

ALGORITMA *DECISION TREE C4.5* DIGUNAKAN UNTUK MENGLASIFIKASIKAN DATA STROKE

Cecep Maulana Sidiq¹, Ahmad Faqih², Gifthera Dwilestari³

^{1,2}Teknik Informatika, STMIK IKMI Cirebon

³Sistem Informasi, STMIK IKMI Cirebon

Jalan Perjuangan No. 10 B Cirebon, Indonesia

18cecepmaulana@gmail.com

ABSTRAK

Tujuan utama dari penelitian ini adalah untuk menggunakan algoritma pohon keputusan atau Decision Tree C4.5 untuk mengkategorikan data stroke. Upaya penelitian khusus ini dimotivasi oleh masalah mendesak dari tingkat kejadian yang mengkhawatirkan terkait dengan stroke, akibat perlu prediksi yang tepat dari faktor risiko stroke. Dataset yang digunakan untuk penelitian ini diperoleh dari situs web data publik dan platform terkenal www.kaggle.com, yang mencakup total 11 variabel. Metodologi penelitian yang digunakan dalam penyelidikan ini mencakup beberapa langkah integral, termasuk pengumpulan data, pra-pemrosesan data, transformasi data, pemodelan penambangan data dengan algoritma Pohon Keputusan C4.5, dan, terakhir, evaluasi model. Temuan fase evaluasi mengungkapkan tingkat akurasi yang mengesankan 93,64%, dengan nilai presisi 12,50% untuk prediksi stroke positif dan 95,26% untuk prediksi stroke negatif. Namun, penting untuk menyoroti bahwa meskipun tingkat akurasi tinggi dicapai, nilai presisi dan penarikan untuk kasus stroke positif masih relatif rendah, sehingga memerlukan peningkatan dan penyempurnaan model. Secara keseluruhan, analisis komprehensif dan hasil penelitian ini sangat menyiratkan bahwa algoritma Pohon Keputusan C4.5 menunjukkan potensi yang signifikan untuk klasifikasi awal penyakit stroke.

Kata Kunci: *Stroke, Decision tree, Data Stroke-*

1. PENDAHULUAN

Stroke adalah kondisi medis yang terjadi ketika aliran darah ke otak terhenti, biasanya akibat pembuluh darah yang pecah atau tersumbat. Kondisi ini dapat menyebabkan kerusakan permanen pada sel-sel otak dan berdampak serius pada kesehatan dan kualitas hidup seseorang. Di seluruh dunia, jumlah penderita stroke terus meningkat, menjadikannya salah satu penyebab utama kematian dan kecacatan [1]. Faktor risiko stroke dapat dibagi menjadi dua kategori: faktor yang tidak dapat diubah (seperti usia, jenis kelamin, riwayat keluarga, dan ras) dan faktor yang dapat dimodifikasi (seperti hipertensi, obesitas, merokok, diabetes melitus, dan aktivitas fisik). Oleh karena itu, penting untuk mengidentifikasi dan memahami faktor-faktor ini untuk mencegah dan mengendalikan stroke [2].

Keuntungan mengklasifikasikan data stroke adalah dapat memberikan informasi berharga untuk mengembangkan strategi pencegahan dan pengobatan yang lebih efektif. Dengan mengidentifikasi faktor-faktor risiko penting, kita dapat mengambil langkah-langkah untuk mengurangi risiko stroke, seperti menerapkan gaya hidup sehat, mengendalikan tekanan darah dan kolesterol, serta mengelola kondisi lainnya. Obat-obatan dapat meningkatkan risiko stroke dan penyakit serebrovaskular. Selain itu, dengan mempelajari pola dan karakteristik yang terkait dengan stroke, kita dapat meningkatkan pemahaman kita tentang penyakit ini dan mengembangkan metode baru untuk deteksi dini dan intervensi yang lebih efektif.

