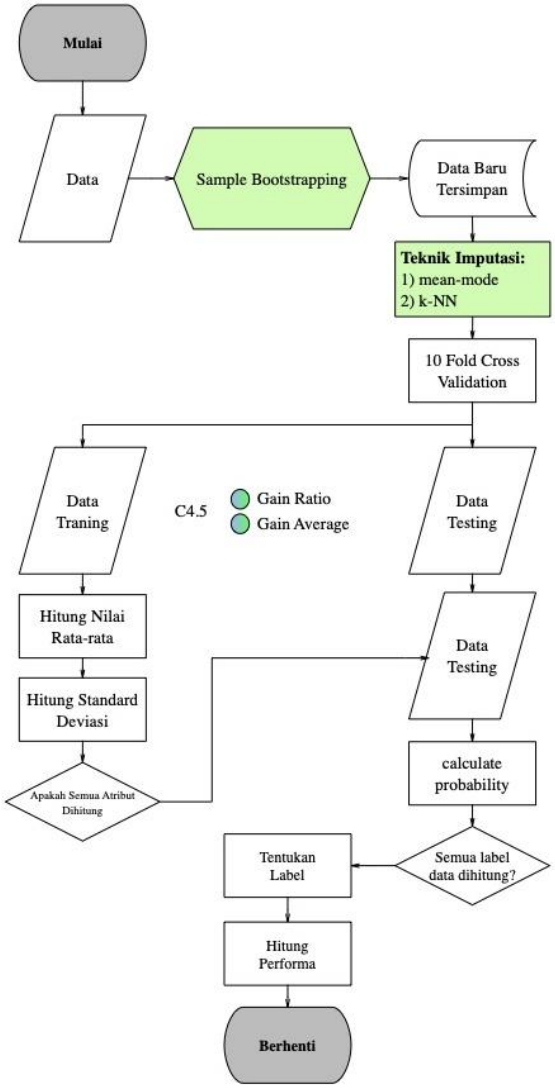
$$AUC = \frac{1+TP_{rate}-FP_{rate}}{2} \quad (8)$$

Gambar 1 menunjukkan kerangka pemikiran penelitian.

METODELOGI PENELITIAN

Gambar 2, menunjukkan flowchart model yang diusulkan. Tujuan Sample Bootstrapping adalah untuk mengambil sampel data dari data asli dengan Teknik penggantian nilai data (*sampling with replacement*). mengatasi missing value dalam data data yang digunakan. Metode imputasi bootstrapping digunakan untuk mengatasi missing values dan C4.5 sebagai model prediksi.

Dalam uji model kami menggunakan validasi menggunakan 10-fold cross validation. Performa kinerja dari metode C4.5 terhadap dataset yang mengandung missing value sebelum dan sesudah dilakukan preprocessing data menjadi obyek pada penelitian ini. Peningkatan kinerja dari model prediksi terhadap hasil yang diperoleh akan diukur menggunakan confusion matrix. Hasil yang dicatat dan dibandingkan melalui confusion matrix dan AUC.



Gambar 2. Flowchart model yang diusulkan.

HASIL DAN PEMBAHASAN

Platform yang digunakan dalam percobaan ini adalah Intel Core i5 Dual-Core 2,5 GHz, 8 GB RAM, dan macOS Cataline 64-bit sebagai sitem operasi. Lingkungan pengembangan adalah Netbeans 7 dengan bahasa pemrograman Java. Perangkat lunak aplikasi adalah Knime 4.1.0.

Dalam penelitian ini, kami menggunakan dataset Chronic Kidney Disease/CKD yang dapat diakses di http://archive.ics.uci.edu/ml/. Deskripsi dataset dapat ditemukan pada Tabel 1, dan Tabel 2.

Tabel 1. Deskripsi data CKD.

Dataset	Chronic Kidney Disease	
Jumlah record	400	
Jumlah fitur	25	
	Fitur nominal	14
	Fitur numerik	11
Missing values	1012	
Jumlah label	2	

Tabel 2. Distribusi label pada data CKD.

