MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 7 | No. 2 | Halaman 151 - 164
DOI : https://doi.org/10.26760/mindjournal.v7i2.151-164 Desember 2022

Prediksi Awal Penyakit Stroke Berdasarkan Rekam Medis menggunakan Metode Algoritma CART(*Classification and Regression Tree*)

AGIEL FADILLAH HERMAWAN, FAJRI RAKHMAT UMBARA, FATAN KASYIDI

Fakultas Sains dan Informatika Universitas Jenderal Achmad Yani, Cimahi, Indonesia.
Email: agielfadillah18@if.unjani.ac.id

Received 23 Agustus 2022 **|** *Revised* 2 November 2022 **|** *Accepted* 19 Desember 2022

ABSTRAK

Seiring perkembangan zaman bidang teknologi dapat membantu banyak hal salah satu contoh nya dapat membantu bidang kesehatan, teknologi seperti machine learning dan data mining dapat membantu dalam melakukan prediksi penyakit stroke. Oleh karena itu, penelitian kali ini akan menerapkan salah satu metode data mining klasifikasi untuk memprediksi penyakit stroke dengan tujuan dapat mengetahui model dari algoritma yang akan digunakan yaitu Algoritma Classification and Regression Tree atau CART. Metode ini melakukan perhitungan menggunakan nilai ginigain dan giniindex untuk membuat sebuah pohon keputusan. Dengan menggunakan Stroke Prediction Dataset dan dilakukan beberapa eksperimen didapatkan hasil akurasi terbesar sebesar 89,83% pada split data 80/20. Pohon keputusan dapat dipangkas untuk mengidentifikasi dan membuang cabang pohon yang tidak diperlukan, pada penelitian kali ini dilakukan pemangkasan untuk dilihat seberapa berpengaruh pemangkasan pada akurasi algoritma ini dan didapatkan hasil akurasi terbesar sebesar 74,73% maka pemangkasan dinilai kurang berpengaruh pada akurasi algoritma ini.

***Kata kunci**: Stroke, Prediksi, Klasifikasi, Data Mining, CART*

ABSTRACT

Along with the times, technology can help many things, one example of which can help the health sector, technology such as machine learning and data mining can help in predicting stroke. Therefore, this study will apply one of the classification data mining methods to predict stroke with the aim of knowing the model of the algorithm to be used, namely the Classification and Regression Tree Algorithm or CART. This method performs calculations using the Ginigain and Ginindex values to create a decision tree. By using the Stroke Prediction Dataset and conducting several experiments, the highest accuracy results were 89.83% in the 80/20 data split. In this study pruning was carried out to see how much pruning had an effect on the accuracy of this algorithm and the highest accuracy result was 74.73%, so pruning was considered to have less effect on the accuracy of this algorithm.