Tujuan dari penelitian ini adalah untuk melihat perbedaan rata-rata umur pasien stroke berdasarkan jenis kelamin dalam populasi tertentu. Stroke adalah salah satu penyakit vaskular yang paling banyak memengaruhi morbiditas dan mortalitas di seluruh dunia. Pemahaman yang lebih baik tentang faktor-faktor yang memengaruhi stroke, termasuk perbedaan jenis kelamin pasien, dapat membantu dalam mengembangkan metode pencegahan, diagnosis, dan perawatan stroke yang lebih baik [3].

Hasil penelitian ini menegaskan bahwa klasifikasi data stroke dengan memanfaatkan analisis statistik dan pembelajaran mesin merupakan pendekatan yang efektif dalam mengidentifikasi faktor risiko utama dan mengembangkan strategi pencegahan serta pengobatan stroke yang lebih efektif. Dengan demikian, penelitian ini memberikan kontribusi signifikan terhadap upaya global dalam mengurangi beban penyakit stroke, melalui peningkatan pemahaman tentang faktor risiko dan penerapan teknologi dalam deteksi serta intervensi dini.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Studi Yohana Tri Dkk. (2021) berfokus pada penurunan jumlah pelanggan setia perusahaan MNC Play, penyedia layanan internet di Kota Semarang. Untuk membedakan pelanggan yang setia dan tidak setia, penelitian ini menggunakan metode klasifikasi data mining. Hasil penelitian menunjukkan bahwa algoritma C4.5 lebih baik daripada Naïve Bayes dalam menentukan tingkat loyalitas pelanggan di dataset

MNC Play Semarang. Algoritma C4.5 memiliki akurasi 85,43% dan AUC 0,859, sedangkan Naïve Bayes memiliki akurasi 77,39% dan AUC 0,50. Oleh karena itu, algoritma C4.5 dapat digunakan sebagai referensi untuk menentukan apakah pelanggan loyal atau tidak loyal pada dataset MNC Play Semarang. [4].

Penelitian Irma Handayani berfokus pada klasifikasi penyakit *Disk Hernia* dan *Spondylolisthesis Kolumna Vertebralis*. Selain itu, penelitian ini menghadapi tantangan untuk menentukan apakah pasien diklasifikasikan sebagai Normal atau Abnormal berdasarkan nilai karakteristik yang dimiliki masing-masing pasien. Hasil penelitian menunjukkan bahwa metode algoritma C4.5 dapat digunakan untuk mengklasifikasikan *Spondylolisthesis* dan penyakit *Disk Hernia*. Selain itu, penelitian ini membagi kelas abnormal menjadi dua subkelas: kelas *Disk Hernia* dan kelas *Spondylolisthesis* berdasarkan nilai fitur seperti incidence pelvic, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, dan tingkat spondylolisthesis [5].

Penelitian yang dilakukan oleh Agus Byna dan Muhammad Basit berfokus pada peningkatan tingkat kejadian stroke di Kota Banjarmasin, yang membutuhkan prediksi yang akurat tentang penyakit tersebut. Untuk memprediksi stroke, metode penelitian menggunakan algoritma naive bayes dengan optimasi adaboost. Hasil studi menunjukkan Naive Bayes tidak dioptimalkan untuk mencapai akurasi 89,65% dalam proses mengklasifikasikan data. Akurasi meningkat ketika ada optimasi menggunakan algoritma adaboost. Ini ditunjukkan oleh penelitian sebelumnya yang menunjukkan peningkatan akurasi dari 78,88% menjadi 82,58% dengan Particle Swarm Optimization. Namun, hasil penelitian juga menunjukkan bahwa peningkatan ini tidak selalu terjadi, seperti yang terjadi dengan peningkatan akurasi dari 68,50% menjadi 71,29% dengan Particle Swarm Optimization. Akibatnya, penelitian ini menunjukkan bahwa algoritma naive bayes dapat dioptimasi untuk meningkatkan akurasi prediksi penyakit stroke dengan menggunakan algoritma adaboost. [6].

