Teknika 16 (02): 221 – 226

JURNAL TEKNIKA
ISSN: 0854-3143
e-ISSN: 2622-3481
Journal homepage: http://jurnal.polsri.ac.id/index.php/teknika
Journal Email: teknika@polsri.ac.id

Klasifikasi Penyakit Stroke Menggunakan Algoritma *Decision Tree* C.45

Randi Estian Pambudi*[1], Sriyanto[2], Firmansyah[3]
[1,2,3] Fakultas Ilmu Komputer, Informatics & Business Institute Darmajaya, Jl. 2.A. Pagar Alam No. 93, Bandar Lampung - Indonesia 35142; Telp. (0721) 787214 Fax. (0721) 700261
e-mail: *[1]**randiestian@darmajaya.ac.id**, [2]sriyanto@darmajaya.ac.id, [3]firmansyahyunialfi@darmajaya.ac.id

Abstrak

Stroke merupakan suatu gangguan fungsi otak baik lokal maupun menyeluruh yang akan menyebabkan pasokan darah ke otak terganggu secara cepat dan berlangsung lebih dari 24 jam atau berakhir dengan kematian. Stroke juga merupakan salah satu jenis penyakit yang paling mematikan di Indonesia. Pentingnya mengetahui gejala penyakit stroke sejak dini merupakan pencegahan awal. Maka dari itu, dilakukan penelitian untuk menganalisa data terkait dengan penyebab stroke. Adapun atribut yang terlibat dalam penyebab terjadinya stroke yakni, usia, jenis kelamin, kadar glukosa, riwayat penyakit jantung, hipertensi, tipe pekerjaan, tipe tempat tinggal, status merokok, index masa tubuh dan status pernikahan. Diperlukan algoritma tertentu untuk mengklasifikasikan semua atribut tersebut. Decision Tree C4.5 merupakan Algoritma yang paling banyak digunakan, dalam kasus ini akurasi dari algoritma Decision Tree C4.5 sebesar 99.07%.

Kata kunci*— Stroke, Akurasi, Algoritma Decision Tree C4.5.*

Abstract

Stroke is a disorder of brain function, both local and comprehensive, which will cause the blood supply to the brain to be disrupted quickly and last more than 24 hours or end in death. Stroke is also one of the deadliest types of disease in Indonesia. The importance of knowing the symptoms of stroke early is an early prevention. Therefore, a study was conducted to analyze data related to the causes of stroke. The attributes involved in the cause of stroke are age, gender, glucose level, history of heart disease, hypertension, type of work, type of residence, smoking status, body mass index and marital status. A certain algorithm is needed to classify all these attributes. Decision Tree C4.5 is the most widely used algorithm, in this case the accuracy of the Decision Tree C4.5 algorithm is 99.07%.

Keywords*— Stroke, Accuracy, Decision Tree C4.5 Algorithm*

1. PENDAHULUAN

Stroke merupakan masalah kesehatan utama bagi masyarakat modern. Pada saat ini, stroke semakin menjadi masalah serius yang dihadapi hampir diseluruh dunia. Hal tersebut dikarenakan serangan stroke yang mendadak dapat mengakibatkan kematian, kecacatan fisik dan mental baik pada usia produktif maupun usia lanjut [1].

Penyakit Stroke merupakan jenis penyakit yang mematikan dimana masuk kedalam 10 jenis penyakit yang paling mematikan di Indonesia. Hal ini dilihat berdasarkan pada data yang dikumpulkan dari sampel yang mewakili Indonesia, meliputi 41.590 kematian sepanjang 2014 dan pada semua kematian itu dilakukan autopsi verbal, sesuai pedoman Badan Kesehatan Dunia (WHO) secara *real-time* oleh dokter dan petugas terlatih [2].

Stroke merupakan salah satu penyakit yang paling banyak diderita oleh masyarakat Indonesia dan menjadi urutan pertama penyebab kematian tertinggi disusul oleh diabetes dan hipertensi [3].

