UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 1 Tahun 2024 Halaman 84-89.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Classification of Stroke Disease at Rumah Sakit Otak Dr. Drs. M. Hatta Bukittinggi With C4.5 Decision Tree Algorithm

Futiah Salsabila, Zamahsary Martha*, Atus Amadi Putra, dan Admi Salma
Departemen Statistika, Universitas Negeri Padang, Kota Padang, Negara Indonesia
*Corresponding author: zamahsarymartha@fmipa.unp.ac.id

Submitted : 04 Desember 2023
Revised : 02 Februari 2024
Accepted : 16 Februari 2024

ABSTRACT

Stroke is a health condition that has vascular disorders where brain function is related to problems with blood vessels that carry blood to the brain. Several factors that can influence stroke include unhealthy eating habits, lack of physical activity, smoking behavior, alcohol consumption, and obesity. The symptoms experienced are headache, nausea, vomiting, blurred vision and difficulty swallowing. The researcher's aim is to determine the risk faktors that affect the incidence of stroke hospitalization based on stroke diagnoses at Rumah Sakit Otak Dr. Drs. M. Hatta Bukittinggi city by classifying each variable using a decision tree. A decision tree is a flowchart that resembles a branching tree. The C4.5 algorithm is used in this research, which can process numerical and categorical data, can handle missing attribute values, and produces rules that are easy to interpret. The results of the analysis show that the attribute that is a risk factor for stroke is the heart. The model created using the C4.5 algorithm was tested using a counfusion matrix resulting in an accuracy of 64.54%, a precision of 53.34% for classifying ischemic stroke patients correctly, and a recall of 72.73% for classifying hemorrhagic patients correctly.

Keywords: *C4.5 Algorithm*, *Classification*, *Stroke, Decision Tree*

I. PENDAHULUAN

Kesehatan berperan penting dalam kehidupan, dan gaya hidup yang tidak sehat memiliki dampak negatif pada kesehatan fisik. Gaya hidup tidak sehat mencakup pola makan yang tidak sehat, kurangnya aktivitas fisik, kelebihan berat badan, dan konsumsi alkohol yang berlebihan, yang semuanya merupakan ancaman serius bagi kesehatan dan dapat menyebabkan berbagai jenis penyakit salah satunya adalah penyakit stroke (Kemenkes, 2018).

Cerebrovascular Disease atau Stroke adalah penyakit *cerebrovascular* di mana munculnya gangguan fungsi otak dikaitkan dengan gangguan pada pembuluh darah yang mensuplai darah ke otak (Widyaswara dkk, 2019). Menurut *World Health Organization* (WHO) tahun 2020, stroke adalah masalah kesehatan yang masih menjadi perhatian kesehatan di dunia. Penyakit stroke termasuk penyakit tidak menular yang dialami secara tiba-tiba. Penyakit stroke terjadi akibat pembuluh darah diotak pecah atau mengakibatkan penyumbatan, sehingga aliran darah terganggu dan mengakibatkan adanya bagian diotak tidak mendapatkan pasokan oksigen. Hal ini terjadi karena sel atau jaringan di otak tidak berfungsi (P2PTM Kemenkes RI, 2018).

Penyakit stroke diklasifikasikan sebagai stroke hemoragik dan stroke iskemik. Penyebab Stroke hemoragik yaitu pecahnya pembuluh darah keotak. Kondisi ini menimbulkan gangguan pada sistem saraf yang terjadi secara mendadak serta nyeri kepala pada saat melakukan aktivitas akibat tekanan yang diberikan oleh otak. Sedangkan stroke iskemik adalah gangguan sirkulasi darah keotak tanpa perdarahan yang disebabkan kelemahan disemua bagian tubuh (Wanhari, 2018).

