JURNAL INFORMATIKA, Vol. 9 No. 1 April 2022, Halaman 13-17
ISSN: 2355-6579 | E-ISSN**:** 2528-2247

Analisis Perbandingan Algoritma C4.5 Dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular

Kelvin Leonardi Kohsasih[1], Zakarias Situmorang [2]

[1,2] Program Studi Pascasarjana Ilmu Komputer, Fakultas Teknik dan Ilmu Komputer
Universitas Potensi Utama, Jl.K.L. Yos Sudarso KM.6,5 No.3A, Tj Mulia, Medan

e-mail: [1]ceokelvin12@gmail.com, [2]zakarias65@yahoo.com

Informasi Artikel	Diterima: 21-12-2021	Direvisi: 23-02-2022	Disetujui: 03-03-2022

Abstrak

Penyakit *Cerebrovascular* atau stroke merupakan salah satu penyebab utama kematian di dunia. stroke adalah penyakit yang disebabkan oleh gangguan pada pembuluh darah yang mensuplai darah ke otak. Machine learning merupakan teknologi yang dapat digunakan untuk memprediksi stroke. Algoritma machine learning bersifat konstruktif dalam membuat prediksi yang akurat dan memberikan analisis yang akurat. Salah satu algoritma klasifikasi machine learning yang dapat digunakan untuk melakukan prediksi adalah Algoritma *Decision Tree* C4.5 dan Algoritma Naive Bayes. Tujuan dalam penelitian ini yaitu untuk membandingkan akurasi dan kinerja dua algoritma untuk memprediksi Penyakit Cerebrovascular atau stroke. Berdasarkan hasil penelitian didapatkan bahwa algoritma C4.5 memperoleh tingkat akurasi yang lebih tinggi yaitu 95% sedangkan algoritma Naive Bayes memperoleh tingkat akurasi 91%

Kata Kunci: Penyakit Cerebrovascular; Machine learning; Prediksi Stroke

Abstract

Cerebrovascular accidents or strokes are one of the leading causes of death worldwide. Stroke is a disease caused by a malfunction of the blood vessels that supply the blood to the brain. Machine learning is a technology that can be used to predict stroke. Machine learning algorithms are constructive when making accurate predictions and providing accurate analysis. One of the machine learning classification algorithms that can be used for prediction is the Decision Tree C4.5 algorithm and the Naive Bayes algorithm. The purpose of this study is to compare the accuracy and performance of the two algorithms for predicting cerebrovascular disease or stroke. Based on the results of the study, it was found that the C4.5 algorithm achieved a higher accuracy of 95% and the Naive Bayes algorithm achieved a precision of 91%.

Keywords*: Cerebrovascular Disease; Machine learning; Stroke Prediction*

1. Pendahuluan

Cerebrovascular Disease atau Stroke adalah penyakit *cerebrovascular* di mana munculnya disfungsi otak dikaitkan dengan gangguan pada pembuluh darah yang mensuplai darah ke otak (Widyaswara Suwaryo et al., 2019). Menurut *Centers for Disease Control and Prevention* (CDC), Stroke merupakan salah satu penyebab utama kematian di Amerika Serikat. Stroke adalah penyakit tidak menular yang menyumbang sekitar 11% dari semua kematian dan Lebih dari 795.000 orang di AS mengalami efek samping stroke(Sailasya & Kumari, 2021). Selain itu, Menurut data *World Health Organization* (WHO) 7,9% dari seluruh kematian di Indonesia disebabkan oleh stroke (Mutiarasari et al., 2019).