Sudah banyak penelitian yang dilakukan diantaranya: Menurut jurnal dari Sigit Abdillah yang berjudul Penerapan Algoritma *Decision Tree* C4.5 Untuk Diagnosa Penyakit Stroke Dengan Klasifikasi Data Mining Pada Rumah Sakit Santa Maria Pemalang, memaparkan bahwa untuk studi kasus penyakit stroke dapat memanfaatkan teknik klasifikasi data mining dengan algoritma C4.5 sebagai klasifikasi stroke atau non-stroke. Dari metode klasifikasi data *mining* dengan algoritma C4.5 dan pengaplikasian pohon keputusan yang membentuk aturan tersebut terdapat akurasi pada data *training* yang berjumlah 130 dari 156 data pasien sebesar 82,31% sedangkan akurasi pada data testing yang berjumlah 26 dari 156 data pasien sebesar 76,92%. Perhitungan keduanya menggunakan *confusion matrix*. Penelitian ini bertujuan untuk menerapkan algoritma *Decision Tree* C4.5 dalam klasifikasi penyakit stroke.

2. METODE PENELITIAN

2.1 Jenis Penelitian

Dalam penelitian ini terdapat dua pendekatan utama, yaitu pendekatan kualitatif dan pendekatan kuantitatif. Pendekatan kualitatif digunakan untuk menganalisa kajian literatur yang berkenaan dengan variabel-variabel yang digunakan dalam pengumpulan data [4].

Sedangkan pendekatan kuantitatif merupakan metode penelitian yang digunakan sebagai penelitian populasi atau sempel tertentu, pengumpulan data menggunakan instrumen penelitian dan analisis data bersifat statistik atau kuantitatif dengan tujuan untuk menguji hipotesis yang telah ditetapkan [5].

2. 2 Metode Pengumpulan Data

Pada tahap ini metode pengumpulan data dibagi menjadi dua sumber data yakni data primer dan data skunder. Adapun data primer yaitu data yang dikumpulkan dari sumbernya langsung. Sedangkan data skunder yaitu data yang dikumpulkan dari peneliti sebelumnya artinya peneliti tidak harus mengambil datanya langsung ke lapangan. Dalam penelitian ini metode pengumpulan data untuk mendapatkan sumber data yang akan digunakan adalah dengan metode pengumpulan data skunder. Data dari penelitian ini diambil dari situs kaggle *dataset repository* (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset) sedangkan untuk data pendukung didapatkan dari buku jurnal dan publikasi lainnya.

2. 3 Validasi Data

Validasi data dimana data penelitian ini diambil dari *Kaggle Dataset* yang terdiri 11 atribut dimana terdiri dari 10 atribut Fitur dan 1 atribut label. Pada *dataset* tersebut dilakukan valiadasi data atau *preprocessing* dengan cara menghilangkan data *missing value*. Data yang didapat dari situs *Kaggle Dataset* sebanyak 5110 *record*, dari data tersebut diantaranya 204 diidentifikasi sebagai data

missing value pada atribut BMI. Kemudian data *missing value* tersebut akan dihilangkan agar data menjadi valid sehingga data yang diolah tanpa *missing value* adalah 4.906 *record*.

Tabel 1 Atribut, *Type* Data Dan Nilai Kategori Dalam Klasifikasi Penyakit Stroke

Atribut Label		
Atribut	Type	Nilai
Stroke	Binominal	Yes
		No

Atribut Fitur		
Atribut	Type	Nilai
Gender	Binominal	Male
		Female
Age	Integer	1-82
Hypertension	Binominal	1
		0
Heart_disease	Binominal	1
		0
Ever_married	Binominal	Yes
		No
Work_type	Polynominal	Private
		Self-employed
		Govt_job
		Children
Residence_type	Binominal	Urban
		Rural
Avg_glucose_level	Real	59-160
Bmi	Real	12-78
Smoking_status	Polynominal	Formerly smoked
		Never smoked
		smoked

Berikut merupakan deskripsi dari atribut Tabel 1.

