JURNAL DISPROTEK
ISSN: 2088-6500 (p); 2548-4168 (e)
Vol 15, No. 2, Juli 2024, hlm. 152-160
DOI: https://doi.org/10.34001/jdpt
https://ejournal.unisnu.ac.id/JDPT

OPTIMALISASI DIAGNOSIS STROKE DENGAN ALGORITMA C4.5 DAN STRATEGI IMPUTASI *k-NN* UNTUK MENGATASI *MISSING VALUE*

OPTIMIZATION OF STROKE DIAGNOSIS USING C4.5 ALGORITHM AND k-NN IMPUTATION STRATEGY TO OVERCOME MISSING VALUE

Zainal Abidin[1*], Teguh Tamrin[2], Vilanda Harsono[3], Duwi Nur Aziza[4], Istianti Kansania[5]

[1]Sekolah Tinggi Teknik Pati, Universitas Islam Nahdlatul Ulama Jepara[2], Sekolah Tinggi Teknik Pati[3], Universitas Gadjah Mada[4], Univ. Jederal Soederman[5]
Email : [1*]zainal.frsd@yahoo.co.id

Abstrak – Penyakit yang menyerang pembuluh darah didalam Otak(*Stroke*), mengakibatkan terhambatnya aliran darah dan oksigen ke otak. Gejala stroke bisa berbeda-beda, namu umumnya meliputi kelemahan atau mati rasa pada wajah, lengan, atau kaki, kebingungan, kesulitan berbicara, dan kesulitan berjalan. Diagnosa stroke yang cepat dan tepat sangatlah penting untuk mendapatkan penanganan yang tepat dan mencegah komplikasi. Salah satu faktor untuk mendiagnosis stroke dengan cepat dan akurat dengan menggunakan penerapan algoritma C4.5. Algoritma C4.6 yaitu algoritma klasifikasi yang efektif digunakan untuk membangun pohon keputusan dalam memprediksi. Penelitian ini menggunakan data dari Kaggle stroke prediksi dengan jumlah 15.000 record, 22 atribut dan 2500 data hilang . hasil penelitian menunjukan bahwa algoritma C4.5 dapat digunakan untuk membangun system diagnosis gejala penyakit stroke yang akurat. System ini mampu mengkategorikan pasien stroke dengan metode imputasi k-NN dengan nilai akurasi 91.40%. Pohon keputusan algoritma C4.5 juga dapat digunakan untuk memenuhi factor yang penting dalam diagnosis stroke.

Kata kunci: Stroke; Klasifikasi; Algoritma C4.5; K-NN

Abstract - Diseases that affect blood vessels in the brain (Stroke) result in the obstruction of blood flow and oxygen to the brain. Stroke symptoms can vary but generally include weakness or numbness in the face, arms, or legs, confusion, difficulty speaking, and difficulty walking. Rapid and accurate diagnosis of stroke is crucial to obtain appropriate treatment and prevent complications. One factor for diagnosing stroke quickly and accurately is the application of the C4.5 algorithm. The C4.5 algorithm is a classification algorithm effectively used to build decision trees for prediction. This study uses data from the Kaggle stroke prediction dataset with a total of 297520 records 22 atribut and 2500 missing value. The results of the study indicate that the C4.5 algorithm can be used to build an accurate stroke symptom diagnosis system. This system can categorize stroke patients with an accuracy rate of K-NN Imputation K-NN with classification values of accuration 91,40%. The C4.5 algorithm decision tree can also be used to fulfill important factors in stroke diagnosis.

***Keywords**: Strokel; Classification; Algorithm C4.5; K-NN*

1. PENDAHULUAN

Stroke merupakan penyakit yang menyerang pembuluh darah otak, menyebabkan gangguan aliran darah dan oksigen ke otak[1]. Gejala stroke dapat bervariasi, namun umumnya termasuk kelemahan atau mati rasa pada wajah, lengan, atau kaki, kebingungan, kesulitan berbicara, dan kesulitan berjalan. Diagnosis stroke yang cepat dan akurat sangat penting untuk mendapatkan pengobatan yang tepat dan mencegah komplikasi[2].Identifikasi dan pengenalan awal gejala stroke amatlah krusial untuk memperoleh penanganan medis yang tepat dan memperkecil kemungkinan komplikasi lanjutan.