Label	Record	Keterangan
ckd	250	Jumlah sampel yang dinyatakan sebagai penderita penyakit ginjal kronis
no-ckd	150	Jumlah sampel yang dinyatakan tidak menderita penyakit ginjal kronis.
Total	400	

Seperti yang ditunjukkan pada Tabel 1, dan Tabel 2, dataset SKD mengandung 400 record data, fitur data berjumlah 25 di mana terdiri dari 14 fitur nominal dan 11 fitur numerik. Total missing value adalah 1.012, artinya jumlah missing value sangat besar dan dapat berpengaruh terhadap hasil model prediksi sehingga perlu penangan serius. Data SKD memiliki dua label data yaitu ckd (250 record) dan no-ckd (150 record).

Kami menggunakan crossvalidation bertingkat 10 kali lipat yang canggih untuk pembelajaran dan pengujian data. Ini berarti bahwa kami membagi data pelatihan menjadi 10 bagian yang sama dan kemudian melakukan proses pembelajaran 10 kali. Kami menggunakan validasi silang bertingkat 10 kali lipat, karena metode ini telah menjadi metode standar dalam istilah praktis. Beberapa tes juga menunjukkan

bahwa penggunaan stratifikasi sedikit mening-katkan hasil.

Sebagai indikator akurasi dalam mengevaluasi kinerja pengklasifikasi, dalam percobaan ini, kami menerapkan area under curve (AUC). Lessmann et al. [26] menganjurkan penggunaan AUC untuk meningkatkan komparabilitas studi silang. AUC memiliki potensi untuk secara signifikan meningkatkan konvergensi lintas eksperimen empiris dalam prediksi cacat perangkat lunak, karena memisahkan kinerja prediksi dari kondisi operasi, dan mewakili ukuran umum prediksi.

Pertama-tama, kami melakukan percobaan pada dataset CKD tanpa metode bootstrapping dan metode C4.5 sebagai pengklasifikasi, hasilnya ditunjukkan pada Tabel 3.

Tabel 3. Nilai standard error pada dataset SKD sebelum dan setelah resampling menggunakan metode bootstrap.

Fitur	Standard error	
	Sebelum Bootstrap	Setelah Bootstrap
Age	0.858486	**0.829621**
Blood Pressure	0.684182	**0.686361**
Specific Gravity	0.000286	**0.000280**
Albumin	0.067634	**0.065592**
Sugar	0.054960	**0.05496**

Tabel 3, menunjukkan perbandingan nilai standard error antara data asli dan data resampling dari metode bootstrap. Nilai tebal pada tabel diatas adalah metode resampling yang diusulkan melaporkan rata-rata standard error lebih kecil dari data aslinya. Hal ini berarti, uji hipotesis data sebelum dan setelah resampling menggunakan metode bootstrap lebih baik. Hasil ini telah dikonfirmasi dengan baik oleh [27] bahwa bootstrapping hasilnya lebih stabil dalam sehingga jika dinilai dari signifikansi menunjukkan sifat yang konsisten walaupun nilai *t* berbeda-beda.

Setelah dataset baru didapatkan, kemudian diamati dan dihitung berapakah jumlah missing value pada setiap fitur. Data yang masih mengandung missing selanjutnya dimputasi untuk mengganti nilai yang missing. Tahapan pertama adalah Imputasi dengan mean dilakukan untuk mencari nilai rata-rata dari masing-masing atribut. Nilai rata-rata tersebut digunakan untuk mengganti missing value pada data numerik, sedangkan imputasi dengan mode digunakan untuk mencari nilai yang sering muncul untuk mengganti missing value pada data nominal. Hasilnya dilaporkan pada Tabel 4.

Tabel 4. Nilai mean dan mode pada masing-masing atribut dataset CKD.