Salah satu upaya yang harus diperhatikan dalam mencegah angka kejadian stroke yaitu mengetahui faktor risiko yang dimiliki seseorang. Penyakit stroke dipengaruhi oleh faktor risiko yaitu faktor tidak bisa diubah mencakup umur dan jenis kelamin, sementara faktor yang diubah melibatkan hipertensi, diabetes mellitus, riwayat jantung, *body mass index* (BMI), dan kebiasaan merokok termasuk atribut klasifikasi. Atribut merupakan faktor risiko penyakit stroke. Berdasarkan hal tersebut, metode klasifikasi dapat digunakan dengan tujuan untuk mendapatkan faktor risiko utama penyakit stroke. Salah satu metode klasifikasi adalah pohon keputusan (*decision tree*) yang merupakan sebuah diagram alur mirip dengan struktur pohon, dimana setiap simpul menotasikan atribut yang akan diuji, cabang

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/135

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

menunjukkan hasil dari atribut dan simpul daun mempresentasikan kelas-kelas tertentu (Hand dkk, 2001). Menurut Sifaunajah dkk (2022) keistimewaan dari pohon keputusan ini adalah visualisasi yang disajikan dalam bentuk pohon sehingga prosedur prediksinya dapat diamati dengan mudah. Algoritma yang digunakan dalam klasifikasi yakni C4.5. Menurut Youn dkk (2006) algoritma C4.5 dapat mengolah data numerik dan kategorik, serta dapat menangani nilai hilang, dan menghasilkan aturan-aturan yang mudah diinterpretasikan. Untuk menghasilkan fakktor risko yang mempengaruhi penyakit stroke yang dilakukan dengan cara mengklasifikasikan masing-masing kelas menggunakan pohon keputusan.

Algoritma C4.5 yaitu pengembangan dari algortima ID3, kekurangan algoritma tersebut dapat disembunyikan oleh algoritma C4.5 (Aldino dan Sulistiani, 2020). Algoritma C4.5 memiliki kelebihan yaitu memperoleh pengukuran akurasi yang baik, dapat menangani variabel kategori dan numerik, menangani data hilang, dan menghasilkan aturan yang dapat dijelaskan (Pangaribuan dkk, 2019). Untuk menganalisis apa faktor risiko utama penyakit stroke dapat dilakukan dengan cara mengklasifikasikan masing-masing kelas menggunakan pohon keputusan.

II. METODE PENELITIAN

A. Sumber Data dan Variabel Penelitian

Data yang diperoleh adalah data sekunder. Data digunakan yaitu riwayat medis pasien stroke yang dirawat di Rumah Sakit Otak DR. Drs. M. Hatta Kota Bukittinggi (RSOMH) Tahun 2022. Atribut sebagai berikut yaitu umur, jenis kelamin, hipertensi, jantung, diabetes mellitus, merokok, BMI, dan diagnosa stroke (stroke iskemik dan hemoragik).

B. Algoritma C4.5

Algoritma C4.5 menerapkan untuk membentuk pohon keputusan, data yang digunakan pada algoritma C4.5 yaitu data yang bersifat kategorik dan numerik. Hasil yang telah dilakukan dari proses pengelompokan dapat membentuk aturan untuk memprediksi nilai atribut kategori dari *record* yang baru. Algoritma C4.5 dapat mengatasi data hilang (Elisa, 2017). Terdapat beberapa langkah membentuk pohon keputusan algoritma C4.5 (Melina, 2016):

1. Menghitung nilai *entropy*.

$$Entropy\ (S) = -\sum_{i=1}^{n} p_i * log_2 p_i \tag{1}$$

Keterangan:
p_i : proporsi setiap kategori
S : jumlah data
n : jumlah partisi S

2. Menghitung nilai *gain information*.

$$Gain\ information\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \tag{2}$$

Keterangan:
$|S_i|$: proporsi setiap kategori terhadap jumlah data
$|S|$: jumlah data
A : jumlah atribut

3. Menghitung nilai *split information*.

$$Split\ Information\ (S, A) = -\sum_{i=1}^{n} \left|\frac{S_i}{S}\right| + log_2 \left|\frac{S_i}{S}\right| \tag{3}$$