Umumnya, diagnosis stroke dilakukan oleh tenaga medis melalui rangkaian pemeriksaan fisik dan pencitraan medis[3]. Akan tetapi, kemajuan teknologi informasi telah memungkinkan penerapan metode berbasis data dan algoritma kecerdasan buatan untuk membantu dalam diagnosis medis, termasuk stroke[4]. Deteksi dan diagnosis dini terhadap gejala stroke sangatlah penting untuk mendapatkan penanganan medis yang tepat waktu dan mengurangi risiko komplikasi lebih lanjut. Diagnosa yang cepat dan akurat dapat meningkatkan peluang pasien untuk pulih serta mengurangi tingkat keparahan kerusakan otak yang terjadi. Saat ini, diagnosis stroke umumnya

dilakukan oleh tenaga medis melalui serangkaian tes fisik dan pencitraan medis seperti CT Scan atau MRI. Namun, metode ini sering kali memerlukan waktu yang cukup lama dan ketersediaan peralatan yang canggih[5][6].

Namun, proses diagnosis stroke seringkali terhambat oleh kekurangan data (*missing value)* dalam rekam medis. Keberadaan *missing value* dapat mempengaruhi akurasi model klasifikasi yang digunakan untuk diagnosis stroke, sehingga menghasilkan prediksi yang tidak akurat dan membahayakan pasien.

Pemanfaatan metode berbasis data dan algoritma kecerdasan buatan (AI) dapat mendukung proses diagnosis medis, termasuk stroke. Algoritma pembelajaran mesin (machine learning) memungkinkan analisis data historis pasien untuk mengidentifikasi pola dan menghasilkan prediksi yang akurat. Salah satu algoritma yang efektif dalam klasifikasi dan prediksi adalah algoritma C4.5. Algoritma ini membangun pohon keputusan (decision tree) berdasarkan dataset yang tersedia. Kelebihan algoritma C4.5 terletak pada kemampuannya untuk menangani atribut kontinu dan diskrit serta mengatasi data yang hilang (missing values), menjadikannya sangat berguna dalam berbagai aplikasi, termasuk diagnosis penyakit. Berbagai sistem klasifikasi telah dikembangkan dalam literatur, seperti Radial Basis Function (RBF), Learning Vector Quantization (LVQ), C4.5, Classification 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), serta kombinasi FMM-CART-RF untuk mendukung diagnosis diabetes dan kanker payudara.

Para peneliti telah mengembangkan sistem klasifikasi untuk meningkatkan akurasi prediksi klasifikasi. Model yang diusulkan menggunakan K-means clustering untuk menganalisis teknik-teknik imputasi nilai yang hilang yang telah diteliti dan memilih metode imputasi terbaik[7]. Metode imputasi terbaik diterapkan pada dataset sebelum ekstraksi pola dan prediksi dilakukan. Salah satu metode klasifikasi yang digunakan adalah algoritma Pohon Keputusan (Decision Tree), yang telah diterapkan dalam berbagai bidang, seperti pengobatan[8], bidang bisnis [9] dan deteksi kegagalan [10]. Di bidang kesehatan contohnya penerapan *Decision Tree* untuk memprediksi pasien kanker payudara [11]. Algoritma C4.5 *Decision Tree* merupakan pengembangan dari metode *Iterative Dichotomiser 3* (ID3) yang dapat bekerja pada variabel kontinyu dan *missing value*.

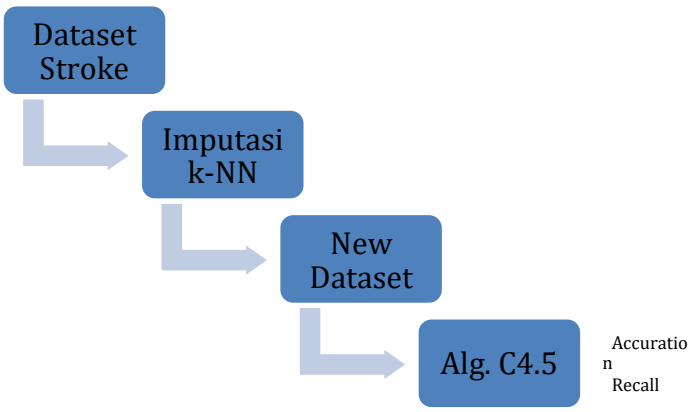
Untuk mengatasi masalah ini, dua pendekatan ditemukan dalam literatur. Pendekatan pertama terdiri dari teknik toleransi *missing value* yang mengintegrasikan teknik penanganan *missing value* dalam algoritma data mining tertentu seperti dalam klasifikasi [12], [13], clustering [14] dan seleksi fitur [15]. Pendekatan kedua terdiri dari teknik imputasi data dengan cara mengisi *missing value* sebelum menggunakan metode data yang lengkap. Salah satu keuntungan dari imputasi adalah bahwa penanganan *missing value* tidak bergantung pada pemilihan algoritma tetapi dapat memilih algoritma pembelajaran yang sesuai setelah imputasi [16].