***Keywords**: Stroke, Prediction, Classification, Data Mining, CART*

1. PENDAHULUAN

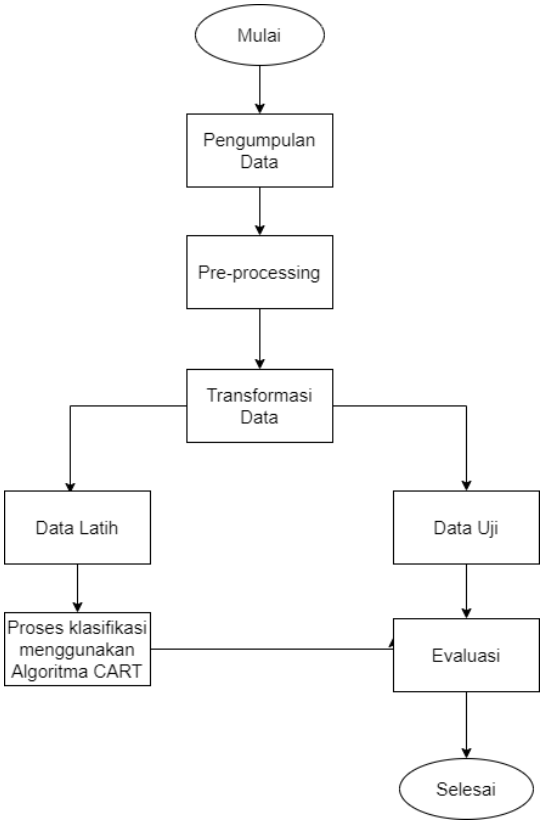
Seiring perkembangan zaman bidang teknologi dapat membantu banyak hal salah satu contoh nya membantu pada bidang kesehatan **(Tjandrawinata, 2016)**, teknologi seperti *machine learning* dan *data mining* dapat membantu untuk memprediksi penyakit stroke. Oleh karena itu, metode *data mining* klasifikasi dapat dilakukan untuk mengukur seberapa besar tingkat akurasi dari algoritma yang akan digunakan pada dataset stroke yang akan digunakan pada penelitian ini. Algoritma *Classification and Regression Tree* (CART) yang akan digunakan pada penelitian ini. Algoritma CART ini membentuk sebuah pohon keputusan atau *decision tree,* Algoritma CART dikembangkan untuk melakukan analisis klasifikasi yang menghasilkan kelas berdasarkan pohon yang dibuat **(Nafi'iyah, 2015)**. Algoritma CART ini memiliki beberapa kelebihan diantaranya adalah hasilnya lebih mudah ditafsirkan dan lebih mudah pada perhitungan karena hanya menggunakan dua persamaan, juga algoritma CART ini dapat diterapkan pada himpunan data yang besar **(Pratiwi & Zain, 2014)**. Pada pohon keputusan juga dapat dilakukan pemangkasan atau pruning untuk mengidentifikasi dan membuang cabang pohon yang tidak diperlukan pada pohon yang telah dibentuk, dan untuk dilihat seberapa berpengaruh pruning pada tingkat akurasi dari algortima **(Rahayu et al., 2015)**.

Berdasarkan latar belakang diatas, maka penulis menerapakan prediksi menggunakan *Classification and Regression Tree* (CART). CART yang diusulkan oleh Breiman pada tahun 1984 telah menjadi metode pembelajaran pohon keputusan yang banyak digunakan **(Zhang et al., 2018)**. Metode ini memiliki asumsi bahwasanya pohon keputusan adalah pohon biner. Metode ini merupakan Teknik klasifikasi dengan menggunakan algoritma rekonsiliasi rekursif biner. *Classification and Regression Tree* ini akan mengeluarkan hasil pohon yang berbentuk klasifikasi jika variable respon memiliki skala kategorikal dan akan mengeluatkan hasil pohon regresi jika variable respon merupakan data kontinu. Tujuan dari metode ini untuk mendapatkan kumpulan data yang akurat sebagai pin klasifikasi **(Subarkah et al., 2018)**.
Banyak penelitian telah dilakukan untuk memprediksikan penyakit stroke ini salah satu penelitian yang dilakukan untuk memprediksi penyakit stroke ini adalah menggunakan metode Naïve Bayes **(Byna & Basit, 2020)**. Pada penelitian yang menggunakan metode naïve bayes ini tingkat akurasinya adalah sebesar 0.976 dengan *split data* 80/20. Berdasarkan penjelasan diatas, pada penelitian ini akan dilakukan prediksi dengan menggunakan metode yang berbeda yaitu menggunakan metode *Classification and Regression Tree* (CART).

Oleh karena itu, dalam penelitian ini menggunakan metode yang berbeda dengan penelitian yang sebelumnya yaitu metode Algoritma *Classification and Regression Tree* atau bisa disingkat Algoritma CART. Metode Algoritma CART ini dipilih dikarenakan algoritma ini memiliki kelebihan yaitu metode Algoritma CART ini bersifat fleksibel dan dapat diatur sesuai kebutuhan. CART tidak memiliki asumsi dan cepat untuk menghitung. CART menentukan variable yang dianggap penting dan membuang variabel yang dianggap tidak penting **(Indah Prabawati et al., 2019)**.