2.2. Stroke

Stroke adalah kondisi kesehatan yang terjadi ketika suplai darah ke otak terganggu atau terhenti, menyebabkan kerusakan pada sel-sel otak. Hal ini dapat menyebabkan berbagai gejala, seperti kelemahan otot, kesulitan berbicara, dan masalah kognitif. Deteksi dini dan klasifikasi stroke dapat membantu dalam pengelolaan dan perawatan yang tepat waktu. 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. Penyakit Stroke merupakan jenis penyakit yang mematikan dimana masuk kedalam 10 di antara jenis penyakit yang paling berbahaya 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. Stroke merupakan salah satu penyakit yang paling banyak diderita oleh masyarakat Indonesia dan menjadi urutan pertama penyebab kematian tertinggi disusul oleh diabetes dan hipertens [7].

2.3. Klasifikasi

Klasifikasi adalah teknik mendalam dan rumit yang digunakan untuk membangun model data yang tidak diklasifikasikan yang komprehensif dan rumit, sehingga memberdayakan penerapan model tersebut untuk mengklasifikasikan data baru dan yang belum dijelajahi. Dalam bidang perawatan kesehatan, klasifikasi berfungsi sebagai alat yang tak ternilai untuk mengelompokkan pasien secara efektif, sehingga memfasilitasi pemeriksaan yang cermat terhadap faktor risiko dan gejala penyakit masing-masing. Sejumlah penyelidikan sebelumnya telah berhasil menggunakan klasifikasi sebagai sarana untuk secara akurat memprediksi kejadian dan manifestasi penyakit stroke, terutama melalui analisis berbagai faktor risiko, termasuk tetapi tidak terbatas pada usia, jenis kelamin, tekanan darah, diabetes mellitus, dislipidemia, kadar asam urat, dan penyakit jantung. [8][9]

2.4. Decision Tree

Decision Tree adalah metode klasifikasi dalam machine learning dan data mining yang menggunakan struktur pohon atau model hirarkis untuk mengambil keputusan. Struktur pohon ini terdiri dari node, cabang, dan daun. Node mewakili fitur atau atribut, cabang mewakili aturan keputusan, dan daun mewakili hasil keputusan. Decision Tree bekerja dengan membagi dataset menjadi subset yang lebih kecil berdasarkan atribut [10].

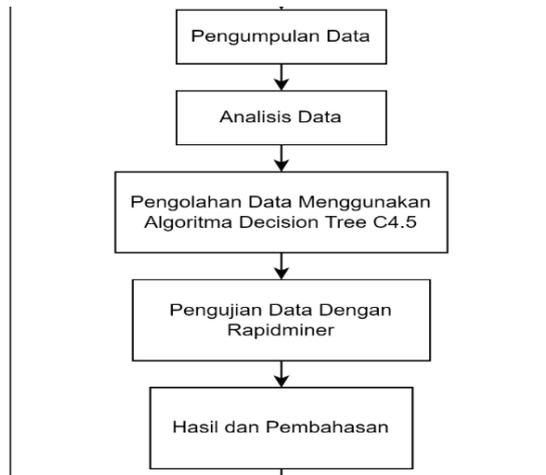
2.5. Algoritma C4.5

Algoritma C4.5 adalah pengembangan dari algoritma ID3 yang dibuat oleh Ross Quinlan. Algoritma ini digunakan untuk menghasilkan Decision Tree dengan membagi data berdasarkan atribut yang memiliki nilai informasi gain tertinggi. Informasi gain dihitung berdasarkan konsep entropi dalam teori informasi. Algoritma C4.5 memiliki beberapa peningkatan dibandingkan ID3, termasuk penanganan atribut dengan nilai kontinu, penanganan data yang hilang, dan pemangkasan pohon (pruning) untuk menghindari overfitting. Overfitting adalah kondisi di mana model terlalu fit dengan data latih sehingga performanya menurun saat diuji dengan data yang belum pernah dilihat sebelumnya. Dengan fitur-fitur ini, algoritma C4.5 menjadi salah satu algoritma pembuatan Decision Tree yang paling populer dan

banyak digunakan dalam berbagai aplikasi klasifikasi [11].