Metode yang diusulkan pada penelitian ini untuk proses klasifikasi menggunakan algoritma C4.5. Sedangkan masalah *missing value* akan diatasi dengan melakukan *preprocessing* data menggunakan ukuran pemusatan data yaitu imputasi menggunakan algoritma *k*-NN.

Jadi pada penelitian ini akan menerapkan *preprocessing data* menggunakan metode imputasi pada *dataset* yang mengandung *missing value* terhadap algoritma C4.5 sebagai pengklasifikasi berdasarkan *dataset* yang terbentuk dari hasil *preprocessing data*. Oleh karena itu penelitian ini dilakukan untuk menerapkan strategi yang efektif untuk menangani *missing value* dalam data medis, dengan demikian akurasi diagnosis stroke dapat ditingkatkan dan kualitas layanan Kesehatan bagi pasien stroke dapat dijamin.

2. METODE PENELITIAN

Metode yang akan digunakan dalam penelitian ini dapat ditunjukan pada gambar dibawah ini:



Gambar 1. Kerangka Kerja Metode yang Diusulkan

1. **Pengumpulan Data**
   Dataset yang digunakan pada penelitian ini berasal dari Kagle yaitu *Stroke Prediction* yang dapat diakses di *https://www.kaggle.com/datasets/teamincribo/stroke-prediction*. *Dataset Stroke Prediction* mulai tersedia pada bulan April 2024 dan diambil oleh peneliti pada tanggal 14 Juni 2024. *Dataset* ini dipilih karena belum ada peneliti yang melakukan penelitian lanjut tentang Stroke dan *dataset* tersebut juga mengandung *missing value* yang mana dapat menjadi masalah pada saat dilakukan proses klasifikasi.
2. **Imputasi k-NN**
   Selain metode statistik yang digunakan untuk mengkompensasi nilai yang hilang pada bagian sebelumnya, metode komputasi menggunakan Machine Learning juga telah dikembangkan oleh para peneliti. Batista [21] melakukan pengujian dengan cara menguji akurasi klasifikasi dari dua pengklasifikasi populer (C4.5 dan CN2) dengan mempertimbangkan metode *k*-NN sebagai metode imputasi dan *Most Common* (MC). Baik algoritma C4.5 dan CN2 (seperti penelitian [22]) memiliki estimasi *missing value* tersendiri. Hasil penelitian menunjukkan bahwa tidak ada pengaruh negatif dari adanya proses imputasi bahkan metode imputasi *k*-NN lebih *robust* dibandingkan metode yang lain dengan peningkatan jumlah *missing value* di dalam *dataset*.
   Imputasi *k*-NN adalah metode imputasi *missing value* standar dan populer dalam menangani *missing value* [23]. Keunggulan dari metode imputasi *k*-NN 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 [21]. Kelemahan dari imputasi k-NN adalah ketika melakukan pengamatan untuk mencari nilai yang withering sesuai terhadap lost esteem, 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 lost esteem [24]. Tahapan algoritma imputasi *k*-NN dalam melakukan imputasi *missing value* adalah sebagai berikut:
   1. Menentukan jumlah *k*, yaitu jumlah observasi terdekat yang akan digunakan.
   2. Menghitung jarak antara observasi yang mengandung *missing value* pada variabel ke-*i* dengan observasi lainnya yang tidak mengandung *missing value* pada variabel yang bersesuaian dengan menggunakan formula:

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

      dengan:

      $d(X_a, X_b)$ merupakan jarak antara target observasi X_a dan observasi X_b

      X_{ai} adalah nilai variabel ke-i pada target observasi X_a, $i = 1, 2, \ldots, m$

      X_{bi} adalah nilai variabel ke-i pada target observasi lainnya X_b, $i = 1,2, \ldots, m$

   3. Mencari *k* observasi terdekat berdasarkan nilai jarak terkecil. Nilai variabel pada *k* observasi terdekat ini akan digunakan untuk proses imputasi pada observasi yang mengandung *missing value*.
   4. Menghitung bobot (*weight*) pada setiap *k* observasi terdekat. Observasi yang paling dekat akan mendapatkan bobot yang paling besar.
   5. Menghitung nilai rata-rata pada *k* observasi terdekat yang tidak mengandung *missing value* dengan prosedur *weighted mean estimation* yaitu dengan formula

