Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202294778
Vol. 9, No. 4, Agustus 2022, hlm. 779-786 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENGARUH PREDIKSI MISSING VALUE PADA KLASIFIKASI DECISION TREE C4.5

Aji Seto Arifianto[*1], Kursita Dewi Safitri[2] , Khafidurrohman Agustianto[3], I Gede Wiryawan[4]

[1,2,3,4]Politeknik Negeri Jember, Jember
Email: [1]ajiseto@polije.ac.id, [2]kursitadewi36@gmail.com, [3]khafid@polije.ac.id, [4]wiryawan@polije.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 23 Februari 2021, diterima untuk diterbitkan: 19 Agustus 2022)

Abstrak

Pendekatan klasifikasi data bersifat supervised learning menuntut adanya dataset yang lengkap. Permasalahan yang muncul adanya missing value yaitu hilangnya nilai suatu atribut yang diakibatkan kesalahan dalam pengumpulan data, kesalahan saat memasukkan data, dan ketidakmampuan responden untuk memberikan jawaban yang akurat. Penelitian ini melakukan uji coba pengembangan rule decision tree C4.5 untuk data penyakit ginjal kronis. Dataset terdiri dari 400 record, 24 atribut dan 1 kelas target. Karakteristik data yang digunakan meliputi 11 data bertipe numerik dan 14 data bertipe nominal. Jumlah data kelas positif penyakit ginjal kronis 250, sedangkan negatif ginjal kronis 150. Total data yang tidak lengkap (missing value) 1012 records. Perlakuan pertama dibangun rule dengan menghitung entropy dan gain pada 360 data training yang terdapat missing value diperoleh 21 rules. Kemudian pada perlakuan kedua diterapkan prediksi missing value menggunakan rumus mean dan modus sebelum pembentukan rule tree, didapatkan 24 rules. Peneliti melakukan pengujian akurasi tree C.45 dengan 40 data uji, hasilnya 90% untuk rule dengan missing value dan 95% untuk dataset yang telah diprediksi nilainya.

Kata kunci: *decision tree C4.5; missing value; classification, rule*

THE EFFECT OF MISSING VALUE PREDICTION ON DECISION TREE C4.5 CLASSIFICATION

Abstract

The supervised learning approach to data classification requires a complete dataset. The problem that arises was the existence of missing value, namely the loss of the value of an attribute due to errors in data collection, errors when entering data, and the inability of respondents to provide accurate answers. This study conducted a trial on the development of the C4.5 rule decision tree for chronic kidney disease data. The dataset consisted of 400 records, 24 attributes and 1 target class. The data characteristics included 11 numeric data and 14 nominal data types. The number of positive data for kidney disease was 250, while the number of negative for kidney disease was 150 and the total of missing value was 1012 records. The first treatment was building a rule by calculating the entropy and gain on 360 training data where missing value was obtained, it was 21 rules. Then in the second treatment, the prediction of missing value was applied using the mean and mode formula before the formation of the rule tree, obtained 24 rules. Researcher was measuring the accuracy of the two rules tree C4.5 is done by using 40 data-testing, the result is 90% for rules with missing value and 95% for datasets whose value has been predicted.

Keywords: *decision tree C4.5; missing value; classification, rule*

1. PENDAHULUAN

Decision Tree merupakan salah satu metode klasifikasi yang paling sering digunakan dalam penelitian seperti pada bidang kedokteran, astronomi dan biologi molekuler (Sari and Mahmudy, 2019). Data mining adalah proses menemukan informasi penting dari database yang besar. Salah satu teknik dalam data mining adalah klasifikasi, metode yang digunakan adalah *decision tree*, algoritma yang digunakan adalah algoritma C4.5 (Wahyuni, 2018). *Decision tree* merupakan representasi sederhana dari teknik klasifikasi. Algoritma *decision tree* memiliki beberapa jenis diantaranya adalah ID3, C4.5 dan J48. Metode *decision tree* berguna untuk mengeksplorasi data, serta menemukan hubungan tersembunyi antara sejumlah variabel input potensial dengan variabel target. Keunggulan *decision tree* C4.5 mampu menangani *countinuous attribute*, *missing data*, membangkitkan *rule* dari sebuah pohon (*tree*), dan