Keterangan:
S_i : banyaknya sampel untuk atribut A dengan kelas ke-i

4. Menghitung nilai *gain ratio*.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/135

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$Gain\ Ratio(S,A) = \frac{Gain\ Information\ (S,A)}{Split\ Information\ (S,A)} \quad (4)$$

5. Mengulang semua langkah 1 sampai 4 untuk setiap cabang, sehingga semua cabang mempunyai daun keputusan.
6. Membuat aturan berdasarkan pohon keputusan.

C. Evaluasi Ketepatan Hasil

Menurut Liu (2015) *confusion matrix* yaitu salah satu metode yang digunakan untuk mengukur kinerja suatu metode klasifikasi. Ada 4 istilah sebagai represntasi proses pengelompokan. Pada *confusion matrix* dihasilkan nilai *True Positive* (TP), *False Negative* (FN), *True Negative* (TN), *False positive* (FP). Nilai yang akan diperoleh disajikan pada Tabel 1.

Tabel 1. *Confussion Matrix*

Classification		*Predicted Class*	
		Hemoragik	*Iskemik*
Aktual Class	*Hemoragik*	*True Positif* (TP)	*False Positif* (FP)
	Iskemik	*False Negatif* (FN)	*True Negatif* (TN)

Metode *confusion matrix* digunakan untuk menguji nilai *precision, recall*, akurasi dari algoritma C4.5. Untuk menghitung semua nilai-nilai kinerja pada rumus yang disajikan pada Persamaan 5,6, dan 7.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (5)$$

$$precision = \frac{TP}{TP + FP} * 100\% \quad (6)$$

$$recall = \frac{TP}{TP + FN} * 100\% \quad (7)$$

Agar sistem dapat memberikan nilai informasi dan tingkat keakuratan pada peneliti, maka terbagi tiga informasi yaitu *accuracy*, *precision*, *recall*. *Accuracy* dijelaskan sebagai tingkat kesesuaian antara nilai yang diproyeksikan dengan nilai sebenarnya dengan menilai data yang telah diklasifikasikan dengan benar ke seluruh data set. *Precision* adalah jumlah data stroke diklasifikasikan dengan benar dibagi dengan data golongkan positif. *Recall* adalah jumlah persentase data berkategori positif yang dapat diklasifikasikan oleh sistem secara akurat untuk menentukan seberapa suksesnya pemulihan informasi.

III. HASIL DAN PEMBAHASAN

A. Analisis Deskriptif Statistika

Penjelasan karakteristik pasien yang mengalami stroke rawat inap di RSOMH Kota Bukittinggi terdapat di Gambar 1.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/135

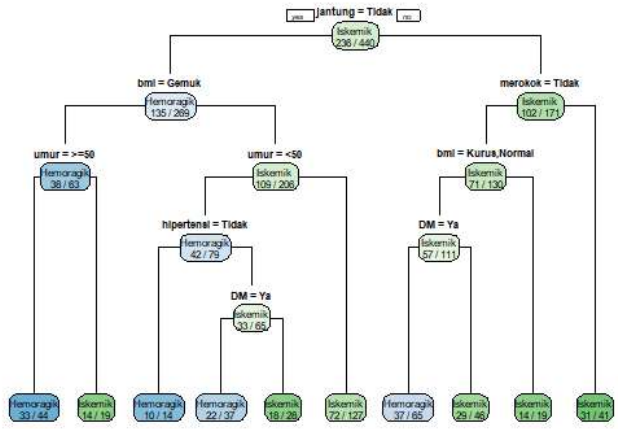
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Gambar 1. Statistik Deskriptif

Berdasarkan Gambar 1, memberikan informasi bahwa dari 550 pasien riwayat medis kasus stroke terdapat 52% atau 286 pasien terkena stroke Iskemik dengan 311 dari 550 pasien penyakit stroke memiliki jenis kelamin laki-laki. Mayoritas pasien terkena stroke memiliki riwayat hipertensi yaitu sebesar 429 pasien dan sebanyak 337 pasien tidak memiliki riwayat jantung dan sebanyak 285 pasien tidak memiliki riwayat diabetes mellitus dan sebanyak 348 pasien memiliki *body mass index* normal dengan 435 dari 550 pasien tidak merokok.