      $\widehat{X}_i = \frac{1}{KW} \sum_{K=1}^{K} W_k V_{ki}$

      Dengan:
      V_{kj} adalah nilai variabel ke-*i* pada observasi ke-*k*, $k = 1, 2, \ldots, K$
      $W = \sum_{K=1}^{K} W_k$

      W_k adalah bobot observasi tetangga terdekat ke-*k*, yang dirumuskan sebagai berikut:
      $W_k = \frac{1}{d(X,V_k)^2}$
   6. Melakukan proses imputasi *missing value* pada nilai-nilai yang mengandung *missing value* dengan nilai rata-rata yang diperoleh pada tahap 5.
3. **Data Baru (New Dataset)**
   Setelah melakukan imputasi k-NN, data yang hilang menghasilkan data baru yang siap untuk dianalisis. Data baru ini lebih lengkap dan informatif, sehingga memungkinkan analisis yang lebih akurat dan

terpercaya. Dengan menggunakan data yang lengkap, peneliti dapat memperoleh pemahaman yang lebih baik tentang fenomena yang diteliti dan membuat kesimpulan yang lebih valid.

4. *Algoritma C4.5*

Decision Tree merupakan algoritma pengklasifikasian yang mempunyai struktur yang sederhana dan mudah untuk di interpretasikan [25]. Pohon Keputusan (Decision Tree) menyerupai struktur pohon yang digunakan sebagai metode penalaran untuk menemukan solusi dari masalah yang diberikan. Struktur pohon yang terbentuk tidak selalu berupa pohon biner. Jika semua fitur dalam dataset hanya memiliki dua nilai kategorikal, maka pohon yang terbentuk akan berbentuk pohon biner. Namun, jika fitur-fitur tersebut memiliki lebih dari dua nilai kategorikal atau menggunakan tipe numerik, maka pohon yang terbentuk biasanya tidak berbentuk pohon biner.Kefleksibelan membuat metode ini atraktif, khususnya karena memberikan keuntungan berupa visualisasi saran (dalam bentuk **choice** tree) yang membuat prosedur prediksinya dapat diamati [26]. Decision Tree banyak digunakan untuk menyelesaikan kasus penentuan keputusan seperti di bidang kedokteran (diagnosis penyakit pasien) [8], [11], bidang bisnis [9], ilmu komputer (struktur data), psikologi (teori pengambilan keputusan), dan sebagainya.

Karakteristik dari *Decision Tree* membentuk sejumlah elemen sebagai berikut [27]: 1) *Node* Akar, tidak mempunyai lengan masukan dan mempunyai nol atau lebih lengan keluaran. 2) Node internal, setiap node yang bukan daun (nonterminal) yang mempunyai tepat satu lengan masukan dan dua atau lebih lengan keluaran. 3) Lengan, setiap cabang menyatakan nilai hasil pengujian di node bukan daun. 4) Node daun (terminal), node yang mempunyai tepat satu lengan masukan dan tidak mempunyai lengan keluaran. *Node* ini menyatakan label kelas (keputusan).

Keuntungan dari penggunaan *Decision Tre*e [26] adalah sebagai berikut:

1. Mudah dipahami dan diinterpretasikan.
2. Dalam pembentukan pohon biayanya murah.
3. Tidak ada batasan untuk data numerik dan kategorikal.
4. Logika yang mendasari proses pengambilan keputusan dapat diikuti dengan mudah dan aturan pengklasifikasiannya sejak awal dapat dipahami.
5. Menggunakan teknik statistik klasik dalam membuat model validasi.
6. Kuat, cepat dan dapat memproses dengan baik *dataset* yang besar.
7. Mempunyai akurasi sebanding dengan teknik pengklasifikasian untuk *dataset* yang simple.

Beberapa algoritma yang telah dikembangkan berdasarkan *Decision Tree* antara lain CHAID (Chi-squares Automatic Interaction Detection), CART (Classification and Regression Tree), C4.5 merupakan variasi pengembangan dari ID3 (Iterative Dichotomiser 3) [26]. Dalam ID3, induksi *Decision Tree* hanya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal), sedangkan tipe numerik (interval dan rasio) tidak dapat digunakan. Perbaikan yang membedakan algoritma C4.5 dari ID3 adalah dapat menangani fitur tipe numerik, melakukan pemotongan (pruning) *Decision Tree*, penurunan (deriving) *rule set* dan dapat mengatasi masalah *dataset* yang mengandung *Missing Value* [28].

Konsep yang digunakan dalam algoritma C4.5 dalam penentuan *split* yang optimal yaitu menggunakan *information gain* dan *entropy reduction* [29]. *Split* yang terpilih adalah yang mempunyai nilai *gain* dan *information gain* yang terbesar.