2. METODE PENELITIAN

Adapun langkah-langkah yang dilakukan pada penelitain ini dapat dilihat pada Gambar 1 dibawah ini.



Gambar 1. Alur Metode Penelitian

2.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data yang diambil dari situs web *kaggle.com* dengan nama *Stroke Prediction Dataset* dari versi 1. Data tersebut dibuat pada tanggal 27 Januari 2021, yang terdiri dari 5107 *record* data. Dataset ini digunakan untuk memprediksi kemungkinan pasien yang terkena stroke berdasarkan rekam medis seperti berat badan, status merokok, dan berbagai penyakit. Data pada penelitian ini tidak seimbang atau *imbalance* antara data stroke dan data tidak stroke maka pada penelitian ini dilakukan *upsampling* menggunakan metode *Simple Random Sampling*. Setiap baris data memberikan informasi yang relevan tentang pasien. Data yang diperoleh memiliki atribut data sebagai berikut:

Tabel 1. Atribut Dataset

No.	Nama Atribut	Keterangan
1.	*Gender*	Merupakan jenis kelamin.
2.	*Age*	Merupakan Usia.
3.	*Hypertension*	Merupakan status pasien apakah memiliki Riwayat penyakit hipertansi
4.	*Heart Disaese*	Merupakan status pasien apakah memiliki Riwayat penyakit jantung
5.	*Ever Married*	Merupakan status pasien apakah pasien pernah menikah
6.	*Work Type*	Merupakan status pekerjaan dari pasien
7.	*Residence Type*	Merupakan status tipe rumah yang dimiliki oleh pasien.
8.	*Average Glucose Level*	Merupakan status gula darah yang dimiliki oleh pasien
9.	*BMI*	Merupakan index masa tubuh
10.	*Smoking Status*	Merupakan status pasien apakah merokok atau tidak
11.	*Stroke*	Merupakan status pasien tersebut apakah mempunyai penyakit stroke atau tidak

2.2. *Preprocessing*

Setelah data terkumpul lalu tahap berikutnya dalam penelitian kali ini yaitu tahapan *pre-pocessing.* Tahapan *pre-processing* adalah proses mengubah data mentah menjadi data yang dapat digunakan untuk *data mining*. Tahapan ini merupakan langkah yang penting dalam melakukan *data mining,* karena *data mining* tidak dapat dilakukan apabila data yang ada masih merupakan data yang mentah. Ada beberapa proses yang dilakukan pada tahap *pre-processing* diantaranya:

a. Pembersihan Data

Pembersihan data adalah tahapan awal dalam melakukan proses *data mining* **(Sulastri & Gufroni, 2017)***.* Dalam proses pembersihan data ini bisa dilakukan untuk mengisi data yang kosong atau *missing value,* pada penelitian ini pengisian data yang kosong menggunakan metode interpolasi data pada *Microsoft excel.* Kolom Index Masa Tubuh diisi berdasarkan usia karena Index Masa Tubuh pasien dapat juga dilihat dari usia pasien.

b. Pemilihan Data

Pemilihan data adalah proses untuk meminimalisir jumlah data yang nantinya akan digunakan untuk proses *data mining* dengan masih mengidentifikasikan data aslinya **(Pratama et al., 2019)**. Seleksi data ini memilih atribut-atribut atau yang akan digunakan pada penelitian. Atribut yang digunakan pada penelitian ini adalah *Gender, Hypertension, heart_disaese, ever_married, work_type, residence_type, smoking_status, stroke,* usia, rata-rata_gula_darah, Index masa tubuh. Atribut yang tidak digunakan yaitu *id.*

Tabel 2. Data Sebelum Dipreprocessing

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1

c. Oversampling Data

Pada dataset yang digunakan pada penelitian ini yaitu *stroke prediction dataset* terjadi ketidakseimbangan data, oleh karena itu pada peneltian ini dilakukan *Oversampling data* **(Bima et al., 2013)** untuk mengambil data dari kelas minoritas sedemikian rupa sehingga proporsinya dalam sampel lebih besar dibandingkan dengan skala awalnya. *Oversampling data* ini menggunakan metode *simple random sampling* karena metode ini tidak membutuhkan informasi tambahan pada *sample data* **(Arieska & Herdiani, 2018)**. *Oversampling data* ini menggunakan data baru sebanyak 4800 data.