pemangkasan pada saat pembangunan *tree* atau pada saat proses pembangunan *tree* selesai (Wulandari, 2010). Di samping itu, hasil dari algoritma *decision tree* C4.5 ini mudah dipahami, membutuhkan lebih sedikit data eksperimen dibanding algoritma klasifikasi lainnya, mampu mengolah data nominal dan kontinyu, waktu komputasi relatif cepat (Rahim *et al.*, 2018). Algoritma *decision tree* C4.5 sangat baik digunakan untuk analisis rekomendasi calon debitur dengan akurasi 97,96% (Nurellisa and Fitrianah, 2020). Klasifikasi *decision tree* bersifat *supervised learning* mutlak membutuhkan adanya dataset untuk prosesnya. Dataset yang dimaksud dibagi menjadi data latih dan uji. Data latih digunakan untuk membangun model klasifikasi dan data uji digunakan untuk validasi hasil klasifikasi. Pengumpulan data latih menimbulkan berbagai masalah seperti jumlah data yang minim serta adanya data kosong (*missing value*). *Missing value* berarti hilangnya nilai dari suatu atribut tertentu, hal ini dapat disebabkan kesalahan saat pengumpulan data, kesalahan pada saat entri data dan ketidakmampuan responden dalam memberikan jawaban yang akurat (Mukarromah, Martha and Ilhamsyah, 2015). *Missing value* menyebabkan tingkat akurasi data berkurang dan cenderung bias oleh karenanya harus di solusikan dengan nilai prediksi pada kelompok data (Rizaldi, Purnomo and Arifianto, 2019).

Penyakit ginjal kronis atau *chronic kidney disease* (CKD) merupakan penurunan fungsi ginjal secara bertahap karena kerusakan ginjal. Secara medis didefinisikan terjadinya penurunan *glomerular filtration rate* (GFR) kurang dari 60 mL/menit per 1,73 m2 selama 3 bulan atau lebih. Diabetes dan hipertensi adalah penyebab utama CKD di berbagai negara baik maju maupun berkembang (Romagnani *et al.*, 2017). Mayoritas pasien dengan CKD berada pada risiko kematian yang tinggi. Keterbatasan akses untuk transplantasi ginjal merupakan masalah di banyak bagian dunia. Faktor yang menjadi penyebab CKD diantaranya kehilangan nefron karena bertambahnya usia, cedera ginjal akut karena paparan racun atau penyakit seperti obesitas dan *diabetes mellitus* (Tjekyan, 2014) (Romagnani *et al.*, 2017) . Terdapat hasil yang menunjukan adanya hubungan antara tekanan darah tinggi, merokok dan konsumsi minuman penambah tenaga dengan penyakit ginjal. Kondisi ini memicu munculnya penelitian terkait ginjal dari multi perspektif (Delima *et al.*, 2014).

Berdasarkan faktor-faktor resiko yang mempengaruhi penyakit ginjal kronik, memungkinkan untuk dilakukan pendeteksian penyakit tersebut dengan model klasifikasi. Namun banyaknya faktor resiko (parameter penelitian) juga meningkatkan potensi terjadinya data yang tidak lengkap atau hilang, hal ini menjadi tantangan tersendiri. Penelitian-penelitian yang menangani permasalahan *missing value* diantaranya penelitian yang implementasi metode penanganan data hilang untuk evaluasi data sejarah perawatan sistem/komponen RSG GAS (Hartini, 2017), penelitian lainnya melakukan pengujian tentang metode *missing data*, dimana penelitian ini membandingan *K-Means* dan *Means*, dengan hasil untuk tingkat peningkatan akurasi *K-Means* lebih unggul namun *Mean* membutuhkan waktu yang lebih cepat walaupun selisih MSE-nya 0,04 (Mukarromah, Martha and Ilhamsyah, 2015). Contoh di bidang pertanian untuk klasifikasi kualitas produksi jagung dengan algoritma C4.5. Hasil akurasi dari penanganan data *missing value* meningkat 2,40% (Moch. Lutfi and Mochamad Hasyim, 2019).

Penelitian dengan topik menangani *missing value* dalam bidang kesehatan. Teknik imputasi *missing values* pada data hepatitis, ditujukan untuk memilih metode yang cocok dalam melakukan imputasi. Tingkat akurasi metode *mean* dan *modus* sebesar 84,615%, metode *K-Nearest Neighbor Imputation* sebesar 82,051% dan metode *Singular Value Decomposition Imputation* sebesar 79,487% (Apriliawan, 2015). Pada penelitian (Arifin and Ariesta, 2019) didapatkan hasil akurasi *confusion matrix* 97% dan AUC 99.8% setelah dilakukan penambahan *Particle Swarm Optimization* hasil yang diperoleh menjadi 98,75% dan 99%. Penelitian ini mengimplementasikan *Mean* untuk menyelesaikan *missing value*¸ pemilihan metode ini dengan tujuan menghasilkan *preprocessing* yang lebih cepat untuk pembacaan data penyakit ginjal. Penelitian melakukan pengembangan *rule decision tree* C4.5 pada dataset yang tidak lengkap dan dataset yang telah diprediksi *missing value*. Dengan dilakukan penelitian ini, maka akan diketahui perbedaan *rule* yang dihasilkan dari data tidak lengkap dan data yang telah diprediksi dengan metode *missing value* berdasarkan nilai akurasi.