Pada algoritma C4.5, tahapan dalam membentuk pohon keputusan (Decision Tree) adalah:

1. Menghitung nilai *entropy.*

Untuk menghitung nilai entropy digunakan formula sebagai berikut:

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

dimana:

S = Himpunan Kasus

n = Jumlah partisi *S*

pi = Proporsi dari Si terhadap *S*

2. Menghitung nilai *gain ratio* pada masing-masing atribut.

Untuk menghitung nilai *gain ratio* digunakan formula sebagai berikut:

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

Dimana:

S = Himpunan Kasus

A = Atribut

n = Jumlah partisi atribut *A*

$|S_i|$ = Jumlah kasus pada partisi ke-*i*

$|S|$ = Jumlah kasus dalam *S*

3. Atribut dengan *gain ratio* tertinggi dipilih menjadi akar (root), sedangkan atribut dengan *gain ratio* terendah akan dipilih menjadi cabang (branches)
4. Menghitung lagi nilai *gain ratio* pada tiap-tiap atribut tetapi tidak mengikutsertakan atribut yang telah menjadi akar (root) pada tahap sebelumnya
5. Atribut dengan *gain ratio* tertinggi dipilih menjadi cabang (branches)
6. Mengulang langkah ke-4 dan ke-5 sampai pada semua atribut yang tersisa nilai *gain* = 0.

Sehingga akan dihasilkan pohon (tree) yang terdiri dari akar (root), cabang (branches) dan daun (leaf) yang ditentukan melalui proses *split* atribut dan juga menghasilkan nilai A*ccuration, Recall dan Precision*.

3. HASIL DAN PEMBAHASAN

3.1. PENGUMPULAN DATA

Data yang digunakan dari Kagle yaitu *Stroke Prediction* yang dapat diakses di *https://www.kaggle.com/datasets/teamincribo/stroke-prediction*. *Dataset Stroke Prediction* mulai tersedia pada bulan April 2024 dan diambil oleh peneliti pada tanggal 14 Juni 2024. Yaitu jumlah data record 15.00, atribut 22 dan 2500 missing value. Data yang digunakan dapat ditunjukan pada table berikut:

Tabel 1 *Dataset* yang digunakan dalam eksperimen

No	*Age*	*Gender*	*Hypertension*	*Heart Disease*	*Marital Status*	…	*Cholesterol Levels*	*Symptoms*	*Diagnosis*
1	*56*	*Male*	*0*	*1*	*Married*	*...*	*HDL: 68, LDL: 133*	*Difficulty Speaking, Headache*	*Stroke*
2	*80*	*Male*	*0*	*0*	*Single*	*...*	*HDL: 63, LDL: 70*	*Loss of Balance, Headache, Dizziness, Confusion*	*Stroke*
3	*26*	*Male*	*1*	*1*	*Married*	*...*	*HDL: 59, LDL: 95*	*Seizures, Dizziness*	*Stroke*
4	*73*	*Male*	*0*	*0*	*Married*	*...*	*HDL: 70, LDL: 137*	*Seizures, Blurred Vision, Severe Fatigue, Headache, Confusion*	*No Stroke*
5	*51*	*Male*	*1*	*1*	*Divorced*	*...*	*HDL: 65, LDL: 68*	*Difficulty Speaking*	*Stroke*
6	*62*	*Female*	*0*	*0*	*Single*	*...*	*HDL: 80, LDL: 69*	*Severe Fatigue*	*Stroke*
...	*...*	*...*	*...*	*...*	*...*	*...*	*...*	*...*	*...*
14.999	*73*	*Male*	*0*	*0*	*Single*	*...*	*HDL: 79, LDL: 91*	*Severe Fatigue, Numbness, Confusion, Dizziness, Loss of Balance*	*No Stroke*
15.000	*64*	*Female*	*0*	*0*	*Single*	*...*	*HDL: 78, LDL: 179*	*Headache*	*Stroke*

3.2. Imputasi K-NN

Dengan menggunakan imputasi K-NN yaitu mengganti nilai missing value dengan menggunakan perkiraan nilai terdekat pada data set yang tidak mengandung missing value. Adapun tahapan imputasi K-NN adalah sebagai berikut:

1. Menentukan nilai *k*, yaitu jumlah observasi terdekat yang akan digunakan. Nilai *k* yang digunakan pada penelitian ini adalah *k* = 3.
2. Menghitung jarak antara observasi yang mengandung *missing value* dengan observasi lain yang tidak mengandung *missing value* pada variabel