No	Atribut	Nilai *mean* dan *mode*	No	Atribut	Nilai *mean* dan *mode*
1	*Age*	52.39	13	*Sodium*	137.39
2	*Blood Pressure*	75.95	14	*Potassium*	4.91
3	*Specific Gravity*	1.02	15	*Hemoglobin*	12.73
4	*Albumin*	0.92	16	*Packet Cell Volume*	39.58
5	*Sugar*	0.4	17	*White Blood Cell Count*	8260.27
6	*Red Blood Cells*	normal	18	*Red Blood Cell Count*	4.80
7	*Pus Cell*	normal	19	*Hypertension*	no
8	*Puss Cell Clumps*	notpresent	20	*Diabetes Mellitus*	no
9	*Bacteria*	notpresent	21	*Coronary Artery Disease*	no
10	*Blood Glucose Random*	147.90	22	*Appetite*	good

Langkah berikutnya, imputasi menggunakan k-NN dengan nilai k = 3 dilakukan untuk mengganti missing value. sehingga data menjadi lengkap tidak ada nilai yang kosong. Hasil sebelum dan setelah imputasi k-NN ditunjukkan pada Tabel 5.

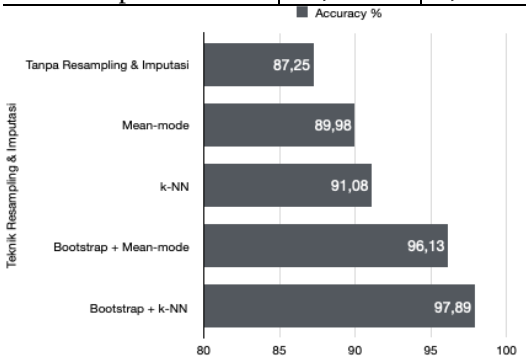
Tabel 5. Perbandingan nilai sebelum dan setelah imputasi berdasarkan jarak observasi k-NN.

Age	Blode Presure (data asli)	Blode Presure (setelah imputasi)	Distance (dari data asli)	Distance (Setelah di imputasi)
48	70	70	24,91987	24,91987
24	?	**80**	-	**24,91987**
52	100	100	37,74917	37,74917
62	80	80	38,34058	38,34058

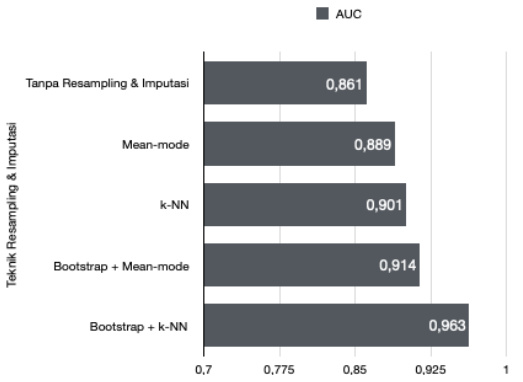
Komparasi dan perbandingan metode-metode imputasi dalam mengatasi missing data pada pemodelan C4.5 ditunjukkan pada Tabel 6.

Tabel 6. Perbandingan evaluasi sebelum dan setelah menggunakan teknik resampling dan imputasi pada C4.5 untuk prediksi penyakit gijal kronis.

Metode Imputasi	Accuracy	AUC
Tanpa resampling dan imputasi	87,25%	0,861
Mean-mode	89,98%	0,889
k-NN	91,08%	0,901
Bootstrap+Mean-mode	96,13%	0,914
Bootstrap+k-NN	97,89%	0,963



Gambar 3. Diagram perbandingan hasil accuracy sebelum dan setelah menggunakan teknik resampling dan imputasi pada C4.5 untuk prediksi penyakit ginjal kronis.