1. Stroke
   Stroke merupakan atribut label yang mencakup pasien yang terkena stroke dengan *kategory Yes* (Menderita Stroke) dan *No* (Tidak Stroke).
2. *Gender* (Jenis Kelamin)
   Gender/Jenis Kelamin merupakan atribut yang mencakup jenis kelamin pasien dengan kategori *Male* (Laki-laki) dan *Female* (Perempuan).
3. *Age* (Umur)
   Age/Umur merupakan atribut yang mencakup usia dataset ini memiliki data umur dari 1 tahun – 82 tahun.
4. *Heypertension* (Hipertensi)
   Hypertension merupakan atribut yang mencakup pasien memiliki penyakit hipertensi/ darah tinggi atau tidak.
5. *Heart_disease* (Riwayat Jantung)
   Heart_disease merupakan atribut yang mengindikasikan apakah pasien yang bersangkutan memiliki riwayat penyakit jantung atau tidak.
6. *Ever_married* (Status Pernikahan)
   Ever_married merupakan atribut yang mengindikasikan apakah pasien yang bersangkutan pernah menikah atau tidak.
7. *Work_type* (Tipe Pekerjaan)
   Work_type merupakan atribut yang mengindikasikan type pekerjaan dari pasien diantaranya:
   a. *Children* (Anak-Anak)
   b. *Private* (Pribadi)
   c. *Self Employed* (Bekerja Sendiri)
   d. *Govt_job* (Pekerja Pemerintahan)
8. *Residence_type* (Tipe Tempat Tinggal)
   Residence_type merupakan atribut yang mencakup tempat tinggal dari pasien yang meliputi *Urban* (Perkotaan) atau *Rural* (Pedesaan).
9. *Avg_glucose_level* (Kadar Glukosa)
   Avg_glucose_level merupakan atribut yang mencakup rata-rata tingkat kadar gula dalam darah.
10. BMI (*Body Masss Index*)
    BMI (*Body Masss Index*) merupakan atribut yang ditentukan dengan cara menghitung berat badan dalam kilogram dibagi dengan tinggi badan dalam meter kuadrat.

$$\text{BMI} = \frac{(\text{weight in kilograms})}{\text{height in meters}^2} \quad (1)$$

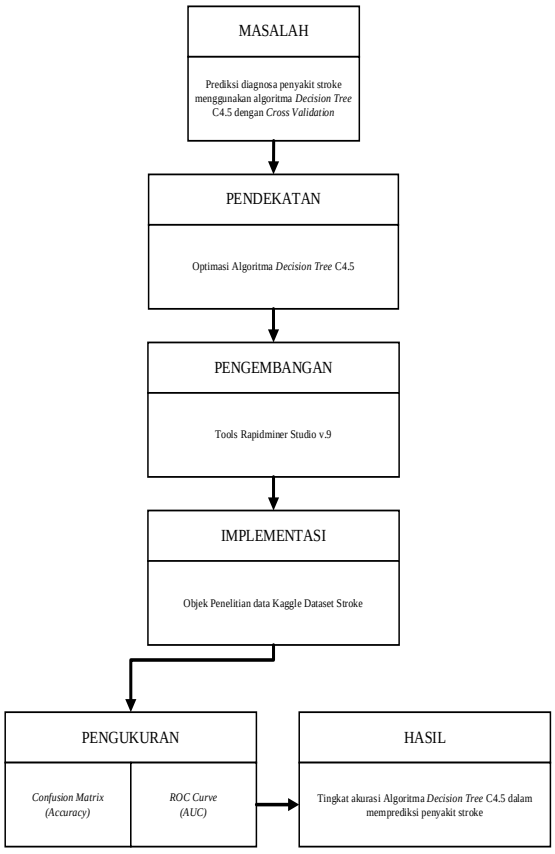
11. *Smoking_status* (Status Merokok)
    Smoking_status merupakan atribut yang mengindikasikan perokok atau bukan dari pasien dengan kategori:

a. *Formerly smoked* (Sebelumnya merokok).
b. *Smoked* (Merokok)
c. *Never Smoked* (Tidak pernah merokok).

2. 4 Kerangka Pemikiran

Penelitian ini terdiri dari beberapa tahapan. Permasalahan pada penelitian ini yaitu membandingkan penelitian sebelumnya mengenai klasifikasi penyakit stroke dengan menggunakan algoritma yang sama yaitu *Decision Tree* C4.5 untuk mendapatkan akurasi yang tinggi.

Atas dasar alasan tersebut, maka dilakukan penelitian menggunakan metode klasifikasi algoritma *Decision Tree* C4.5 dalam memprediksi penyakit stroke. Pengujian metode dilakukan dengan cara *confusion matrix* dan kurva ROC, menggunakan tools Rapidminer studio v.9. berikut dibawah ini kerangka pemikiran yang penulis buat untuk penelitian ini, dapat dilihat pada Gambar 1.



Gambar 1. Kerangka Pemikiran

2. 5 Data Mining

Data mining merupakan suatu istilah yang dapat digunakan untuk menguraikan penemuan pengetahuan di dalam suatu *database*. Data *mining* juga adalah sebuah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [6].