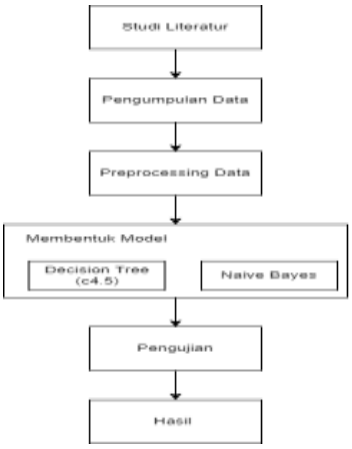
Dengan kemajuan teknologi di bidang medis, machine learning dapat digunakan untuk memprediksi stroke. Algoritma machine learning bersifat konstruktif dalam membuat prediksi yang akurat dan memberikan analisis yang akurat. Banyak peneliti telah menggunakan algoritma berbasis machine learning untuk memprediksi stroke. (Chun-An Cheng et al., 2014), melakukan penelitian untuk memprediksi prognosis stroke iskemik dengan mengumpulkan data dari 82 pasien stroke iskemik. Penelitian ini menggunakan 2 model yaitu model pertama dibangun dengan model Multi-Layer Perceptron (MLP) 13-11-2 yang

mendapatkan hasil *sensitivity* 77.78%, *specificity* 80.43% dan accuracy 79.27%. sedangkan pada model kedus yang dibangun dengan model Multi-Layer Perceptron (MLP) 13-7-2 mendapatkan hasil yang lebih baik yaitu *sensitivity* 94.44%, *specificity* 95.65% dan *accuracy* 95.12%. (Amini et al., 2013)melakukan penelitian untuk memprediksi stroke dengan menggunakan 807 subjek sehat dan sakit dari 50 faktor resiko stroke seperti riwayat penyakit kardiovaskular, diabetes, hiperlipidemia, merokok dan konsumsi alkohol. Proses prediksi menggunakan teknik data mining menggunakan algoritma K-nearest neighbor dan C4.5 *Decision Tree*. Dimana tingkat keakuratan masing masing algorima adalah 94.18% and 95.42%. menurut (Singh & Choudhary, 2017) dalam penelitiannya melakukan prediksi penyakit stroke menggunakan algoritma *Decision Tree* dan algortima klasifikasi *back propagation neural network* dengan menganalisis dan membandingkan efisiensi klasifikasi. Dimana pada penelitian tersebut didapatkan model prediksi yang optimal untuk penyakit stroke dengan akurasi 97,7%.

Berdasarkan beberapa penelitian terdahulu, Pada penelitian ini peneliti akan melakukan penelitian untuk menganalisis dan membandingkan algoritma C4.5 *Decision Tree* dengan algorima Naive Bayes untuk melihat perbandingan tingkat akurasi, presisi, recall dan f1-score dalam memprediksi penyakit stroke. Selain itu, *specificity* dari hasil pengujian juga harus dianalisis untuk melihat seberapa efektif kedua algoritma tersebut.

2. Metode Penelitian

Penelitian ini dilakukan dalam beberapa tahap yaitu tinjauan pustaka, pengumpulan data, preprosesing data, membentuk model, pengujian, dan hasil. alur tahapan penelitian pada penelitian ini dapat dilihat pada gambar 1.



Gambar 1.Alur Tahapan Penelitian

A. Studi Literatur

Pada tahapan ini, peneliti mempelajari masalah yang terkait dengan penyakit *Cerbravascular* atau stroke, Informasi tentang algortima C4.5 dan algoritma Naive Bayes, dan informasi lain yang terkait dengan masalah yang sedang dibahas. Referensi pada penelitian ini diperoleh dari berbagai macam sumber diantaranya yaitu buku, majalah, ebook, dan laporan penelitian sebelumnya.

B. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah dataset *Stroke Prediction Dataset* yang tersedia pada repositori Kaggle yang merupakan salah satu situs yang terkenal di dunia *Data Science* dan *Machine Learning* yang terdiri dari lebih dari 6000 dataset. Total data yang terdapat pada dataset ini yaitu 5110 data observasi dengan 12 atribut (fedesoriano, 2021). atribut data yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Tabel 1. Atribut Dataset

No	Nama	Keterangan
1	id	id pasien
2	gender	jenis kelamin
3	age	usia pasien
4	hypertension	hipertensi
5	heart_disease	penyakit jantung
6	ever_married	pernah menikah
7	work_type	jenis pekerjaan
8	residence_type	tempat tinggal
9	avg_glucose_level	kadar glukosa
10	bmi	massa tubuh
11	smoking_status	status merokok
12	stroke	prediksi stroke