3. METODE PENELITIAN

Desain penelitian merupakan tahapan atau proses yang digunakan untuk merencanakan suatu penelitian. Desain penelitian ini bertujuan untuk mempermudah peneliti dalam melakukan penelitian, tahapan pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Metodologi Penelitian

3.1. Pengumpulan data

Pengumpulan data adalah tahap awal dalam penelitian ini.

3.2. Pra-pemrosesan

Pra-pemrosesan data adalah tahap krusial yang menyiapkan data untuk analisis lebih lanjut. Meskipun data yang dikumpulkan sudah bersih dari noise dan tidak valid, masih perlu dilakukan beberapa langkah pra-pemrosesan. Langkah ini termasuk pembagian data menjadi set pelatihan dan pengujian, yang akan digunakan untuk melatih model dan kemudian menguji kinerjanya. Pembagian ini biasanya dilakukan secara acak dan proporsional untuk memastikan bahwa kedua set data mewakili keseluruhan dataset. Selain itu, normalisasi atau standarisasi data mungkin diperlukan untuk memastikan bahwa semua fitur memiliki skala yang serupa, yang penting untuk beberapa algoritma pembelajaran mesin.

3.3. Transformasi

Transformasi data melibatkan perubahan data ke dalam format yang lebih sesuai untuk analisis. Ini bisa termasuk teknik seperti pengkodean variabel kategorikal menjadi numerik, pengurangan dimensi untuk menghilangkan redundansi dan mengurangi kompleksitas, serta transformasi lain yang dapat membantu dalam mengungkap pola yang lebih jelas dalam data. Transformasi ini penting untuk meningkatkan efektivitas algoritma data mining yang akan digunakan pada tahap berikutnya.

3.4. Modeling Data Mining

Modeling data mining adalah proses di mana algoritma pembelajaran mesin diterapkan pada data yang telah diproses untuk mengidentifikasi pola dan membuat prediksi. Pada tahap ini, berbagai model dan algoritma seperti pohon keputusan, jaringan saraf tiruan, atau mesin vektor pendukung dapat diuji untuk menemukan yang paling cocok dengan data dan masalah penelitian. Model yang diusulkan harus dijelaskan secara rinci, termasuk parameter dan proses pelatihannya.

3.5. Evaluasi

Evaluasi adalah tahap di mana kinerja model yang telah dibuat diuji dan divalidasi. Metrik evaluasi seperti akurasi, presisi, dan recall, digunakan untuk menilai seberapa baik model dapat memprediksi atau mengklasifikasikan data yang tidak terlihat. Penting juga untuk melakukan validasi silang untuk memastikan bahwa model tidak overfitting dan memiliki kemampuan generalisasi yang baik. Hasil evaluasi ini akan menentukan apakah model siap untuk diterapkan atau memerlukan penyesuaian lebih lanjut.

4. HASIL DAN PEMBAHASAN

Pada tahap ini akan dibahas secara mendalam hasil dari proses-proses yang akan digunakan pada penelitian ini.

4.1. Pengumpulan Data

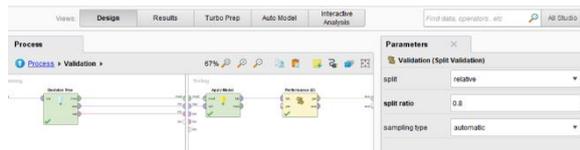
Data yang digunakan untuk studi ini diperoleh dari repositori data yang dapat diakses publik di world wide web, umumnya dikenal sebagai situs web data publik terbuka. Selain itu, untuk menambah dan menambah data yang disebutkan di atas, para peneliti juga memperoleh satu set data kualitas air, yang diperoleh langsung dari situs web dengan URL www.kaggle.com. Penting untuk dicatat bahwa data kualitas air yang diperoleh sendiri ini secara khusus diperoleh untuk tujuan penelitian ini. Pengumpulan data yang digunakan dalam penelitian ini mencakup total sebelas variabel unik, yang masing-masing dipilih dengan cermat dan dianggap relevan dengan klasifikasi penyakit stroke.