2. METODE PENELITIAN

Tahapan penelitian yang diawali dengan studi literatur, kemudian dilanjukan pengumpulan dan pengolahan data, lalu pengembangan perangkat lunak, terakhir analisis hasil penelitian.

2.1. Studi Literatur

Studi literatur adalah tahap awal dalam melakukan penelitian dengan teknik mengumpulkan informasi yang dibutuhkan. Proses ini dilakukan untuk mencapai tujuan penelitian. Literatur yang berkaitan dengan topik penelitian dijadikan sebagai dasar atau sumber rujukan seperti penanganan *missing value* dan gambaran klinis yang terjadi pada pasien penderita penyakit ginjal kronis.

2.2. Pengumpulan Data

Tahap pengumpulan data diambil dari laman UCI Repository: Dataset *Chronic Kidney Disease*. Data tersebut merupakan data sekunder karena data diperoleh berdasarkan hasil dari penelitian orang lain. Pada penelitian (Arifin and Ariesta, 2019). Dataset CKD mempunyai 24 atribut parameter dan 1 atribut

target dengan jumlah sebanyak 400 data. Atribut tersebut mempunyai tipe data yang berbeda yaitu terdapat 11 data bertipe numerik dan 14 data bertipe nominal. Jumlah seluruh data dibagi menjadi 2 kelas yaitu CD (positif penyakit ginjal kronis) sebanyak 250 data dan NCD (negatif penyakit ginjal kronis) sebanyak 150 data. Pada data CKD terdapat data yang hilang atau kosong sebanyak 1.012 data. Dalam penelitian ini *dataset* dibagi menjadi 2 bagian yaitu 90% data latih dan 10% data uji. Data *missing value* yang terdapat dalam data latih sebanyak 599 data. Atribut yang digunakan dalam mengklasifikasikan penyakit ginjal kronis adalah 19 atribut parameter berdasarkan jurnal yang telah dikaji pada tahap sebelumnya.

2.3. Pengolahan Data

Pada tahap ini merupakan proses mengubah data agar dapat diolah sesuai dengan metode yang telah ditentukan. Pengolahan awal data bisa disebut tahap *preprocessing*. Dalam tahap *preprocessing* data terdapat 2 tahapan yaitu :

1. *Cleaning data* mengganti nilai data yang tidak lengkap dengan nilai *mean*.

$$\bar{x} = \frac{\sum_{i=1}^{N} Xi}{N} = \frac{x_1+x_2+\cdots+x_n}{N} \quad (1)$$

Pada persamaan (1) di atas, $\bar{x}$ adalah rata-rata atau *mean*, sedangkan $\sum_{i=1}^{N} x_i$ adalah nilai atribut i, dan n adalah jumlah dari seluruh frekuensi

2. Transformasi data dilakukan dengan *entropy-based discretization*. Metode ini menggunakan *entropy*. Nilai pemisah menggunakan *split point entropy* dan nilai *gain* (2 sampel), ditambah dengan menerapkan teknik diskritisasi atau penyederhanaan.

Nilai *entropy* dapat dihitung dari target yang ingin dicari. Persamaan (2) di bawah ini digunakan menghitung *entropy* untuk satu *attribute*.

$$\text{Entropy(S)} = \sum_{i=1}^{k} -pi * \log_2 pi \quad (2)$$

Dimana S {} (*dataset*) kasus, lalu k n-partisi S, dan Pi adalah probabilitas (ya+tidak/total kasus). Setelah didapatkan nilai *entropy*, nilai gain dapat dihitung dengan menggunakan persamaan (3) berikut ini.

$$\text{Gain(A)} = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(Si) \quad (3)$$

Pada persamaan di atas, A (atribut), n (∑ bagian dari A), $|S_i|$ (∑ kasus pada bagian ke-i), dan |S| (∑ kasus dalam S)

2.4. Pengembangan Perangkat Lunak

Pendekatan yang digunakan untuk mengembangkan perangkat lunak dalam penelitian ini menggunakan Model *Waterfall* milik Ian Sommerville dengan tahapan *Requirement Definition*, *System and Software Design*, *Implementation and Unit Testing* dan *Integration and System Testing* (Muriyatmoko, Harmini and Arrahmantoro, 2021). Tahapan yang digunakan sebagai berikut:

1. *Requirement Definition*
   Tahapan ini menganalisis kebutuhan sistem berdasarkan data yang telah diperoleh.

