

Implementasi Algoritma C5.0 Untuk Klasifikasi Penyakit Gagal Ginjal Kronik

Setyowati Nurhaningsih¹, Yuliana Susanti¹, Sri Sulistijowati Handajani¹

¹Program Studi Statistika, FMIPA Universitas Sebelas Maret, Surakarta 57126, Indonesia

Abstrak

Penyakit gagal ginjal kronik merupakan salah satu penyakit yang mematikan di berbagai negara, termasuk di Indonesia. Penyakit ini memiliki nilai prevalensi meningkat seiring bertambahnya jumlah penduduk. Salah satu metode yang dapat digunakan untuk memprediksi penyakit gagal ginjal kronik dalam bentuk pohon klasifikasi yaitu C5.0. Tujuan dari penelitian ini yaitu untuk menerapkan C5.0 pada klasifikasi penyakit gagal ginjal kronik dan menghitung nilai akurasi. Metode C5.0 merupakan metode klasifikasi dalam pemilihan atributnya akan diproses menggunakan informasi gain. Variabel independen yang berpengaruh dalam penelitian ini adalah eritrosit, urea, kreatinin, dan trombosit. Hasil dari penelitian ini berupa pohon klasifikasi penyakit gagal ginjal kronik. Metode C5.0 menghasilkan 6 segmen klasifikasi dengan nilai akurasi sebesar 99.3%.

Kata Kunci: C5.0, Pohon klasifikasi, Nilai akurasi

Abstract

Chronic kidney failure is one of the deadly diseases in many countries, including in Indonesia. This disease has a prevalence value increasing with the increasing population. Method that can be used to predict chronic kidney failure in the form of classification trees, namely C5.0. The purpose of this study is to apply the C5.0 to the classification of chronic kidney failure and to calculate the accuracy. Method C5.0 is a classification method in selecting its attributes to be processed using gain information. The independent variables that are influential in this study are erythrocytes, urea, creatine, and platelets. The results of this study are in the form of a classification tree for chronic kidney failure. The C5.0 method produces 6 classification segments with an accuracy value of 99.3%.

Keywords: C5.0, Classification Tree, Accuracy Value

1. PENDAHULUAN

Salah satu organ tubuh yang penting bagi manusia namun sangat mudah terserang penyakit yaitu ginjal. Fungsi utama ginjal bagi tubuh manusia adalah membentuk urin untuk mengeluarkan berbagai zat-zat racun dari tubuh. Apabila ginjal mengalami kerusakan maka akan menimbulkan komplikasi penyakit diantaranya hipertensi, anemia, penyakit tulang, gagal jantung, dan penurunan ekskresi (Smeltzer & Bare, 2008). Salah satu penyakit yang dapat menyerang ginjal yaitu gagal ginjal kronik.

Gagal ginjal kronik merupakan salah satu penyakit yang sangat mematikan di berbagai negara, termasuk Indonesia. WHO memperkirakan penderita gagal ginjal kronik di Indonesia akan mengalami peningkatan sebesar 41,4% pada tahun 1995-2025. Menurut hasil *Global Burden of Disease* tahun 2010, gagal ginjal kronik menjadi penyebab kematian urutan ke-18 pada tahun 2010. Berdasarkan data rekam medik RSUD Ir. Soekarno Kabupaten Sukoharjo, pasien gagal ginjal kronik pada tahun 2014 sebanyak 1417 orang, sedangkan bulan Januari-Juli 2015 mengala-

mi peningkatan sebesar 19,7%. Menurut Riset Kesehatan Dasar (Riskesdas, 2013), nilai prevalensi di seluruh Indonesia untuk penyakit gagal ginjal kronik mempunyai nilai rata-rata kurang lebih 0,2%. Namun banyak masyarakat yang masih belum sadar kalau mereka telah menderita penyakit ginjal, dan telah mencapai tahap gagal ginjal kronik sehingga salah satu pengobatannya adalah dengan melakukan cuci darah. Jika masyarakat menyadari dari awal atau masih pada stadium awal dan stadium dua, maka dapat dilakukan terapi tanpa cuci darah. Oleh karena itu, mengetahui keadaan ginjal sangat penting untuk mencegah penurunan fungsi ginjal.