id	gender	age	hypertension	heart_disease	heart_disease_1	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke	
1	9046	Male	67	0	1	Yes	Private	Urban	22869	366	formerly smoked	1
2	51676	Female	61	0	0	Yes	Self-employed	Rural	20221	never smoked	1	
3	31112	Male	80	0	1	Yes	Private	Rural	10592	325	never smoked	1
4	60182	Female	49	0	0	Yes	Private	Urban	17123	344	smokes	1
5	1665	Female	79	1	0	Yes	Self-employed	Rural	17412	24	never smoked	1
6	56669	Male	81	0	0	Yes	Private	Urban	18621	29	formerly smoked	1
7	53882	Male	74	1	1	Yes	Private	Rural	7009	274	never smoked	1
8	10434	Female	69	0	0	No	Private	Urban	9439	228	never smoked	1
9	27419	Female	59	0	0	Yes	Private	Rural	7615	Unknown	1	
10	60491	Female	78	0	0	Yes	Private	Urban	5837	242	Unknown	1
11	12109	Female	81	1	0	Yes	Private	Rural	8043	297	never smoked	1
12	12095	Female	61	0	1	Yes	Govt. job	Rural	12046	368	smokes	1
13	12275	Female	54	0	0	Yes	Private	Urban	10451	273	smokes	1
14	8213	Male	78	0	1	Yes	Private	Urban	21994	Unknown	1	
15	5317	Female	79	0	1	Yes	Private	Urban	21409	282	never smoked	1
16	50202	Female	50	1	0	Yes	Self-employed	Rural	16741	305	never smoked	1
17	56112	Male	64	0	1	Yes	Private	Urban	19161	375	smokes	1
18	34120	Male	75	1	0	Yes	Private	Urban	22129	258	smokes	1
19	27458	Female	60	0	0	No	Private	Urban	8922	378	never smoked	1

Gambar 2. Data Mentah

4.2. Pra-pemrosesan

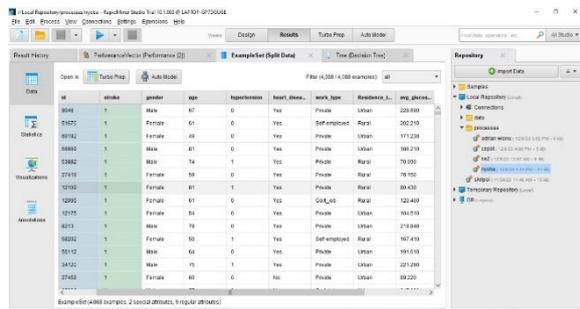
Setelah data berhasil dikumpulkan, tahap selanjutnya adalah pra-pemrosesan data. Pada tahap ini, data dibagi menjadi dua bagian: data pelatihan (training data) dan data pengujian (testing data) dengan rasio 0.8 dan 0.2. Data pelatihan digunakan untuk membangun model pohon keputusan, sedangkan data pengujian digunakan untuk menguji kinerja model. Proses ini penting untuk memastikan bahwa model yang dibuat dapat bekerja dengan baik pada data yang belum pernah dilihat sebelumnya.



Gambar 3. Proses Split Data

4.3. Transformasi

Transformasi data melibatkan perubahan data ke dalam format yang lebih sesuai untuk analisis. Dalam penelitian ini, atribut "heart_disease" dihilangkan karena terjadi error. Selain itu, label "stroke" diubah menjadi tipe binomial. Proses ini penting untuk memastikan bahwa data dalam format yang tepat untuk algoritma yang akan digunakan, dalam hal ini Decision Tree C4.5.