2. *System and Software Design*
   Perancangan sistem ini dilakukan agar dapat diimplementasikan kedalam program pada tahapan selanjutnya.

3. *Implementation and Unit Testing*
   Tahap implementasi dilakukan dengan penulisan kode program bahasa pemrograman yang telah ditentukan.

4. *Integration and System Testing*
   Tahap pengujian dilakukan dengan pengukuran akurasi (*accuracy*). Pada saat pengujian, pengukuran akurasi dapat ditulis dalam Persamaan 4. (Nugraha *et al.*, 2016):

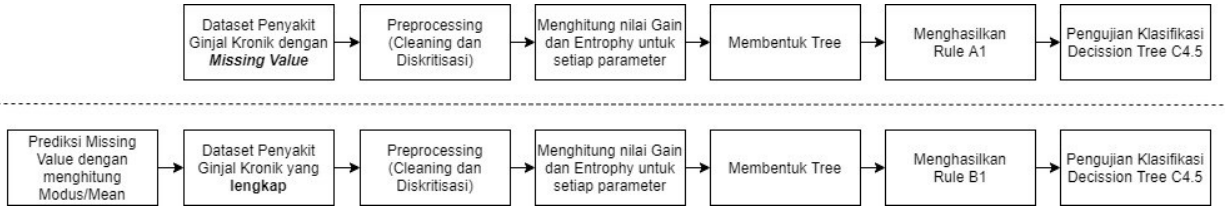
$$Accuracy = \frac{jumlah\_prediksi\_yg\_benar}{jumlah\_keseluruhan} \times 100\% \quad (4)$$

5. *Operation and Maintenance*
   Ditujukan untuk melakukan penyesuaian agar aplikasi yang dikembangkan adaptif terhadap kebutuhan implementasi di lapangan.

2.5. Analisis Hasil Penelitian

Analisis digunakan untuk mengambil kesimpulan hasil dari penelitian dengan memperhatikan tingkat akurasi dan error yang dihasilkan pada proses percobaan.

Berdasarkan tahapan-tahapan penelitian sebelumnya metode penelitian ini, ditunjukkan pada Gambar 1. berikut ini:



Gambar 1. Metode Penelitian

3. HASIL DAN PEMBAHASAN

3.1. Perhitungan Data Tidak Lengkap

Pengembangan *tree* menggunakan data latih tidak lengkap, mengandung 1.012 data kosong, dimulai dengan menghitung *entropy* untuk *node* awal terhadap kelas menggunakan persamaan (2), terdapat total kasus sebanyak 360, jumlah kasus positif penyakit ginjal kronis 225 dan jumlah kasus negatif penyakit ginjal kronis 135. Hasil perhitungan *entropy* untuk *node* awal adalah sebagai berikut.

$$Entropy(S) = -\left(\frac{225}{360}\right) log_2\left(\frac{225}{360}\right) - \left(\frac{135}{360}\right) log_2\left(\frac{135}{360}\right) = 0{,}9544$$

Tabel 1. Perhitungan Entropy

Total kasus	CD (positif)	CND (negatif)	Entropy
360	225	135	0,9544

Tabel 1 di atas menunjukan hasil perhitungan *entropy* sebelumnya. Tahap lanjutan adalah melakukan analisis atribut, nilai dan *entropy*. Salah satu contoh dalam penelitian ini menghitung *entropy* dari atribut *age*. Pada atribut *age* dikategorikan menjadi 3 yaitu ≤44 jumlah kasus positif penyakit ginjal kronis 42 dan negatif 63, >44 jumlah kasus positif ginjal kronis 175 dan negatif 71, serta NA yang diartikan sebagai data kosong memiliki jumlah kasus positif ginjal kronis 8 dan negatif 1. Berikut ini adalah contoh perhitungan *entropy* dari atribut *age*.

$$Entropy(\leq 44) = -\left(\frac{42}{105}\right) log_2 - \left(\frac{63}{105}\right) log_2\left(\frac{63}{105}\right) = 0{,}9710$$

$$Entropy\ (> 44) = -\left(\frac{175}{246}\right) log_2 - \left(\frac{71}{246}\right) log_2\left(\frac{71}{246}\right) = 0{,}8669$$