Tabel 2 *Missing* data

Record	*Age*	Average Glucose Level	Body Mass Index (BMI)	*Smoking*	*Alcohol*
101	48	70	1.005	4	0
102	24	?	1.015	2	4
103	52	100	1.015	3	0
104	62	80	1.010	2	3

Menghitung jarak antara observasi yang mengandung *missing value* dengan observasi lain yang tidak mengandung *missing value* pada variabel yang bersesuaian menggunakan:

$d(1,2) = \sqrt{((48-24)^2) + (70-75)^2 + (1.005-1.015)^2 + (4-2)^2 + (0-4)^2}$
$= 24.91987$

$d(3,2) = \sqrt{((52-24)^2) + (100-75)^2 + (1.015-1.015)^2 + (3-2)^2 + (0-4)^2}$
$= 37.74917$

$d(3,2) = \sqrt{((62-24)^2) + (80-75)^2 + (1.010-1.015)^2 + (2-2)^2 + (3-4)^2}$
$= 38.34058$

Tabel 3 Jarak observasi *k-Nearest Neighbor*

Record	*Age*	Average Glucose Level	Body Mass Index (BMI)	*Smoking*	*Alcohol*	*Distance*
101	48	70	1.005	4	0	24.91987
102	24	?	1.015	2	4	-
103	52	100	1.015	3	0	37.74917
104	62	80	1.010	2	3	38.34058

3. Didapatkan k observasi terdekat 24.91987
4. Menghitung bobot pada setiap k observasi

$W(1, 2) = \frac{1}{24.91987^2} = 0.00161$

$W(3, 2) = \frac{1}{37.74917^2} = 0.00070$

$W(4, 2) = \frac{1}{38.34058^2} = 0.00068$

Tabel 4 Pembobotan pada *k* observasi terdekat

Record	*Age*	Average Glucose Level	Body Mass Index (BMI)	*Smoking*	*Alcohol*	*Distance*	*Weight*
101	48	70	1.005	4	0	24.91987	0.00161
102	24	?	1.015	2	4	-	-
103	52	100	1.015	3	0	37.74917	0.00070
104	62	80	1.010	2	3	38.34058	0.0068

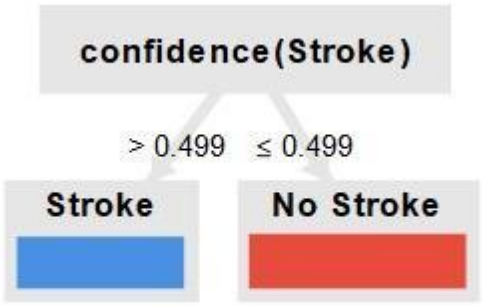
5. Menghitung nilai rata-rata dari k observasi

$$\underline{X} = \frac{\frac{70}{24.91987^2} + \frac{100}{37.74917^2} + \frac{80}{38.34058^2}}{\frac{1}{24.91987^2} + \frac{1}{37.74917^2} + \frac{1}{38.34058^2}} = 79.30891 \cong 80.$$

Sehingga didapatkan nilai imputasi untuk mengisi *missing value* sebesar 80.

6. Melakukan proses imputasi

Setelah melakukan perhitungan, nilai k-NN didapatkan pohon keputusan (decision tree) seperti gambar di bawah ini:



Gambar 2. pohon keputusan data set Stroke dengan imputasi k-NN

Gambar 2 menunjukkan model pohon keputusan (Decision Tree) yang tebentuk pada pada *dataset* dengan metode imputasi nilai *k*-NN. Bentuk aturan *IF THEN* untuk *Decision Tree* adalah sebagai berikut:

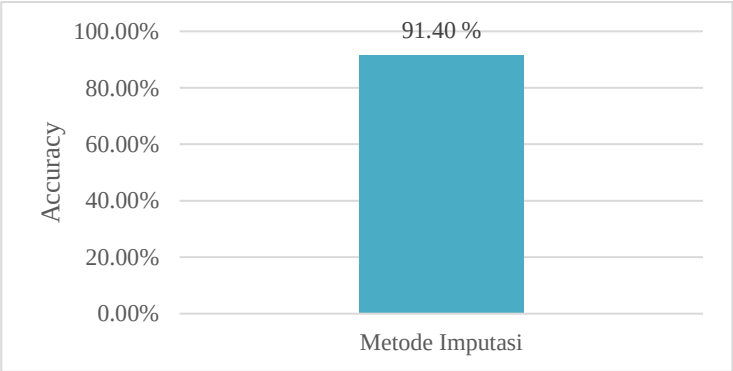
IF Confidence <=12.85 *THEN* stroke
IF Confidence > 12.85 *AND Specific Gravity* <= 0.499*THEN* notstroke
IF Confidence > 12.85 *AND Specific Gravity* > 0.499 *AND Appetite* = *THEN* stroke

Setelah semua tahapan pembentukan pohon keputusan pada metode dilakukan, kemudian dicatat hasil pengukuran metode berdasarkan hasil *accuracy, precision and recall*.