Upaya untuk mengetahui keadaan ginjal diperlukan adanya suatu sistem pendukung keputusan yang dapat melakukan pengambilan keputusan (Amalia, 2018). Salah satu metode pengambilan keputusan yang dapat dilakukan adalah metode klasifikasi. Klasifikasi merupakan pengelompokan secara sistematis ke dalam kelas tertentu berdasarkan karakteristik yang sama. Metode klasifikasi yang dapat digunakan untuk mengklasifikasikan penyakit gagal ginjal kronik dalam bentuk visual adalah pohon klasifikasi. Pohon klasifikasi merupakan metode yang menggunakan aturan untuk menentukan kelas dari suatu objek yang mempunyai nilai-nilai variabel independen (Loh dan Shih, 2001). Pohon klasifikasi bertujuan untuk mengklasifikasikan objek yang ditampilkan melalui gambar daun dan cabang. Metode pohon klasifikasi yang digunakan dalam penelitian ini yaitu *C5.0*. Menurut Patil (2012), *C5.0* dapat mengklasifikasikan model berstruktur pohon dan aturan serta memiliki tingkat akurasi yang lebih baik dibandingkan dengan *ID3* dan *C4.5*. Pemilihan variabel metode *C5.0* akan diproses menggunakan *information gain*.

Pada penelitian sebelumnya, Ocal *et al.*(2015) membandingkan metode *CHAID* dan *C5.0* untuk memprediksi kegagalan finansial pada industri pabrik dan diperoleh hasil akurasi pada metode *C5.0* lebih baik daripada *CHAID*. Revanthy dan Lawrance

(2017) membandingkan algoritma *C4.5* dan *C5.0* pada data hama tanaman dan menghasilkan nilai akurasi pada algoritma *C5.0* lebih besar daripada *C4.5* yaitu sebesar 99.49%. Penelitian Patil *et. al.*(2012) melakukan penelitian dengan menggunakan algoritma *C5.0* dengan algoritma CART dalam memprediksi konsumen untuk memilih kartu keanggotaan dan memiliki hasil algoritma *C5.0* memiliki hasil akurasi yang lebih tinggi mencapai 99,6%.

Pada penelitian ini, diterapkan algoritma *C5.0* pada klasifikasi penyakit gagal ginjal kronik di RSUD Ir. Soekarno Kabupaten Sukoharjo dan menghitung nilai akurasinya.

2. METODE

Penelitian ini menggunakan data sekunder yang diperoleh dari RSUD Ir. Soekarno Kabupaten Sukoharjo. Teknik pengumpulan data melalui studi dokumentasi dengan mengumpulkan dokumen resmi rumah sakit yang berupa data rekam medik pasien. Data sekunder yang digunakan yaitu kelas penyakit gagal ginjal kronik di RSUD Ir. Soekarno Kabupaten Sukoharjo sebagai variabel dependen dan eritrosit, urea, kreatinin, dan trombosit sebagai variabel independen. Adapun langkah-langkah dalam penelitian ini, sebagai berikut:

1. Mendiskripsikan data penyakit gagal ginjal kronik di RSUD Ir. Soekarno Kabupaten Sukoharjo
2. Menyusun dan mengkategorikan data pada variabel dependen dan variabel independen
3. Menghitung nilai *entropy* dan *information gain*. *Entropy* adalah ukuran untuk mengetahui karakteristik dari kumpulan data. Sedangkan *information gain* digunakan untuk memilih variabel uji pada setiap node di dalam tree. Rumus yang digunakan untuk perhitungan *entropy* (Patil, 2012)

$$I(S_1, S_2, \dots, S_m) = - \sum_{i=1}^m p_i \log_2(p_i) \dots\dots\dots (1)$$

Keterangan :

- S = himpunan kasus
- m = jumlah sampel
- p_i = proporsi kelas

Untuk mendapatkan informasi nilai subset dari variabel A maka digunakan rumus sebagai berikut.

$$E(A) = \sum_{j=i}^y \frac{S_{1j} + \dots + S_{mj}}{S} (S_{1j}, \dots, S_{mj}) \dots\dots\dots(2)$$

Information gain selanjutnya didapatkan dari rumus sebagai berikut.

$$Gain(A) = I(S_1, S_2, \dots, S_m) - E(A) \dots\dots\dots(3)$$

4. Menetapkan variabel dengan nilai *information gain* tertinggi sebagai node akar dan variabel dengan *information gain* tertinggi berikutnya sebagai node cabang.
5. Melakukan tahap penghentian apabila semua data telah memperoleh kelas masing-masing.
6. Membentuk pohon klasifikasi menggunakan algoritma *C5.0*
7. Menginterpretasikan pohon klasifikasi yang terbentuk berdasarkan tingkat klasifikasi.
8. Menghitung nilai akurasi pohon klasifikasi pada algoritma *C5.0* dalam memprediksi hasil klasifikasi. Akurasi adalah presentase dari total jumlah data yang diprediksi benar untuk mengetahui seberapa baik model dalam melakukan proses klasifikasi. Tabel *confusion matrix* merupakan tabel yang digunakan untuk mencatat hasil klasifikasi dan menghitung akurasi (Han dan Kamber, 2006). Tabel *confusion matrix* ditunjukkan pada Tabel 1.