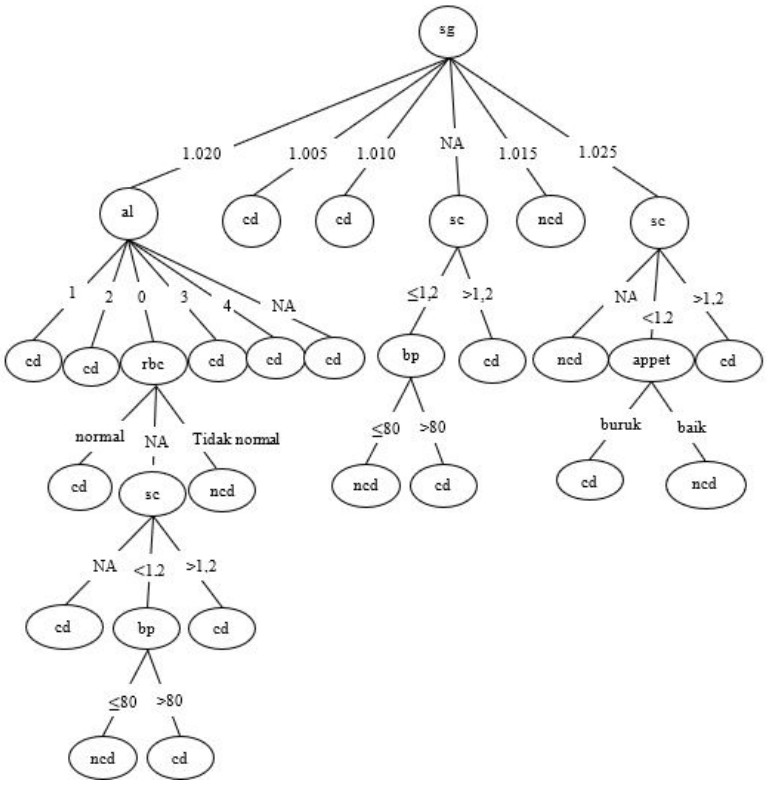
$$Entropy(\text{NA}) = -\left(\frac{8}{9}\right)\ log_2\left(\frac{8}{9}\right) - \left(\frac{1}{9}\right) log_2\ \left(\frac{1}{9}\right) = 0{,}5033$$

Setelah atribut dihitung dan didapat nilai *entropy*, maka langkah selanjutnya menghitung *gain* menggunakan persamaan (3). Salah satu contohnya atribut *age* memiliki nilai *entropy* pada ≤44 sebesar 0,9710 dan jumlah kasus 105, >44 sebesar 0,8669 dan jumlah kasus 246, serta NA sebesar 0,5033 dengan jumlah data 9. Berikut hasil nilai *gain* dari 19 atribut (Tabel 2) yang digunakan untuk mengklasifikasikan penyakit ginjal kronis pada data tidak lengkap, beserta contoh perhitungannya.

$$Gain(age) = 0{,}9544 - \left(\frac{105}{360}\right) * (0{,}9710) - \left(\frac{246}{360}\right) * (0{,}8669) - \left(\frac{9}{360}\right) * (0{,}5033) = 0{,}0663$$

Tabel 2. Nilai Gain Data Tidak Lengkap

Atribut	Nilai Gain	Atribut	Nilai Gain
Age	0,0663	Bgr	0,2736
Blood pressure	0,1642	Bu	0,2831
Specific gravity	**0,5340**	Sc	0,5203
Albumin	0,4572	Sod	0,2330
Sugar	0,1615	Pot	0,1945
Red blood cells	0,3848	Htn	0,3304
Pus cell	0,2056	Dm	0,3099
Pus cell clumps	0,0888	Appet	0,1674
Bacteria	0,0501	Pe	0,1473



Gambar 2. Tree untuk Data yang Tidak Lengkap

Berdasarkan Tabel 2 di atas, nilai *gain* tertinggi adalah atribut *specific gravity* (sg) senilai 0.5340, maka atribut sg menjadi *node* awal pada pohon keputusan data tidak lengkap. Lakukan proses perhitungan *entropy* dan *gain* pada setiap atribut secara berulang-ulang sehingga terbentuk pohon keputusan, seperti yang ditunjukan pada Gambar 2. Setelah pohon keputusan terbentuk akan menghasilkan sebuah pola atau informasi berupa *rule*. Aturan yang dihasilkan oleh data tidak lengkap sebanyak 21 (disebut rule **A1**).