C. Data Preprocessing

Data Preprocessing (Pra-pemrosesan data) adalah proses mengubah data mentah menjadi format yang dapat digunakan. *Data Preprocessing* terdiri dari beberapa bagian, yaitu *cleaning data, data transformation, dan data reduction*. Tahap preprocessing yang digunakan dalam penelitian adalah data cleaning, yang merupakan bagian dari proses preprocessing data sebelum digunakan. Oleh karena itu, pada langkah ini dilakukan pembersihan data untuk menghilangkan atribut dan data yang tidak lengkap, tidak akurat, tidak konsisten dan tidak relevan. (Agarwal, 2015; Sapna Devi & Dr. Arvind Kalia, 2015). Tujuan dilakukan *data cleaning* adalah untuk menghasilkan data yang benar-benar terpakai.

D. Membentuk Model

Pada tahapan penbentukan model data akan dibagi menjadi 2 yaitu *data training* dan

data testing dimana persentase pembagian data bersifat bebas. dataset yang digunakan pada penelitian ini akan dibagi menjadi 60% *data training* dan 40% *data testing*. Pada penelitian ini model akan dibentuk dengan menggunakan algoritma C4.5 decision tree dan algoritma Naive Bayes yang selanjutnya akan dilakukan perbandingan nilai akurasi, presisi, recall dan f1-score masing masing algorima dalam melakukan prediksi terhadap penyakit *Cerbravascular* atau stroke. Akurasi merupakan rasio dari hasil prediksi yang benar. presisi adalah rasio prediksi positif yang benar dengan hasil keseluruhan dari prediksi positif. recall Adalah rasio antara prediksi yang benar-benar positif dengan data global yang sebenarnya positif.Kemudian F1 Score adalah perbandingan antara presisi rata-rata dan recall. Adapun rumus untuk menghitung akurasi, presisi, dan recall masing-masing ditunjukkan pada rumus (1), (2), dan (3)(Goutte & Gaussier, 2005).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (1)$$

$$\text{Pr}\, e\, sisi = \frac{TP}{FP + TP} * 100\% \quad (2)$$

$$Recall = \frac{TP}{FN + TP} * 100\% \quad (3)$$

Keterangan:
TP : Banyaknya data positif yang benar
TN : Banyaknya data negatif yang benar
FN : Banyaknya data negatif yang salah
FP : Banyaknya data positif yang salah

Algoritma C4.5 merupakan salah satu algoritma yang dapat digunakan untuk mengklasifikasikan data berdasarkan atribut numerik dan kategorik. Hasil dari proses klasifikasi dapat digunakan dalam bentuk aturan untuk memprediksi nilai atribut diskrit dari record yang baru. Algoritma C4.5 sendiri merupakan evolusi lebih lanjut dari algoritma ID3, dimana pengembangan dapat dilakukan dalam kasus, data yang hilang dan data persisten dapat diperbaiki (Elisa, 2017).

Langkah-langkah dalam pengolahan data menggunakan algoritma C4.5 antara lain mencari nilai entropi, mencari nilai gain, membentuk pohon keputusan dan aturan. Rumus untuk menghitung entropi dan gain masing-masing dapat dilihat pada persamaan (4) (5) berturut turut (Setyanto dan Hanif Al Fattah, 2017).

$$Entropy(s) = \sum_{i=0}^{n} -pi * \log_2 pi \quad (4)$$

Keterangan:
S : himpunan kasus
n : Jumlah Partisi S
pi : proporsi dari Si terhadap S

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_I) \quad (5)$$

Keterangan:
S : himpunan kasus
n : jumlah partisi atribut A
|S| : jumlah kasus dalam S.
|Si|: jumah kasus pada partisi ke i