2. 5 Algoritma Decision Tree C4.5

Decision Tree adalah struktur *flowchart* yang menyerupai *Tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes,dan simpul daun merepresentasikan kelas atau distribusi kelas [7].

C4.5 merupakan salah satu solusi pemecahan kasus yang sering digunakan dalam pemecahan masalah pada teknik klasifikasi [8]. Pemilihan atribut sebagai simpul, baik simpul akar (*root*) atau simpul internal didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Penghitungan nilai *Gain* digunakan rumus Persamaan 1.

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

S : Himpunan kasus
A : Atribut
n : Jumlah partisi himpunan atribut A
||: Jumlah kasus pada partisi ke- i
| *S* | : Jumlah kasus dalam S

Untuk menghitung nilai *Entropy* dapat dilihat pada Persamaan 2

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 (pi) \quad (3)$$

n : Jumlah partisi S
pi : Proporsi dari terhadap *S*

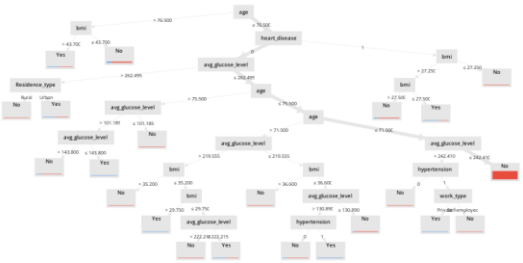
3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk menerapkan algoritma *Decision Tree* C4.5 untuk mendapatkan tingkat akurasi

penyakit stroke. Hasil dari penelitian ini berupa pengolahan data kualitatif dan data kuantitatif dengan perhitungan yang dilakukan pada sebuah dataset publik.

1. *Pemodelan Algoritma Decision Tree C4.5*

Pemodelan algoritma *Decision Tree* C4.5 terdiri dari 11 atribut yang merupakan atribut dari diagnosis penyakit stroke dan *class* yang merupakan hasil akhir prediksi. Model dari algoritma *Decision Tree* C4.5 yaitu berupa pohon keputusan, agar leih mudah dalam membuat pohon keputusan, langkah pertama adalah menghitung jumlah *class* yang berpotensi terkena penyakit stroke dan tidak stroke dari masing-masing *class* berdasarkan atribut yang telah ditentukan dengan menggunakan data *training*.



Gambar 2. Model Pohon Keputusan Klasifikasi Penyakit Stroke Menggunakan Algoritma C4.5

Pada pemodelan pohon keputusan pada Gambar 2 terdapat 23 rule yang dihasilkan dari algoritma klasifikasi *Decision Tree* C4.5, dengan jumlah class sebanyak 14 rule (*no stroke*) dan 9 *rule* (yes *strokes*). Sedangkan dari atribut asli yaitu 11 terseleksi menjadi 8 atribut yang terdiri dari *Age, Heart desease, work Type, Avg Glukose Level, BMI, Residence Status, hypertension* dan *stroke*.

2. Hasil Akurasi Algoritma C4.5

Pada ekperimen algoritma yang telah dilakukan maka didapatkan hasil untuk prediksi penyakit stroke sebesar 96.05%. hasil akurasi dapat dilihat pada Gambar 3.

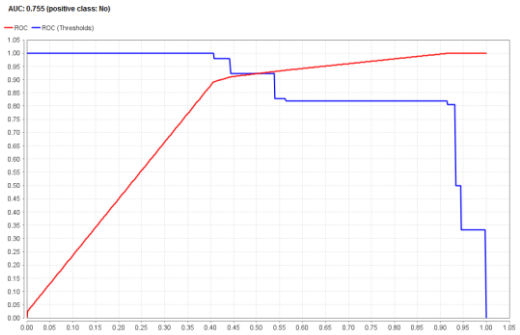
accuracy: 96.05%

	true Yes	true No	class precision
pred. Yes	14	4	77.78%
pred. No	151	3756	96.14%
class recall	8.48%	99.89%	

Gambar 3. Hasil Akurasi Klasifikasi Penyakit Stroke Menggunakan Algoritma C4.5

3. Hasil Visualisasi ROC *Curve* (AUC)

Hasil dari prediksi penyakit stroke juga dapat dilihat dalam kurva ROC (*Receiver Operating Characteristic*). Hasil dari kurva ROC dapat dilihat pada Gambar 4.