B. Pembentukan Pohon Klasifikasi

Algoritma C4.5 menghitung dan mengklasifikasikan 550 data riwayat medis pasien stroke rawat inap di RSMOH Kota Bukittinggi pada tahun 2022. Dimana data latih sebanyak 440 dan data uji sebanyak 110. Analisis algoritma C4.5 menggunakan software Rstudio. Dataset dihitung dengan memasukkan data yang sudah dicleaning, setelah itu masukkan data dan menjadikan algoritma C4.5 untuk mencari nilai pada rumus Persamaan 1 sampai dengan Persamaan 4. Untuk penentuan simpul akar dilihat dari nilai gain rasio tertinggi yang kemudian menjadi simpul akar. Proses perhitungan terus diulang hingga setiap simpul memiliki sebuah keputusan. Setelah memperoleh keputusan pada data latih, algoritma secara cepat menguji keputusan tersebut pada data uji dan membentuk sebuah keputusan dan prediksi. Penerapan analisis menggunakan *confusion matrix* seperti yang disajikan Tabel 1. Pada Gambar 2 berikut hasil dari algoritma C4.5.



Gambar 2. Hasil Algoritma C4.5

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/135

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Hasil pohon keputusan algoritma C4.5 menjelaskan bahwa penyakit jantung merupakan atribut yang menjadi faktor utama penyakit stroke yang rawat inap di RSOMH Kota Bukittinggi, karna penyakit jantung mempunyai nilai *gain ratio* tertinggi maka atribut tersebut yang menjadi simpul akar. Pohon keputusan yang terbentuk terdiri dari satu simpul akar yang menentukan faktor risko utama, delapan simpul internal yang menunjukkan atribut dan sepuluh simpul daun yang menunjukkan simpul internal. Atribut yang ditempatkan pada simpul akar yaitu yang memiliki nilai *information gain* tertinggi seperti jantung berkategori tidak. Pada atribut BMI, umur, diabetes mellitus, merokok, dan hipertensi yang menunjukkan simpul internal. Pada simpul daun adalah diagnosa stroke hemoragik dan iskemik.

C. Evaluasi Ketepatan Hasil Klasifikasi

Ketepatan klasifikasi menggunakan *confusion matrix*. Berikut pada Tabel 2 yang merupakan hasil pengujian *confusion matrix* yang terdiri dari dua kategori yaitu Hemoragik dan Iskemik.

Tabel 2. Hasil Perhitungan *Counfision Matrix*

Classification		*Predicted Class*	
		Hemoragik	*Iskemik*
Aktual Class	*Hemoragik*	32	28
	Iskemik	12	38

Berdasarkan Tabel 2 hasil perhitungan *counfusion matrix* yang dihasilkan jumlah pasien Hemoragik yang juga diperdiksi Hemoragik *recall* sebanyak 32. Sedangkan pasien yang Iskemik diprediksi Iskemik *precision* sebanyak 12 orang, maka hasil perhitungan dihitung dengan Persamaan (5), (6), dan (7).

$$accuracy = \frac{32 + 38}{32 + 38 + 28 + 12} * 100\% = \ 64{,}54\%$$

$$precision = \frac{32}{32 + 28} * 100\% = 53{,}34\ \%$$

$$recall = \frac{32}{32 + 12} * 100\% = 72{,}73\%$$

Dari perhitungan diatas, nilai *accuracy* data prediksi sebesar 64,54% dapat disimpulkan bahwa pohon yang dilatih mampu mengklasifikasikan data sebesar 64,54%. Nilai *precision* yang didapatkan sebesar 53,34% untuk ketepatan klasifikasi pada pasien stroke Iskemik dan nilai *recall* yang diperoleh untuk mengukur ketetapan klasifikasi pada pasien stroke Hemoragik sebesar 72,73%.