Algoritma Naive Bayes adalah pengklasifikasi menggunakan metode probabilistik dan statistik yang diusulkan oleh seorang ilmuwan Inggris, Thomas Bayes. Naive Bayes adalah model yang populer dalam aplikasi pembelajaran mesin karena kesederhanaannya yang memungkinkan semua atribut berkontribusi sama pada keputusan akhir. Kesederhanaan ini setara dengan efisiensi komputasi, yang membuat teknik Naive Bayes menarik dan cocok untuk banyak bidang. Elemen utama dari pengklasifikasi Naive Bayes berkaitan dengan tiga aspek, yaitu, *prior, posterior dan class conditional probability* (Wibawa et al., 2019). Rumus Teorema Bayes dapat dilihat pada persamaan (6):

$$P(Q|X) = \frac{P(X|Q)P(Q)}{P(X)} \quad (6)$$

Keterangan:
X : Data dengan kelas tidak diketahui
Q : Hipotesis X pada class spesifik
P(Q|X) : Probabilitas hipotesis Q berdasarkan kondisi X (posteriori probabilitas) 0
P(Q) : Probabilitas hipotesis Q
P(X|Q) : Probabilitas X terhadap hipotesis Q
P(X) : Probabilitas X

E. Pengujian Model

Pada tahapan ini,peneliti melakukan menguji model yang dibuat menggunakan algoritma pohon keputusan C4.5 dan algoritma naive Bayes. Pengujian model dilakukan dengan menggunakan aplikasi orange Tujuan dari pengujian model adalah untuk membandingkan akurasi dan kinerja dua algoritma untuk memprediksi Penyakit Cerebrovascular atau stroke.

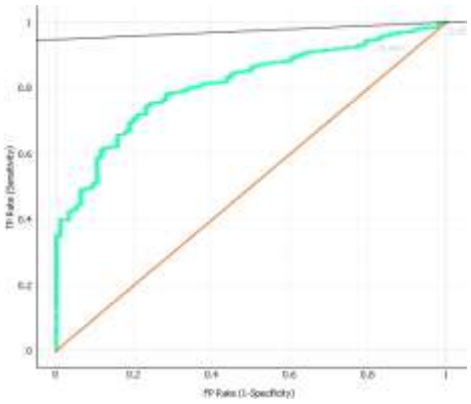
3. Hasil dan Pembahasan

Pada tahap ini peneliti menggunakan metode algoritma C4.5 dan algoritma Naive Bayes untuk mengaplikasikan data yang telah mengalami proses *preprocessing* data atau pembersihan data pada aplikasi orange. Aplikasi orange atau biasa dikenal dengan orange data mining merupakan suatu *software open source* yang digunakan untuk melakukan data analytic. Berdasarkan pengujian yang dilakukan menggunakan aplikasi orange didapatkan hasil yaitu: algoritma C4.5 mencapai akurasi 95% dan algoritma Nave Bayes mencapai hasil akurasi 91%. Tabel 2 menunjukkan perbandingan kinerja Algoritma C4.5 dan Naive Bayes.

Tabel 2. Perbandingan Kinerja

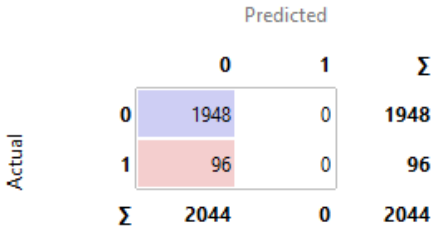
Model	*Accuracy*	Precision	Recall Score	F1-Score
C4.5 *Decision Tree*	0.953	0.908	0.953	0.930
Naïve Bayes	0.913	0.921	0.913	0.917

Selain itu, hasil pengujian menunjukkan area under ROC curve (AUC) pada algoritma C4.5 dan Naïve Bayes berturut-turut adalah 0.500 dan 0.810 . Plot ROC untuk kedua algoritma dapat dilihat pada Gambar 2.

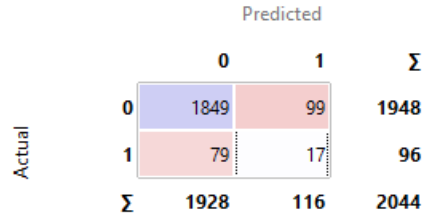


Gambar 2. Plot ROC

Sumbu y pada plot ROC mewakili *True Positive Rate*, sumbu x pada grafik ROC menunjukkan *False Positive Rate*. Di mana garis lurus pada grafik ROC menggambarkan semua kemungkinan TP dan FP jika kita menjalankan ambang batas dari rendah ke tinggi. Confusion matrix dari algoritma C4.5 dan Naïve Bayes dapat dilihat pada Gambar 3 dan 4.