Tabel 1. Confusion Matrix

Observasi	Prediksi	
	+	-
+	True Positive (TP)	False Negative (FN)
-	False Positive (FP)	True Negative (TN)

Nilai akurasi didapat dari rumus sebagai berikut.

$$Akurasi = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \dots\dots\dots(4)$$

3. HASIL DAN PEMBAHASAN

Data yang digunakan yaitu data rekam medik pasien penderita gagal ginjal kronik di RSUD Ir. Soekarno Kabupaten Sukoharjo. Jumlah total data sebanyak 160 dengan 120 pasien terkena penyakit gagal ginjal kronik dan 40 pasien tidak terkena penyakit gagal ginjal kronik. Faktor yang digunakan dalam penelitian ini adalah eritrosit, urea, kreatinin, dan trombosit. Rincian kategori dari masing-masing variabel ditunjukkan pada Tabel 2.

Tabel 2. Tabulasi Kategori Setiap Variabel

Variabel	Kategori	Jumlah
Gagal Ginjal Kronik	Ya	120
	Tidak	40
Eritrosit	<3,8 juta/mm ³	26
	3,8-6,5 juta/mm ³	134
	>6,5 juta/mm ³	
Urea	10-50 mg/dL	46
	>50 mg/dL	114
Kreatinin	0,6-1,3 mg/dL	39
	>1,3 mg/dL	121
Trombosit	<150.000/μL	33
	150.000-350.000/μL	120
	>350.000/μL	7

Berikut ini adalah langkah-langkah dalam pembentukan pohon klasifikasi menggunakan algoritma C5.0:

1. Menentukan node akar. Langkah yang dilakukan yaitu menghitung nilai *entropy* dan *information gain* dari lima variabel data. Hasil perhitungan ditunjukkan pada Tabel 3.

Tabel 3. Perhitungan Node Akar

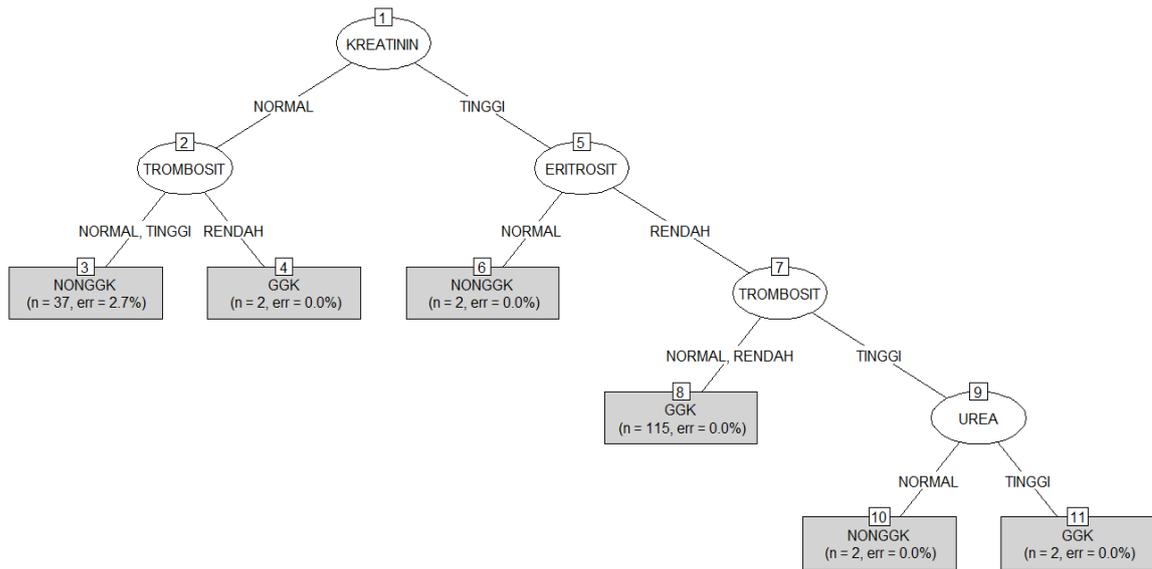
Variabel	Information gain
Eritrosit	0.4067
Urea	0.4812
Kreatinin	0.5575
Trombosit	0.1204

Dari Tabel 2 diperoleh bahwa variabel kreatinin dipilih menjadi node akar karena memiliki *information gain* tertinggi yaitu sebesar 0.5575. Terdapat dua kategori pada variabel kreatinin yaitu normal dan tinggi. Dari kedua kategori tersebut diperlukan perhitungan lebih lanjut karena belum dapat diklasifikasikan kedalam kelas gagal ginjal kronik atau tidak gagal ginjal kronik.