IV. KESIMPULAN

Berdasarkan analisis yang dilakukan dengan klasifikasi menggunakan pohon keputusan algoritma C4.5 pada pasien stroke rawat inap di RSOMH Kota Bukittinggi, yang menjadi faktor risiko utama yaitu pasien yang tidak memiliki penyakit jantung. Faktor lainnya yaitu umur pasien kurang dari 50 tahun, BMI dengan kategori kurus, normal, dan gemuk, tidak memiliki riwayat hipertensi, memiliki riwayat diabetes mellitus, dan tidak merokok. Ketetapan hasil klasifikasi menggunakan *confusion matrix* menghasilkan nilai *accuracy* sebesar 64,54%, *precision* sebesar 53,34% untuk mengkalsifikasian pasien stroke Iskemik dengan benar, dan *recall* sebesar 72,73% untuk mengklasifikasi pasien Hemoragik dengan benar.

DAFTAR PUSTAKA

Elisa, Erlin,. (2017). Analisis Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengindentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Konstruksi PT. Arupadhatu. Jurnal, JOIN Vol. 2 No.1, ISSN: 2527-9165.

Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining. MIT Press.*

Kementerian Kesehatan RI. (2018). Hasil Riset Kesehatan Dasar (Riskesdas) 2018. Jakarta: Badan Penelitian dan Pengembangan Kesehatan Kementerian RI.

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/135

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Liu, B. (2015). Data Mining Exploring Hyperlinks, Contents, And Usage Data. In M. Carey & S. Ceri (Eds.)., *Global Journal of Pure and Applied Mathematics* (Second Edi,Vol.11,Issue 5). Springer. https://doi.org/10.1007/978-3-642-19460-3.

Melina, Agustina, D,. (2016). Analisis Perbandingan Algoritma ID3 Dan C4.5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum Pada PDAM Kabupaten Kendal. *Journal of Applied Intelligent System*, 234–244.

P2PTM, Kemenkes RI. (2018). Definisi Asma Direktorat Pencegahan Dan Pengendalian Penyakit Tidak Menular Kementerian Kesehatan Republik Indonesia from diambil dari

Quinlan, J. R. (1996). *Learning Decision Tree Classifiers*. *ACM Computing Surveys*, 71–72.

Pangaribuan, J. J, Tedja, C, & Wibowo, S. (2019). Membandingkan Algoritma C4.5 Dan Extreme Learning Machine Untuk Mendiagnosis Penyakit Jantung Koroner. In *PSDKU Medan Jurusan Teknik Informatika Informatics Engineering Research And Technology.*

Sifaunajah, A., dan Wahyuningtyas, R. D. (2022). Penggunaan Algoritma ID3 untuk Klasifikasi Data Calon Perserta Didik. CSRID Journal.

Wanhari, M. A. (2008). Asuhan Keperawatan Stroke. Diakses 29 Juli 2012.

Widyaswara ,S, P. A., Widodo, WT, & Setianingsih, E. (2019). Faktor Risiko yang Mempengaruhi Kejadian Stroke.Jurnal Keperawatan, 11 (4), 251 – 260. https://doi.org/10.32583/keperawatan.v11i.

WHO. World Health Organization. (2020). Definition of Stroke. https://www.publichealth.com.ng/world-health-organization-who-definition-of-stroke

Youn, S. D. Mcleod, A Comparative Study for Email Classification. Proceedings of International Joint Conferences on Computer, Information, System Sciences and Engineering, Bridgeport, CT, (2006).

DOI: https://doi.org/10.24036/ujsds/vol2-iss1/135