Gambar 3. Confusion Matrix Algoritma C4.5



Gambar 4. Confusion Matrix Algoritma Naïve Bayes

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan yaitu dengan membagi dataset menjadi 60% data training dan 40% data testing maka dapat simpulkan bahwa algortitma C4.5 memiliki performa yang lebih baik yaitu dengan tingkat akurasi sebesar 95% serta nilai presisi, recall dan f1-score masing masing yaitu 90% , 95% dan 93%. sedangkan algoritma naïve bayes mendapatkan tingkat akurasi sebesar 91%, presisi 92%, recall 91% dan f1-score sebesar 92%. selain itu hasil log loss dan specificity dari algorima naïve bayes yaitu 0.205 dan 0.213 sedangkan algoritma C4.5 mendapatkan nilai masing masing yaitu 0.190 dan 0.047.

Referensi

Agarwal, V. (2015). Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis. *International Journal of Computer Applications*, *131*(4), 30–36. https://doi.org/10.5120/ijca2015907309

Amini, L., Azarpazhouh, R., Farzadfar, M. T., Mousavi, S. A., Jazaieri, F., Khorvash, F., Norouzi, R., & Toghianfar, N. (2013). Prediction and control of stroke by data mining. *International Journal of Preventive Medicine*, *4*(Suppl 2), S245-9. http://www.ncbi.nlm.nih.gov/pubmed/23776732

Chun-An Cheng, Yi-Ching Lin, & Hung-Wen Chiu. (2014). Prediction of the Prognosis of Ischemic Stroke Patients after Intravenous

Thrombolysis Using Artificial Neural Networks. *Integrating Information Technology and Management for Quality of Care*, *202*, 115–118. https://doi.org/10.3233/978-1-61499-423-7-115

Elisa, E. (2017). *Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti*. *2*(1).

fedesoriano. (2021, January 27). *Stroke Prediction Dataset*. Kaggle. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata

Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Lecture Notes in Computer Science*, *3408*, 345–359. https://doi.org/10.1007/978-3-540-31865-1_25

Mutiarasari, D., Kesehatan, B. I., Komunitas, M.-K., & Kedokteran, F. (2019). ISCHEMIC STROKE: SYMPTOMS, RISK FACTORS, AND PREVENTION. In *Jurnal Ilmiah Kedokteran* (Vol. 6, Issue 1).

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, *12*(6), 2021. https://doi.org/10.14569/IJACSA.2021.0120662

Sapna Devi, & Dr. Arvind Kalia. (2015). Study of Data Cleaning & Comparison of Data Cleaning Tools. *International Journal of Computer Science and Mobile Computing*, *4*(3), 360–370. http://www.ijcsmc.com/

Setyanto dan Hanif Al Fattah, A. (2017). Analisis Perbandingan Algoritma Decision Tree (C4.5) Dan K-Naive Bayes Untuk Mengklasifikasi Penerimaan Mahasiswa Baru Tingkat Universitas. In *Indonesian Journal of Applied Informatics* (Vol. 2, Issue 1).

Singh, M. S., & Choudhary, P. (2017). Stroke prediction using artificial intelligence. *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, 158–161. https://doi.org/10.1109/IEMECON.2017.8079581

Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. *International Journal of Recent Contributions from Engineering, Science & IT (IJES)*, *7*(2), 91. https://doi.org/10.3991/ijes.v7i2.10659

Widyaswara Suwaryo, P. A., Widodo, W. T., & Setianingsih, E. (2019). Faktor Risiko yang Mempengaruhi Kejadian Stroke. *Jurnal Keperawatan*, *11*(4), 251–260. https://doi.org/10.32583/keperawatan.v11i4.530