Tabel 5 hasil pengukuran pengklasifikasian C4.5 pada dataset Stroke

Algoritma	***Accuracy***	***Precission***		***Recall***	
		Stroke	***No Stroke***	***Stroke***	***No Stroke***
C4.5	91.40 %	90.53 %	92,22%	91,67%	91,15%

Diagram akurasi Gambar 3 menunjukkan bahwa metode imputasi dengan menunjukkan akurasi yang lebih baik dibandingkan dengan metode imputasi *k-NN*.



Gambar 3. Diagram Algoritma C4.5

Gambar 4. Kurva AUC imputasi dengan nilai *k*-NN

Pada penelitian terkait *missing value* oleh Farhangfar et al [19] menyimpulkan bahwa metode imputasi dengan nilai *k-NN* menjadi metode yang paling menguntungkan dibanding metode lain, sedangkan salah satu kesimpulan dari penelitian Song et al [20] adalah kinerja dari C4.5 dapat dipengaruhi oleh mekanisme, pola dan persentase *missing value*. Sedangkan pada penelitian ini menggunakan metode imputasi *missing value* dengan *k-NN* terbukti dapat meningkatkan kinerja dari pengklasifikasi C4.5.

4. KESIMPULAN

Dalam penelitian ini dilakukan pengujian metode pendekatan penanganan *missing value* pada *preprocessing data* terhadap algoritma C4.5 menggunakan teknik imputasi dengan *k*-NN. Teknik imputasi yang dilakukan berguna untuk mengisi/mengganti *missing value* dengan nilai-nilai tertentu sehingga didapatkan *dataset* yang lengkap tanpa *missing value* yang kemudian dilakukan proses klasifikasi pada tahapan berikutnya. Namun, perlu diperhatikan bahwa model C4.5 cenderung berpotensi *overfitting* pada data pelatihan. Ini berarti model mungkin menjadi terlalu kompleks dan tidak umum jika digunakan pada data yang tidak terlihat sebelumnya. Namun, hasil pengujian menunjukkan bahwa dalam penanganan *missing value* diperoleh nilai *missing value* dengan teknik imputasi k-NN yang dapat meningkatkan dan memberikan nilai kinerja dari pengklasifikasi C4.5 secara signifikan.

Penelitian ini telah memberikan kontribusi untuk penanganan *missing value* pada *preprocessing data* terhadap algoritma C4.5. Hasil dari penelitian ini menunjukkan bahwa metode imputasi terhadap *missing value* dapat mempengaruhi dan meningkatkan kinerja dari pengklasifikasi C4.5 sehingga disarankan untuk penelitian selanjutnya dapat mencoba menggunakan teknik *feature selection* dan teknik imputasi yang lain dengan pendekatan statistik, misalnya imputasi dengan nilai standar deviasi dan simpangan rata-rata maupun metode imputasi dengan *machine learning* misalnya *Weighted k-Nearest Neighbor* (W*k*-NN) dan *Clustering k-Nearest Neighbor* (C*k*-NN).

DAFTAR PUSTAKA

[1] S. R. Laily, "Hubungan Karakteristik Penderita dan Hipertensi dengan Kejadian Stroke Iskemik Relationship Between Characteristic and Hypertension With Incidence of Ischemic Stroke," *Berkali Epidemiol.*, vol. 5, no. February, pp. 48–59, 2018, doi: 10.20473/jbe.v5i1.

[2] R. S. Rohman, R. A. Saputra, and D. A. Firmansaha, "Komparasi Algoritma C4.5 Berbasis PSO Dan GA Untuk Diagnosa Penyakit Stroke," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 1, p. 155, 2020, doi: 10.24114/cess.v5i1.15225.

[3] N. Gusrialni Fitri, S. Adilya, and F. Azizi, "Comparison of the Naive Bayes Classification System and C4.5 for the Diagnosis of Stroke Perbandingan Sistem Klasifikasi Naive Bayes dan C4.5 Untuk Diagnosa Penyakit stoke," *SENTIMAS Semin. Nas. Penelit. dan Pengabdi. Masy.*, pp. 49–55, 2023.