Gambar 4. Hasil Visualisasi ROC *Curve* (AUC)

Pada Gambar 4 dijelaskan bahwa nilai AUC:0,755 (*positive class* :No) garis merah merupakan ROC dan garis biru merupakan ROC (Thresholds).

4. KESIMPULAN

Dari penjelasan diatas ini dapat diambil kesimpulan penelitian dengan menggunakan *Decision tree* C4.5 berikut ini:

- Untuk mengetahui akurasi nilai *confusion matrix* dan nilai AUC yang dihasilkan.
- Atribut yang digunakan yaitu *gender, age, hypertension, heart desease, ever married, work type, avg glukose level, bmi, residence status, smoking status, stroke*.
- Dari 11 atribut yang terdapat diatas, kemudian atribut tersebut selanjutnya diseleksi mana atribut terpenting yang mempengaruhi penyakit stroke sehingga menjadi hanya 8 atribut yang

digunakan untuk prediksi penyakit stroke, atribut-atribut tersebut yaitu : *Age, Heart desease, work Type, Avg Glukose Level, BMI, Residence Status, hypertension* dan *stroke*.

- Algoritma *Decision Tree* C4.5, menghasilkan tingkat akurasi diagnosis penyakit stroke yang sangat baik.

5. SARAN

- Dataset *public* masih terdapat eror sehingga diharapkan nantinya akan diperoleh analisis yang lebih tepat.
- Klasifikasi penyakit stroke menggunakan algoritma *Decision Tree* C4.5 dalam mengimprove akurasi dapat mengubah parameter sampling linier disesuaikan dengan dataset.
- Data *missing value* tidak dapat dihindarkan maka untuk penelitian selanjutnya dapat dilakukan pengembangan metode yang lain untuk data *missing value* dan penentuan parameter tanpa trial and error yang diharapkan nantinya akan memberikan akurasi yang lebih tinggi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tim Redaksi Jurnal Teknika Politeknik Negeri Sriwijaya yang telah memberi memberi kesempatan, sehingga artikel ilmiah ini dapat diterbitkan.

DAFTAR PUSTAKA

[1] J. I. Hipertensi, *Pengenalan, Pencegahan, dan Pengobatan*. Jakarta: PT Bhuana Ilmu Populer, 2016.

[2] U. Widowati, "10 Penyakit Paling Mematikan Di Indonesia," *CNN Indones.*, pp. 1–5, 2015, [Online]. Available: http://www.cnnindonesia.com/gaya-hidup/20150513163407-255-53129/10-penyakit-paling-mematikan-di-indonesia/

[3] Jawapos.com, "Inilah Penyakit yang Paling Banyak Menyerang Masyarakat Indonesia," *Jawapos.com*, 2017, [Online]. Available: https://www.jawapos.com/kesehatan/21/11/2017/inilah-penyakit-yang-paling-banyak-menyerang-masyarakat-indonesia/

[4] A. Fauzi, B. Suharjo, and M. Syamsun, "Pengaruh Sumber Daya Finansial, Aset Tidak Berwujud dan Keunggulan Bersaing yang Berimplikasi Terhadap Kinerja Usaha Mikro, Kecil dan Menengah di Lombok NTB," *Manaj. IKM J. Manaj. Pengemb. Ind. Kecil Menengah*, vol. 11, no. 2, pp. 151–158, 2017, doi: 10.29244/mikm.11.2.151-158.

[5] A. Syaifullah, "Analisis pengaruh financial leverage dan operating leverage terhadap stock return," *Inovasi*, vol. 14, no. 2, p. 53, 2018, doi: 10.29264/jinv.v14i2.1928.

[6] E. Turban and dkk., *Decicion Support Systems and Intelegent Systems*. Andi Offset, 2005.

[7] P. Kasih, "Pemodelan Data Mining Decision Tree Dengan Classification Error Untuk Seleksi Calon Anggota Tim Paduan Suara," *Innov. Res. Informatics*, vol. 1, no. 2, pp. 63–69, 2019, doi: 10.37058/innovatics.v1i2.918.

[8] D. Nofriansyah and G. W. Nurcahyo, *Algoritma Data Mining Dan Pengujian*. Yogyakarta: Deepublish, 2019.