2.3. Transformasi Data

Transformasi data adalah teknik yang digunakan untuk mengubah data mentah menjadi format yang sesuai yang secara efisien memudahkan *data mining* **(Anggraeni et al., 2013)**. Transfomasi data dilakukan dengan menggunakan tiga skenario *split data* yaitu untuk skenario pertama dengan menggunakan 60% data latih dan 40% data uji, skenario kedua dengan menggunakan 70% data latih dan 30% data uji, dan skenario terakhir yaitu 80% data latih dan 20% data uji. Ketiga skenario ini dipilih untuk dilihat seberapa baik akurasi algoritma CART pada masing-masing skenario. Masing-masing skenario *split data* dipisahkan langsung melalui sistem yang telah dibuat. Sistem akan merubah data mentah menjadi data yang dapat digunakan untuk dilakukan proses klasifikasi.

2.4. Proses Klasifikasi

Setelah melakukan tahapan diatas maka tahap selanjutnya yaitu menerapkan teknik *data mining* yang telah dibahas sebelumnya, teknik yang digunakan yaitu teknik klasifikasi menggunakan metode Algoritma CART atau *Classification and Regression Tree* terhadap data yang sebelumnya telah dilakukan *pre-processing* dan tahap transformasi **(Zhang et al., 2018)**. Ada beberapa tahapan pada proses klasifikasi menggunakan metode CART ini, yang pertama yaitu menyiapkan dataset yang akan digunakan, lalu menghitung *IndexGini* dari keseluruhan atribut, setelah didapatkan *IndexGini* lalu menghitung nilai *GiniGain*, setelah didapatkan nilai *GiniGain* maka dapat dibuat pohon keputusan **(Aribowo et al., 2021)**.

2.4.1. Menyiapkan Dataset yang akan Digunakan

Pada tahapan ini data yang akan digunakan adalah *dataset* yang telah dipilih secara acak dengan menggunakan 11 *sample data* yang dapat dilihat pada Tabel 3.

Tabel 3. Sample Data

Gender	***Hypertension***	**smoking_status**	**……..**	**Index_Masa_Tubuh**	**stroke**
Male	tidak	formerly smoked	……..	Obesitas	Ya
Female	tidak	never smoked	……..	Normal	tidak
Male	ya	smokes	……..	Berlebih	tidak
Female	tidak	never smoked	……..	Berlebih	tidak
Female	tidak	formerly smoked	……..	Obesitas	Ya
Female	ya	smokes	……..	Berlebih	tidak
Male	tidak	smokes	……..	Berlebih	Ya
Female	tidak	never smoked	……..	Obesitas	tidak
Male	tidak	never smoked	……..	Dibawah Normal	tidak
Female	tidak	never smoked	……..	Dibawah Normal	tidak
Female	tidak	never smoked	……..	Normal	tidak

2.4.2. Menghitung IndexGini semua atribut

Lalu setelah seluruh *dataset* telah disiapkan hitung seluruh *giniindex* yang dimiliki dari seluruh atribut tersebut menggunakan Persamaan(1) **(Aribowo et al., 2021)** :

$$IndexGini(S) = 1 - \sum_{i=1}^{k} Pi^2 \quad (1)$$

- Diketahui sample data (S) = 11

1. Atribut *Stroke*

$$IndexGini(Semua) = 1 - (\left(\frac{Ya}{Jumlah\ Data}\right)^2 + \left(\frac{Tidak}{Jumlah\ Data}\right)^2)$$
$$IndexGini(Semua) = 1 - (\left(\frac{3}{11}\right)^2 + \left(\frac{8}{11}\right)^2)$$
$$IndexGini(Semua) = 0{,}396694$$