Gambar 4. Transformasi Data

Dari data tersebut terdiri variabel berikut:

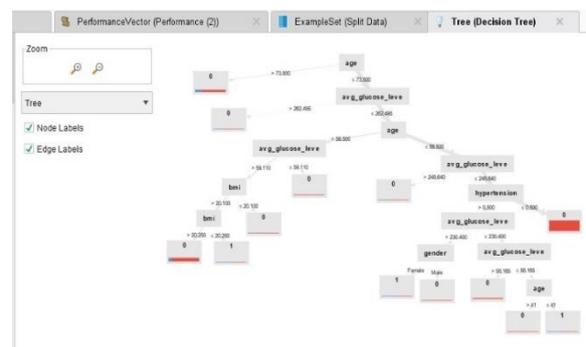
- 1) id: pengenal unik
- 2) jenis kelamin: "Pria", "Wanita" atau "Lainnya"
- 3) usia: usia pasien
- 4) hipertensi: 0 jika pasien tidak memiliki hipertensi, 1 jika pasien memiliki hipertensi
- 5) penyakit_jantung: 0 jika pasien tidak memiliki penyakit jantung, 1 jika pasien memiliki penyakit jantung
- 6) pernah_menikah: "Tidak" atau "Ya"
- 7) jenis_pekerjaan: "anak-anak", "PNS", "Tidak pernah bekerja", "Swasta" atau "Wiraswasta"
- 8) Jenis_Tempat_Tinggal: "Pedesaan" atau "Perkotaan"
- 9) avg_glucose_level: kadar glukosa rata-rata dalam darah
- 10) bmi: indeks massa tubuh

- 11) status_merokok: "pernah merokok", "tidak pernah merokok", "merokok" atau "Tidak diketahui "*"
- 12) stroke: 1 jika pasien mengalami stroke atau 0 jika tidak

*Catatan: "Tidak diketahui" pada status_merokok berarti informasi tersebut tidak tersedia untuk pasien ini.

4.4. Modeling Data Mining

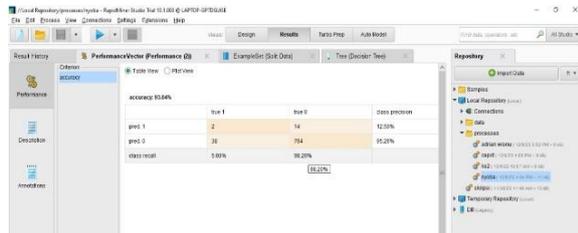
Pada tahap ini, dua model data mining, yaitu Decision Tree dan C4.5, dibuat menggunakan data yang telah diproses. Pemodelan Algoritma Decision Tree dilakukan dengan menggunakan 11 atribut yang merupakan faktor-faktor yang dapat mempengaruhi risiko stroke. Model ini berbentuk pohon keputusan, yang merupakan representasi grafis dari kemungkinan hasil yang dapat terjadi berdasarkan serangkaian pilihan yang berkaitan dengan atribut-atribut tersebut. Sebelum pembuatan model, langkah pertama yang dilakukan adalah penanganan missing value. Missing value adalah kondisi di mana beberapa data tidak memiliki nilai dan perlu ditangani dengan hati-hati karena dapat mempengaruhi kinerja model. Metode yang digunakan untuk menangani missing value dapat bervariasi, mulai dari penghapusan sampai pengisian dengan nilai tertentu, tergantung pada konteks dan jumlah missing value. Setelah itu, data dibagi menjadi dua bagian: data yang berpotensi terkena stroke dan data yang tidak berpotensi terkena stroke. Pembagian ini dilakukan berdasarkan atribut-atribut yang telah ditentukan dan menggunakan data training. Tujuannya adalah untuk melatih model dengan data yang berpotensi terkena stroke dan kemudian menguji kinerjanya dengan data yang tidak berpotensi terkena stroke. Hasil Pemodelan decision tree dapat dilihat pada Gambar 5.