Gambar 4. Diagram perbandingan Accuracy sebelum dan setelah menggunakan teknik resampling dan imputasi pada C4.5 untuk prediksi penyakit ginjal kronis

Berdasarkan evaluasi menggunakan Accuracy pada Tabel 6 dan Gambar 3, kinerja C4.5 sebelum dan setelah menggunakan hibrid metode bootstrap sebagai teknik resampling dan k-NN sebagai teknik imputasi mengalami peningkatan yang signifikan dari 91,08% ke 97,89% unggul dibandingkan teknik lainnya. Teknik unggul kedua disusul oleh bootstrap + mean-mode lalu terakhir tanpa resampling masing-masing mendapatkan akurasi 89,98% ke 96,13% dan 87,25%. Temuan ini menyimpulkan bahwa penggabungan teknik resampling dan teknik imputasi pada pemodelan C4.5 mengalami peningkatan yang signifikan dibandingkan pemodelan tanpa dua teknik tersebut dalam mengatasi missing value untuk prediksi penyakit ginjal kronis.

Begitupun dengan evaluasi menggunakan AUC seperti yang dapat dilihat pada Tabel 6, dan Gambar 4 ditemukan peningkatan yang mirip. Kinerja C4.5 sebelum dan setelah menggunakan hibrid metode bootstrap sebagai teknik resampling dan k-NN sebagai teknik imputasi mengalami peningkatan yang signifikan dari 0,901 ke 0,963 unggul dibandingkan teknik lainnya. Teknik unggul kedua disusul oleh bootstrap + mean-mode lalu terakhir tanpa resampling masing-masing mendapatkan akurasi 0,889 ke 0,914 dan 0,861.

KESIMPULAN

Dalam penelitian ini, kami telah mempresentasikan metode baru yang kami usulkan yang disebut B-kNN-C4.5 untuk mengatasi missing values untuk prediksi penyakit ginjal kronis. Temuan utama kami adalah bahwa bootstrap menjadi alternative yang baik dalam menyiapkan data sebelum di modelkan, dan k-NN dengan teknik imputasi yang handal mampu menyediakan data tanpa missing values lagi, akhirnya terbentuk sebuha data baru yang siap dimodelkan menggunakan metode C4.5. Hasil penelitian ini menunjukkan metode yang diusulkan lebih unggul dibandingkan

metode pembanding lainnya. Dapat disimpulkan bahwa metode yang diusulkan dapat diterakan untuk mengatasi data yang missing untuk prediksi penyakit ginjal kronis.

DAFTAR PUSTAKA

[1] A. A. Al-Sayyari and F. A. Shaheen,2011,End stage chronic kidney disease in Saudi Arabia. A rapidly changing scene., *Saudi Med. J.*, vol. 32, no. 4, pp. 339–46, http://www.ncbi.nlm.nih.gov/pubmed/21483990.

[2] Z.Chen,X.Zhang,Z. Zhang, 2016, Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models, *Int. Urol. Nephrol.*, vol. 48, no.12,pp.2069–2075, doi:10.1007/s11255-016-1346-4.

[3] M. Seera, C. P. Lim, 2014,A hybrid intelligent system for medical data classification,*Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, doi: 10.1016/j.eswa.2013.09.022.

[4] D.Setsirichok *et al.*, 2012, Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, *Biomed. Signal Process. Control*, doi: 10.1016/j.bspc.2011.03.007.

[5] P. Duchessi and E. J. M. Lauría, 2013, Decision tree models for profiling ski resorts' promotional and advertising strategies and the impact on sales, *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5822–5829, doi: 10.1016/j.eswa.2013.05.017.

[6] Y. Sahin, S. Bulkan, and E. Duman, 23013, A cost-sensitive decision tree approach for fraud detection, *Expert Syst.Appl.*,doi:10.1016/j.eswa.2013.05.021.

[7] M. Ture, F. Tokatli, and I. Kurt, 2009, Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Expert Syst. Appl.*, doi:10.1016/j.eswa.2007.12.002.

[8] L. Guelman, 2012, Gradient boosting trees for auto insurance loss cost modeling and prediction, *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3659–3667, doi:10.1016/j.eswa.2011.09.058.