2. Atribut *Gender*

$$IndexGini(Gender, Female) = 1 - \left(\left(\frac{1}{7}\right)^2 + \left(\frac{6}{7}\right)^2\right) = 0{,}244898$$
$$IndexGini(Gender, Male) = 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right) = 0{,}5$$

3. Atribut *Hypertension*

$$IndexGini(Hypertension, Ya) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right) = 0$$
$$IndexGini(Hypertension, Tidak) = 1 - \left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right) = 0{,}444444$$

4. Atribut *Heart Disease*

$$IndexGini(Heart\ Disease, Ya) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 0{,}444444$$
$$IndexGini(Heart\ Disease, Tidak) = 1 - \left(\left(\frac{1}{8}\right)^2 + \left(\frac{1}{8}\right)^2\right) = 0{,}21875$$

5. Atribut Jenis Pekerjaan

$$IndexGini(Jenis\ Pekerjaan, Children) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 0$$

$$IndexGini(Jenis\ Pekerjaan, Govt\ Job) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right) = 0$$

$$IndexGini(Jenis\ Pekerjaan,\ Never\ Worked) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right) = 0$$

$$IndexGini(Jenis\ Pekerjaan, Private) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = 0{,}375$$

$$IndexGini(Jenis\ Pekerjaan,\ Self - Employee) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0{,}5$$

2.4.3. Menghitung GiniGain

Setelah nilai *indexgini* didapatkan maka tahapan selanjutnya yaitu menghitung nilai *GiniGain* menggunakan rumus berikut.

$$GiniGain = Gini(A,S) - \sum_{i=1}^{n} \frac{|Si|}{|S|}\ X\ Gini(Si) \quad (2)$$

1. Atribut *Gender*

$$GiniGain = 0{,}396694 - ((\left(\frac{7}{11}\right).\ 0{,}244898\ ) + (\left(\frac{4}{11}\right).0{,}5))) = 0{,}05930319$$

2. Atribut *Hypertension*

$$GiniGain = 0.396694 - ((\left(\frac{2}{11}\right).\ 0\ ) + (\left(\frac{9}{11}\right).0{,}44444))) = 0{,}0330579$$

3. Atribut *Heart Disease*

$$GiniGain = 0.396694 - ((\left(\frac{3}{11}\right).\ 0{,}44444) + (\left(\frac{8}{11}\right).0{,}21875))) = 0,1163912$$

4. Atribut *Ever Married*

$$GiniGain = 0.3966942 - ((\left(\frac{7}{11}\right).\ 0{,}489796) + (\left(\frac{4}{11}\right).0))) = 0{,}0850059$$

5. Atribut Jenis Pekerjaan

$$GiniGain = 0.396694 - \left(\left(\left(\frac{3}{11}\right).\ 0\ \right) + \left(\left(\frac{1}{11}\right).0\right)\right) + (\left(\frac{1}{11}\right).\ 0\ ) + (\left(\frac{4}{11}\right).0{,}375)) + (\left(\frac{2}{11}\right).0{,}5))) = 0{,}1694215$$