3.2. Perhitungan Data yang Telah Diprediksi

Untuk pengembangan *tree* berikutnya dengan menggunakan data yang telah diprediksi, hal pertama yang dilakukan pengolahan data, yaitu *cleaning data* ditunjukkan pada Tabel 3. Data tidak lengkap yang bertipe numerik diganti dengan nilai rata-rata (*mean*), sedangkan data yang bertipe nominal diganti dengan nilai yang sering muncul (*modus*).

Tabel 3. Nilai Mean dan Modus Atribut Penyakit Ginjal Kronis

Atribut	Nilai Mean/ Modus	Atribut	Nilai Mean/ Modus
Age	52	Blood glucose random	146
Blood pressure	80	Blood urea	57
Specific gravity	1.020	Serum creatinine	3
Albumin	0	Sodium	138
Sugar	0	Potassium	5
Red blood cells	Normal	Hypertension	No
Pus cell	Normal	Diabetes mellitus	No
Pus cell clumps	not present	Appetite	Good
Bacteria	not present	Pedal edema	No

Langkah selanjutnya adalah tahap transformasi data, *dataset* di diskritisasi dengan menggunakan *entropy-based discretization*. Diskritisasi diterapkan pada 7 dataset yang bertipe numerik. Tahapan ini dilakukan agar dapat melanjutkan proses perhitungan menggunakan *decision tree* C4.5. Tabel 4 berikut ini adalah merupakan hasil dari diskritisasi. Selanjutnya menghitung data penyakit ginjal kronis menggunakan metode *decision tree* C4.5.

Langkah berikutnya melakukan perhitungan untuk node akar (semua data) seperti Tabel 4. Setelah itu melakukan perhitungan *entropy* pada setiap atribut salah satu contohnya yaitu menghitung atribut *age*. Nilai pada atribut *age* bertipe numerik maka dikategorikan menjadi atribut ≤44 dan >44. Perhitungan *entropy* pada data yang telah diprediksi dilakukan seperti perhitungan pada data tidak lengkap. Berikut adalah contoh perhitungan entropy dari atribut *red blood cells* (rbc).

Tabel 4. Hasil Diskritisasi

Atribut	Diskritisasi
Age	(≤44) , (>44)
Blood pressure (Bp)	(≤80) , (>80)
Blood glucose random (Bgr)	(≤140) , (>140)
Blood urea (Bu)	(≤50) , (>50)
Serum creatinine (Sc)	(≤1,2), (>1,2)
Sodium (Sod)	(≤142) , (>142)
Potassium (Pot)	(≤5) , (>5)

$$Entropy(normal) = -\left(\frac{184}{319}\right) log_2\left(\frac{184}{319}\right) - \left(\frac{135}{319}\right) log_2\left(\frac{135}{319}\right) = 0{,}9829$$

$$Entropy(abnormal) = -\left(\frac{41}{41}\right) log_2\left(\frac{41}{41}\right) - \left(\frac{0}{41}\right) log_2\left(\frac{0}{41}\right) = 0$$

Jika semua atribut telah dihitung dan diketahui nilai *entropy* dari setiap atribut, maka langkah selanjutnya menghitung *gain* menggunakan persamaan (3), perhitungan *gain* pada data yang telah diprediksi sama dengan perhitungan pada data tidak lengkap. Berikut ini adalah perhitungan dari gain untuk atribut *red blood cells* (rbc).

$$Gain(rbc) = 0{,}9436 - \left(\frac{317}{360}\right) * (0{,}0719) - \left(\frac{43}{360}\right) * (0) = 0{,}0835$$