[9] E. M. Mirkes, T. J. Coats, J. Levesley, and A. N. Gorban, 2016, Handling missing data in large healthcare dataset : A case study of unknown trauma outcomes," *Comput. Biol. Med.*, vol. 75, pp. 203–216, doi:10.1016/j.compbiomed. 2016.06.004

[10] Y. Hayashi and S. Yukita, 2016, Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset, *Informatics Med. Unlocked*, vol. 2, pp.92–104, doi:10.1016/j.imu.2016.02.001.

[11] J. A. Sáez, J. Derrac, J. Luengo, and F. Herrera, 2014, Statistical computation of feature weighting schemes through data estimation for nearest neighbor classifiers, *Pattern Recognit.*, vol. 47, no. 12, pp. 3941–3948, doi:10.1016/j.patcog.2014.06.012.

[12] D. Williams, X. Liao, Y. Xue, L. Carin, and B. Krishnapuram,2007, On Classification with Incomplete Data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 427–436, doi:10.1109/TPAMI.2007.52.

[13] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, 2007, Semi-parametric optimization for missing data imputation,*Appl. Intell.*, vol. 27, no. 1, pp. 79–88, doi:10.1007/s10489-006-0032-0.

[14] F. O. De França, G. P. Coelho, and F. J. Von Zuben,2013, Predicting missing values with biclustering: A coherence-based approach, *Pattern Recognit.*, vol. 46, no. 5, pp. 1255–1266, doi:10.1016/j.patcog.2012.10.022.

[15] A. Aussem and S. Rodrigues de Morais, 2010, A conservative feature subset selection algorithm with missing data, *Neurocomputing*, doi: 10.1016/j.neucom.2009.05.019.

[16] Y. Ding and J. S. Simonoff, 2010, An investigation of missing data methods for classification trees applied to binary response data, *J. Mach. Learn. Res.*

[17] J. Han, M. Kamber, and J. Pei, 2012, *Data Mining Concepts and Techniques*. Elsevier.

[18] I. B. Aydilek and A. Arslan, 2013, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Inf. Sci. (Ny).*, vol. 233, pp. 25–35, doi: 10.1016/j.ins.2013.01.021.

[19] R. J. Little and D. B. Rubin, 2015, Missing Data, in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, pp. 602–607.

[20] J. L. Schafer and J. W. Graham, 2002, Missing data: Our view of the state of the art, *Psychol. Methods*, doi: 10.1037/1082-989X.7.2.147.

[21] T. Astuti, H. A. Nugroho, and T. B. Adji, 2015, The impact of different fold for cross validation of missing values imputation method on hepatitis dataset, in *2015 International Conference on Quality in Research (QiR)*, pp. 51–55, doi: 10.1109/QiR.2015.7374894.

[22] B. Twala, 2009, AN EMPIRICAL COMPARISON OF TECHNIQUES FOR HANDLING INCOMPLETE DATA USING DECISION TREES, *Appl. Artif. Intell.*, vol. 23, no. 5, pp. 373–405, doi:10.1080/08839510902872223.

[23] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, 2010, Pattern classification with missing data: a review,*Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, doi: 10.1007/s00521-009-0295-6.

[24] R. W. Johnson,An Introduction to the Bootstrap, Jun. 2001,*Teach. Stat.*, vol. 23, no. 2, pp. 49–54, doi: 10.1111/1467-9639.00050.

[25] A. P. Bradley, Jul. 1997, The use of the area under the ROC curve in the evaluation of machine learning algorithms,*Pattern Recognit*., vol. 30, no. 7, pp. 1145–1159,doi: 10.1016/S0031-3203(96)00142-2.

[26] S. Lessmann, B. Baesens, C. Mues, S. Pietsch,Jul. 2008, Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings,*IEEE Trans. Softw. Eng.*, vol. 34, no. 4, pp. 485–496.

[27] W. Hung, M. Yang, and D. Chen, 2008, Bootstrapping approach to feature-weight selection in fuzzy c - means algorithms with an application in color image segmentation, *Pattern Recognit. Lett.*, vol. 29, pp. 1317–1325, doi: 10.1016/j.patrec.2008.02.003.