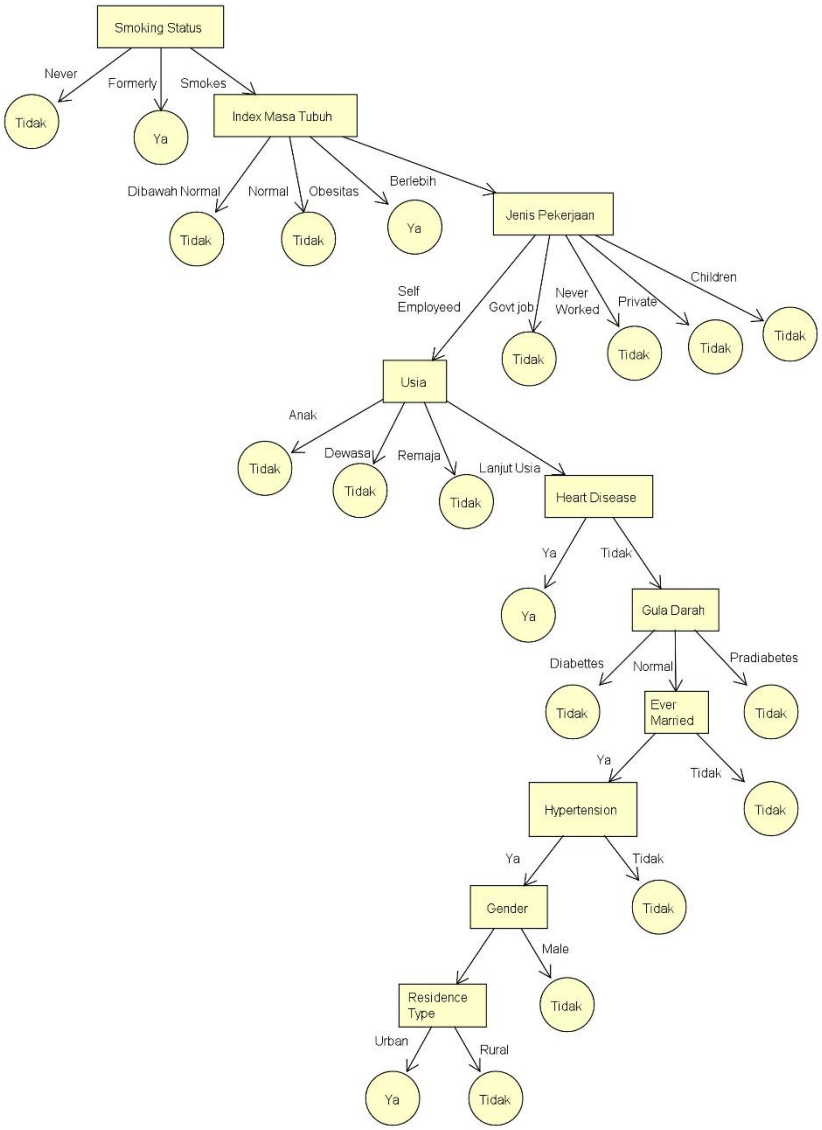
2.4.4. Membuat pohon

Dari hasil perhitungan *GiniIndex* dan *GiniGain* yang sudah dihitung maka nilai *GiniGain* tertinggilah yang akan menjadi node akar. Maka yang akan menjadi node akar yaitu atribut Smoking Status dengan nilai *GiniGain* 0,2754281, lalu untuk lengkapnya dapat dilihat pada Tabel 4.

Tabel 4. Hasil Keseluruhan

	Jumlah Kasus	**Stroke(Ya)**	**Stroke(Tidak)**	**IndexGini**	**GiniGain**
Stroke	11	3	8	0.396694	
Gender					
Female	7	1	6	0.244898	0.0590319

	Jumlah Kasus	**Stroke(Ya)**	**Stroke(Tidak)**	**IndexGini**	**GiniGain**
Male	4	2	2	0.5	
Hypertension					
Ya	2	0	2	0	
Tidak	9	3	6	0.444444	0.0330579
Heart Disaese					
Ya	3	2	1	0.444444	
Tidak	8	1	7	0.21875	0.1163912
Ever Married					
Ya	7	3	4	0.489796	
Tidak	4	0	4	0	0.0850059
Jenis Pekerjaan					
Children	3	3	0	0	
Govt Job	1	0	1	0	
Never Worked	1	0	1	0	
Private	4	1	3	0.375	
Self-Employeed	2	1	1	0.5	0.1694215
Smoking Status					
Never	6	0	6	0	
Formerly	2	2	0	0	
Smokes	3	1	2	0.444444	0.2754821
Usia					
Anak-anak	1	0	1	0	
Remaja	3	0	3	0	
Dewasa	4	1	3	0.375	
Lanjut Usia	3	2	1	0.444444	0.1391185
Gula Darah					
Diabetes	1	0	1	0	
Normal	9	2	7	0.345679	
Pradiabetes	1	1	0	0	0.1138659
Index Masa Tubuh					
Berlebih	4	1	3	0.375	
Dibawah Normal	2	0	2	0	
Normal	2	0	2	0	
Obesitas	3	2	1	0.444444	0.1391185
Residence Type					
Rural	6	1	5	0.277778	
Urban	5	2	3	0.48	0.0269972