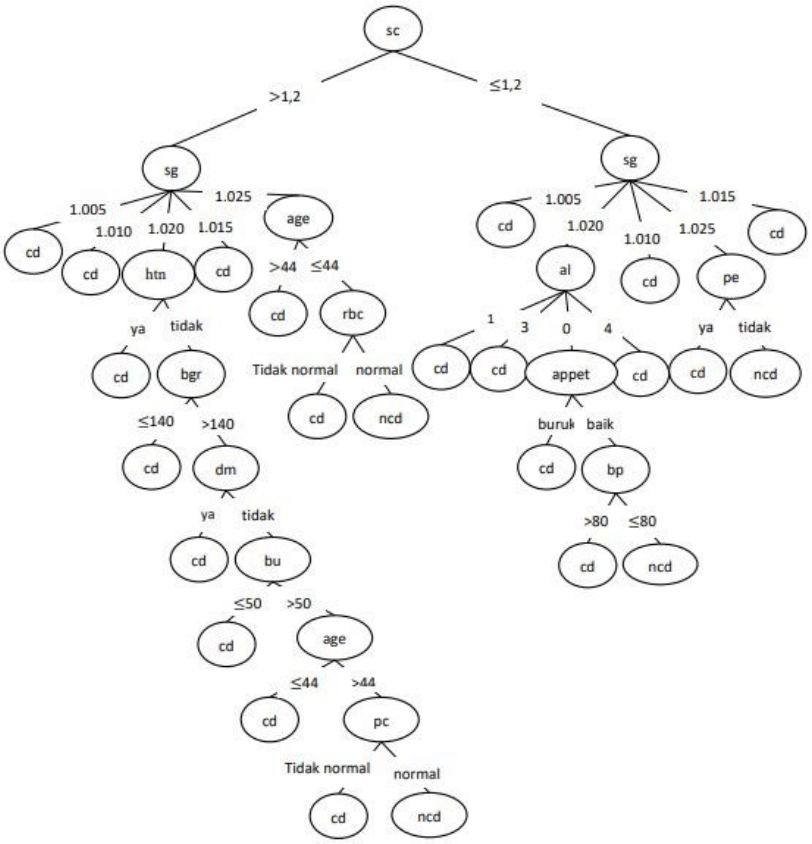
Tabel 5. Nilai Gain Data yang Telah Diprediksi

Atribut	Nilai Gain	Atribut	Nilai Gain
Age	0,0630	Bgr	0,2440
Blood pressure	0,1542	Bu	0,2312
Specific gravity	0,4478	Sc	**0,4679**
Albumin	0,3550	Sod	0,1750
Sugar	0,1106	Pot	0,0640
Red blood cells	0,0835	Htn	0,3260
Pus cell	0,1417	Dm	0,3053
Pus cell clumps	0,0747	Appet	0,1643
Bacteria	0,0350	Pe	0,1442

Tabel 5 di atas menunjukkan hasil dari nilai *gain* pada setiap atribut, dapat diketahui bahwa atribut *Serum creatinine* (sc) memiliki nilai *gain* tertinggi, senilai 0.4679, dibandingkan atribut yang lain sehingga atribut sc terpilih sebagai node awal pada perhitungan data yang telah diprediksi. Rule atau aturan yang dihasilkan pada data yang telah diprediksi sebanyak 24 (disebut rule **B1**) ditunjukan pada Gambar 3 berikut ini.

3.3. Hasil Analisis

Pada penelitian ini menggunakan 360 *data training* dengan 2 kelas yaitu positif penyakit ginjal kronis (CD) dan negatif ginjal kronis (NCD). Percobaan pertama menggunakan data yang asli dan masih terdapat *missing value* menghasilkan 21 rule (rule **A1**) dengan atribut *Specific gravity* (sg) sebagai node awal. Atribut *Specific gravity* (sg) terdapat 6 jenis nilai, namun dari 360 dataset hanya mengandung 3 nilai saja sedangkan 3 jenis nilai lainnya tidak pernah muncul. Dari keseluruhan data atribut sg terdapat 42 *missing value*. Pengujian yang dilakukan pada hasil rule **A1** menggunakan 40 *data testing*.



Gambar 3. Hasil Rule pada Data yang telah Diprediksi

Pengujian ini menghasilkan akurasi sebesar 90%, hasil akurasi pada rule **A1** dipengaruhi oleh atribut jumlah sel darah merah pada urine (rbc) yang memiliki data kosong lebih banyak dibandingkan atribut yang lain.

Pada percobaan kedua dengan penanganan *missing value* menghasilkan 24 aturan (rule **B1**) dengan atribut pengukuran kreatinin serum dalam darah (sc) yang menjadi node awal. Atribut pengukuran kreatinin serum (sc) mempunyai 2 jenis nilai yaitu ≤1,2 dan >1,2, atribut sc menjadi node awal dikarenakan tidak ada yang memiliki nilai nol. Jumlah nilai >1,2 dikelas negatif penyakit ginjal kronis lebih sedikit dibandingkan yang di kelas positif penyakit ginjal kronis. Pengujian rule **B1** menghasilkan akurasi sebesar 95%.

Dari analisis hasil yang sudah diuraikan maka terlihat sebuah perbedaan pada penerapan metode *decision tree* C4.5 dengan data tidak lengkap dan data yang telah diprediksi *missing value*. Perbedaan terdapat pada pohon keputusan dan *rule* yang dihasilkan serta nilai akurasi. Data yang mendapat penanganan *missing value* memiliki hasil lebih baik dibandingkan data yang tidak lengkap.