Gambar 5. Hasil Pemodelan Decision Tree

Hasil Akurasi Decision Tree C4.5. Setelah model dibuat dan diuji, hasil akurasi dari model tersebut ditampilkan. Akurasi adalah metrik evaluasi yang mengukur seberapa sering model membuat prediksi yang benar. Dalam eksperimen ini, model Algoritma Decision Tree C4.5 berhasil mencapai akurasi sebesar 93.64%. Nilai ini menunjukkan bahwa model ini cukup baik dalam membuat prediksi atau klasifikasi

risiko stroke berdasarkan 11 atribut yang digunakan. Namun, penting untuk diingat bahwa akurasi bukanlah satu-satunya metrik evaluasi yang penting dan harus dipertimbangkan bersama dengan metrik lain seperti presisi, dan recall. Untuk mendapatkan gambaran yang lebih lengkap tentang kinerja model.



Gambar 6. Hasil Akurasi

4.5. Evaluasi

Evaluasi dari algoritma pohon keputusan C4.5 Setelah melakukan proses decision tree c4.5 yang dilakukan terdapat hasil dengan menampilkan nilai accuracy sebesar 93.64%, pred.1 (positif stroke) terdapat 2 pasien stroke dan tidak stroke (negatif stroke) 14 dengan class precision 12.50 % class recall 5.00% kemudian pada pred.0 terdapat 38 pasien stoke dan tidak stroke 764 dengan class precision 95.26% class recall 98.20%. Decision TreeC4.5 untuk mengklasifikasikan penyakit stroke merupakan salah satu cara untuk memprediksi penyakit stroke yang bagus untuk di gunakan untuk klasifikasi awal penyakit stroke. Prediksi penyakit menggunakan Decision Tree C4.5 memiliki hasil yang bagus dapat dilihat dari hasil perhitungan confusion matrix yang mendapatkan hasil accuracy sebesar 93,64%. Dapat dilihat pada Gambar 7.



Gambar 7. Hasil Evaluasi

Evaluasi algoritma pohon keputusan atau decision Tree C4.5 menggunakan confusion matrix untuk menghitung metrik kinerja seperti akurasi, presisi, dan recall. Confusion matrix adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi pada set data uji yang nilai sebenarnya diketahui. Berdasarkan confusion matrix yang direpresentasikan pada Gambar 7 gambar, dapat diidentifikasi nilai-nilai berikut:

- 1) True Positives (TP) untuk stroke: 2
- 2) False Negatives (FN) untuk stroke: 14
- 3) True Negatives (TN) untuk tidak stroke: 764
- 4) False Positives (FP) untuk tidak stroke: 38

Dengan nilai-nilai ini, peneliti dapat menghitung metrik kinerja model:

$$Akurasi = \frac{2 + 764}{2 + 14 + 38 + 764} \times 100 = 93.64\%$$

$$Precision\ Stroke = \frac{2}{2 + 38} = 5\%$$

$$Recall\ Stroke = \frac{2}{2 + 14} = 12.5\%$$

$$Precision\ Tidak\ Stroke = \frac{764}{14 + 764} = 98.2\%$$

$$Recall\ Tidak\ Stroke = \frac{764}{764 + 38} = 95.26\%$$

Dari perhitungan ini, dapat dilihat bahwa model memiliki akurasi yang tinggi secara keseluruhan, namun presisi dan recall untuk kasus stroke positif (pred.1) relatif rendah, yang menunjukkan bahwa model mungkin kurang efektif dalam mengidentifikasi semua kasus stroke sebenarnya. Ini penting untuk diperhatikan karena dalam konteks medis, kegagalan untuk mengidentifikasi kasus stroke dapat memiliki konsekuensi serius. Oleh karena itu, meskipun akurasi tinggi, model mungkin perlu ditingkatkan dalam hal presisi dan recall untuk kasus stroke positif.