[4] Suryani, D. Rahmadani, A. A. Muzafar, A. Hamid, R. Annisa, and Mustakim, "Analisis Perbandingan Algoritma C4.5 dan CART untuk Klasifikasi Penyakit Stroke," *SENTIMAS Semin. Nas. Penelit. dan Pengabdi. Masy.*, pp. 197–206, 2022, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas

[5] Prita Prita, I Made Lana Prasetya, and Rahmat Widodo, "Prosedur Pemeriksaan MRI Brain Pada Kasus Stroke Hemoragik," *J. Ris. Rumpun Ilmu Kedokt.*, vol. 2, no. 2, pp. 82–91, 2023, doi: 10.55606/jurrike.v2i2.1859.

[6] P. K. Kognisi *et al.*, "No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title," *Ind. High. Educ.*, vol. 3, no. 1, pp. 1689–1699, 2021, [Online]. Available: http://journal.unilak.ac.id/index.php/JIEB/article/view/3845%0Ahttp://dspace.uc.ac.id/handle/123456789/1288

[7] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5621–5631, 2015, doi: 10.1016/j.eswa.2015.02.050.

[8] D. Setsirichok *et al.*, "Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for thalassaemia screening," *Biomed. Signal Process. Control*, vol. 7, no. 2, pp. 202–212, 2012, doi: 10.1016/j.bspc.2011.03.007.

[9] P. Duchessi and E. J. M. Lauría, "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, 2013, doi: 10.1016/j.eswa.2013.05.017.

[10] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5916–5923, 2013, doi: 10.1016/j.eswa.2013.05.021.

[11] M. Ture, F. Tokatli, and I. Kurt, "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.*, vol. 36, no. 2 PART 1, pp. 2017–2026, 2009, doi: 10.1016/j.eswa.2007.12.002.

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

[13] M. Saar-Tsechansky and F. Provost, "Handling Missing Values when Applying Classification Models," *J. Mach. Learn. Res.*, vol. 8, pp. 1625–1657, 2007, doi: 10.1.1.72.3271.

[14] R. J. Hathaway and J. C. Bezdek, "Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm," *Pattern Recognit. Lett.*, vol. 23, no. 1–3, pp. 151–160, 2002, doi: 10.1016/S0167-8655(01)00115-5.

[15] A. Aussem and S. Rodrigues de Morais, "A conservative feature subset selection algorithm with missing data," *Neurocomputing*, vol. 73, no. 4–6, pp. 585–590, 2010, doi: 10.1016/j.neucom.2009.05.019.

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

[17] E. R. Hruschka, E. R. Hruschka, and N. F. F. Ebecken, "Bayesian networks for imputation in classification problems," *J. Intell. Inf. Syst.*, vol. 29, no. 3, pp. 231–252, 2007, doi: 10.1007/s10844-006-0016-x.

[18] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," vol. 41, pp. 3692–3705, 2008, doi: 10.1016/j.patcog.2008.05.019.

[19] Q. Song, M. Shepperd, X. Chen, and J. Liu, "Can k -NN Imputation Improve the Performance of C4 . 5 With Small Software Project Data Sets ? A Comparative Evaluation," pp. 1–31, 2008.

[20] B. Twala, "An Empirical Comparison of Techniques for Handling Incomplete Data Using Decision Trees," *Model. Digit. Intel.*, no. Ml, pp. 1–35, 1998.

[21] G. Batista and M. C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method," *Hybrid Intell. Syst.*, vol. 87, no. 48, pp. 251–260, 2002.

[22] J. W. Grzymala-Busse, "A comparison of traditional and rough set approaches to missing attribute values in data mining," in *WIT Transactions on Information and Communication Technologies*, May 2009, pp. 155–163. doi: 10.2495/DATA090161.

[23] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001, doi: 10.1093/bioinformatics/17.6.520.

[24] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2009, doi: 10.1007/s00521-009-0295-6.

[25] C. J. Mantas and J. Abellán, "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4625–4637, 2014, doi: 10.1016/j.eswa.2014.01.017.

[26] F. Gorunescu, *Data Mining Concept, Model and Techniques*, vol. 12. in Intelligent Systems Reference Library, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19721-5.

[27] T. Pang-Ning, M. Steinbach, and V. Kumar, "Introduction to data mining," *Libr. Congr.*, p. 796, 2006, doi: 10.1016/0022-4405(81)90007-8.

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

[29] D. T. Larose, *Data Mining Methods and Models*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005. doi: 10.1002/0471756482.