4. KESIMPULAN

Hasil analisis penelitian menemukan tiga poin kesimpulan sebagai berikut:

1. Prediksi missing value dapat dilakukan dengan metode mean pada tipe data numerik dan modus pada tipe data nominal.
2. Pengembangan rule decision tree C4.5 yang dihasilkan pada data tidak lengkap adalah 21 dan data dengan prediksi missing value adalah 24.
3. Akurasi yang didapatkan pada data yang tidak lengkap sebesar 90% dan akurasi pada data dengan missing value sebesar 95%.

UCAPAN TERIMA KASIH

Kolaborasi riset anggota Grup Keahlian dan Riset Laboratorium Komputasi Sistem Informasi dan Laboratorium Multimedia Cerdas Jurusan Teknologi Informasi, Politeknik Negeri Jember.

DAFTAR PUSTAKA

APRILIAWAN, Y. E., 2015. Teknik Imputasi Missing Values pada Data Mining', pp. 1–5.

ARIFIN, T. AND ARIESTA, D., 2019. 'Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization. *Jurnal Tekno Insentif*, 13(1), pp. 26–30.

DELIMA, E. T. *et al.*, 2014. Risk Factors for Chronic Kidney Disease: A Case Control Study in Four Hospitals in Jakarta in 2014. *Buletin Penelitian Kesehatan*, 45(1), pp. 17–26.

HARTINI, E., 2017. Implementation of Missing Values Handling Method for Evaluating the System/Component Maintenance Historical Data. *Jurnal Teknologi Reaktor Nuklir Tri Dasa Mega*, 19(1), pp. 11–18.

MOCH. LUTFI AND MOCHAMAD HASYIM, 2019. Penanganan Data Missing Value Pada Kualitas Produksi Jagung Dengan Menggunakan Metode K-Nn Imputation Pada Algoritma C4.5. *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, 2(2), pp. 89–104.

MUKARROMAH, M., MARTHA, S. AND ILHAMSYAH, I., 2015. Perbandingan Imputasi Missing Data Menggunakan Metode Mean Dan Metode Algoritma K-Means. *BIMASTER*, 4(3), pp. 305–312.

MURIYATMOKO, D., HARMINI, T. AND ARRAHMANTORO, M. N., 2021. Penerapan Metode Weighted Product Untuk Seleksi Kelulusan Santri Pada Sistem Informasi Wisuda Taman Pendidikan Al-Quran (TPA) Universitas Darussalam Gontor. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(2), pp. 349–356.

NUGRAHA, P. G. S. C. *et al.*, 2016. Penerapan Metode Decision Tree(Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa Smpn1 Kintamani. *Seminar Nasional Vokasi dan Teknologi (SEMNASVOKTEK)*, pp. 35–44.

NURELLISA, L. AND FITRIANAH, D., 2020. Analisis Rekomendasi Calon Debitur Motor Pada PT. Xyz Analysis of Motorcycle Debitor Recommendations in PT. Xyz Using. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 7(4), pp. 673–682.

RAHIM, R. *et al.*, 2018. C4.5 classification data mining for inventory control. *International Journal of Engineering and Technology (UAE)*, 7(2.3), pp. 68–72.

RIZALDI, T., PURNOMO, F. E. AND ARIFIANTO, A. S., 2019. Perbandingan Metode K-Nn Dan Bayes Pada Missing Imputation. *Jurnal Teknologi Informasi dan Terapan*, 5(2), pp. 85–90.

ROMAGNANI, P. *et al.*, 2017. Chronic kidney disease. *Nature Reviews Disease Primers*, 3(1), p. 17088.

SARI, S. K. AND MAHMUDY, W. F., 2019. Penerapan Metode Decision Tree dan Algoritme Genetika Untuk Klasifikasi Risiko Hipertensi', 3(3), pp. 2867–2873.

TJEKYAN, S., 2014. Prevalensi dan Faktor Risiko Penyakit Ginjal Kronik di RSUP Dr. Mohammad Hoesin Palembang Tahun 2012. *Majalah Kedokteran Sriwijaya*, 46(4), pp. 275–282.

WAHYUNI, S., 2018. Implementation of Data Mining to Analyze Drug Cases Using C4 . 5 Decision Tree Implementation of Data Mining to Analyze Drug Cases Using C4 . 5 Decision Tree. *Journal of Physics: Conference Series*, 970(1), p. 012030.

WULANDARI, R. T., 2010. Pengertian Data Mining. *Data Mining*, 7(3), pp. 3–9.

Halaman ini sengaja dikosongkan