2. Menentukan node cabang. Node cabang dipilih berdasarkan nilai *information gain* tertinggi setelah menghapus variabel yang sudah dipilih sebagai node akar.
 - a. Iterasi 1. Node kreatinin kategori normal, didapatkan nilai *information gain* tertinggi adalah variabel trombosit sebesar 0.2244. Node

trombosit kategori normal dan tinggi telah mengklasifikasikan kedalam grup kelas tidak gagal ginjal kronik. Sedangkan node trombosit kategori rendah mengklasifikasikan kedalam grup kelas gagal ginjal kronik.

- b. Iterasi 2. Node kreatinin kategori tinggi, didapatkan nilai *information gain* tertinggi adalah variabel eritrosit sebesar 0.1406. Node eritrosit kategori normal telah mengklasifikasikan kedalam grup kelas tidak gagal ginjal kronik. Sedangkan node eritrosit kategori rendah perlu dilakukan perhitungan lebih lanjut.
 - c. Iterasi 3. Node eritrosit kategori rendah, didapatkan nilai *information gain* tertinggi adalah variabel trombosit sebesar 0.0895. Node trombosit kategori normal dan rendah telah mengklasifikasikan kedalam grup kelas gagal ginjal kronik. Sedangkan node trombosit kategori tinggi perlu dilakukan perhitungan lebih lanjut.
 - d. Iterasi 4. Node trombosit kategori tinggi, didapatkan nilai *information gain* tertinggi adalah variabel urea sebesar 1. Node urea kategori normal telah mengklasifikasikan kedalam grup kelas tidak gagal ginjal kronik. Sedangkan node urea kategori tinggi telah mengklasifikasikan kedalam grup kelas gagal ginjal kronik.
3. Hasil pohon klasifikasi ditunjukkan pada Gambar 1.



Gambar 1. Pohon Klasifikasi Algoritma C5.0

Pohon klasifikasi yang dibentuk dapat digunakan untuk menentukan klasifikasi penyakit gagal ginjal kronik sebagai berikut:

- Klasifikasi 1: Pasien yang memiliki kadar kreatinin normal dengan trombosit normal dan tinggi
- Klasifikasi 2: Pasien yang memiliki kadar kreatinin normal dengan trombosit rendah
- Klasifikasi 3: Pasien yang memiliki kadar kreatinin tinggi dan eritrosit normal
- Klasifikasi 4: Pasien yang memiliki kadar kreatinin tinggi, eritrosit rendah, dengan trombosit normal dan rendah
- Klasifikasi 5: Pasien yang memiliki kadar kreatinin tinggi, eritrosit rendah, trombosit tinggi, dan urea normal
- Klasifikasi 6: Pasien yang memiliki kadar kreatinin tinggi, eritrosit rendah, trombosit tinggi, dan urea tinggi

Berdasarkan tingkatan klasifikasi diketahui bahwa pasien yang memiliki risiko penyakit gagal ginjal kronik adalah pada klasifikasi yang ke 2, 4, dan 6 dengan hasil presentasinya sebesar 100%. Sedangkan jumlah pasien yang tidak memiliki risiko penyakit gagal ginjal kronik adalah pada klasifikasi yang ke 3 dan 5 dengan jumlah dengan hasil presentasinya sebesar 100% dan klasifikasi yang ke 1 dengan hasil presentasinya 97.3%.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, diperoleh kesimpulan bahwa implementasi algoritma C5.0 pada klasifikasi penyakit gagal ginjal kronik dengan faktor eritrosit, urea, kreatinin, dan trombosit memiliki struktur pohon klasifikasi dengan enam tingkatan klasifikasi. Hasil pengujian akurasi yang diperoleh sebesar 99.3%.

DAFTAR PUSTAKA

- Amalia, H, 2018, Perbandingan Metode Data Mining SVM dan NN untuk Klasifikasi Penyakit Ginjal Kronis, *Jurnal Pilar Nusa Mandiri*, Vol. 14 pp. 1, 1-5.
- Loh, W. and Shih, T., 2001, Selection Methods for Classification Trees, *Statistica Sinica* 7, pp. 815-840.
- Ocal, N., Ercan M. K., and Kadioglu, E., 2015, Predicting Financial Failure Using Decision Tree Algorithms: An Empirical Test on the Manufacturing Industry at Borsa Istanbul, *International Journal of Economics and Finance*, Vol. 7 pp. 189-206
- Patil N, Lathi R, Chitre V, 2012, Customer Card Classification Based on C5.0 & CART Algorithms, *International Journal of Engineering Research & Technology (IJERT)*, Vol. 2 pp. 164-167
- Revathy, R and Lawrance, R., 2017, Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 5 pp. 50-58
- Riskesdas, 2013, *Laporan Hasil Riset Kesehatan Dasar (Riskesdas)*, Badan Penelitian dan Pengembangan Kesehatan Kementerian RI, Jakarta
- Smeltzer dan Bare, 2008, *Buku Ajar Keperawatan Medikal Bedah*, EGC, Jakarta