5. KESIMPULAN DAN SARAN

Penelitian ini berfokus pada klasifikasi data stroke menggunakan algoritma pohon keputusan atau Decision Tree C4.5. Hasil penelitian menunjukkan tingkat akurasi yang mengesankan sebesar 93,64%, namun nilai presisi untuk prediksi stroke positif hanya 12,50%, sementara presisi untuk prediksi stroke negatif mencapai 95,26%. Meskipun tingkat akurasi tinggi dicapai, perlu dilakukan peningkatan dan penyempurnaan model untuk meningkatkan presisi prediksi stroke positif. Dalam konteks ini, disarankan untuk melakukan penelitian lebih lanjut dengan mempertimbangkan lebih banyak faktor risiko stroke, seperti riwayat keluarga dan kondisi kesehatan lainnya. Selain itu, penelitian lebih lanjut juga dapat dilakukan untuk membandingkan efektivitas Decision Tree C4.5 dengan algoritma klasifikasi lainnya, seperti Random Forest, Naive Bayes, atau Support Vector Machine (SVM), dalam konteks data stroke. Hal ini akan membantu dalam mengembangkan model prediksi yang lebih akurat dan efisien untuk stroke, yang pada akhirnya dapat berkontribusi pada upaya pencegahan dan pengobatan stroke yang lebih baik.

DAFTAR PUSTAKA

- [1] M. Anggraini, "Berhubungan Dengan Perilaku Pasien Pasca Stroke," *J. Kesehat. Tambusai*, Vol. 3, No. 2, Pp. 260–266, 2022.
- [2] F. I. Komputer And U. D. Nuswantoro, "Penyakit Stroke Dengan Klasifikasi Data Mining Pada," *Comput. Sci.*, 2015.
- [3] A. Faisal And A. Subekti, "Deep Neural Network Untuk Prediksi Stroke," *Jepin (Jurnal Edukasi*

- Dan Penelit. Inform.*, Vol. 7, No. 3, Pp. 443–449, 2021.
- [4] Y. T. Widayanti, Y. Prihati, And S. Widjaja, “Analisis Dan Komparasi Algoritma Naïve Bayes Dan C4.5 Untuk Klasifikasi Loyalitas Pelanggan Mnc Play Kota Semarang,” *Transformtika*, Vol. 18, No. 2, Pp. 161–172, 2021.
- [5] I. Handayani, “Penyakit Disk Hernia Dan Spondylolisthesis Dalam Kolumna Vertebralis,” *Jasiek*, Vol. 1, No. 2, Pp. 83–88, 2019, Doi: 10.12928/Jasiek.V13i2.Xxxx.
- [6] A. Byna And M. Basit, “Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes,” *J. Sisfokom (Sistem Inf. Dan Komputer)*, Vol. 09, No. 03, Pp. 407–411, 2020.
- [7] E. R. Pambudi, Sriyanto, And Firmansyah, “Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C.45,” *J. Tek.*, Vol. 16, No. 02, Pp. 221–226, 2022.
- [8] F. Adha, H. Airi, T. Suprpti, And A. Bahtiar, “Komparasi Metode Klasifikasi Data Mining Untuk Prediksi,” *J. Tek. Elektro Dan Inform.*, Vol. 18, Pp. 73–79, 2023.
- [9] Z. Zuriati And N. Qomariyah, “Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (Knn) Classification Of Stroke Using The K-Nearest Neighbor (Knn) Algorithm,” *Routers J. Sist. Dan Teknol. Inf.*, Vol. 1, No. 1, Pp. 1–8, 2023.
- [10] Z. Sitorus And A. Widarma, “Data Mining Algoritma Decision Tree Iterative Dechotomiser 3 (Id3) Untuk Klasifikasi Penyakit Stroke,” *Cess (Journal Comput. Eng. Syst. Sci.)*, Vol. 8, No. 2, Pp. 554–563, 2023.
- [11] M. Purnamasari And F. D. Hastuti, “Klasifikasi Siswa Berdasarkan Mata Pelajaran Lintas Minat Menggunakan Metode Decision Tree C4.5,” *Jsii / J. Sist. Inf.*, Vol. 8, No. 2, Pp. 141–149, 2021.