Gambar 2. Pohon Keputusan

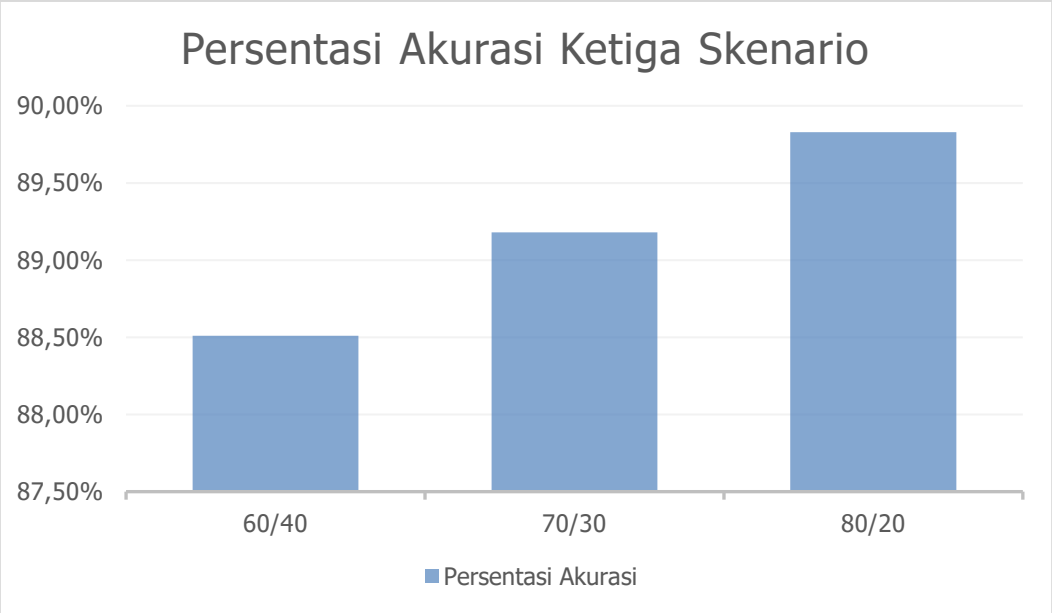
Pohon keputusan ini dihitung berdasarkan Tabel 4. Hasil perhitungan yang telah dilakukan dengan menggunakan 11 data dengan tiga data menggunakan kelas **Ya** dan delapan data menggunakan kelas **Tidak**. Pada akar pohon keputusan berada pada atribut *Smoking status* karena nilai *GiniGain* pada Tabel 4 atribut tersebut memiliki nilai *GiniGain* terbesar, oleh karena itu atribut *smoking status* menjadi akar pohon keputusan.

2.5 Evaluasi

Evaluasi yang dilakukan pada penelitian ini yaitu dengan menerapkan evaluasi *Confusion Matrix* **(Xu et al., 2020)**. *Confusion matrix* dilakukan untuk mengetahui kemampuan dari model algoritma klasifikasi yang telah dibuat yaitu algoritma CART, *confusion matrix* dilakukan untuk mengukur tingkat akurasi dari algoritma CART.

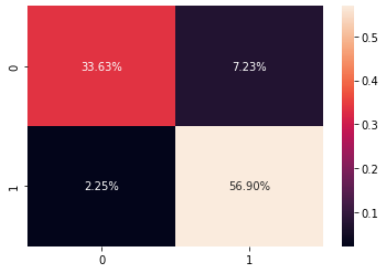
3. HASIL DAN PEMBAHASAN

Hasil dari pengujian model yang telah dilakukan didapatkan hasil dari akurasi, *precision, recall* dan *f1 score* untuk masing-masing skenario *split dataset* yang dilakukan, dengan menggunakan 3 skenario *split dataset* yaitu skenario 60% data latih dan 40% data uji, 70% data latih dan 30% data uji, 80% data latih dan 20% data uji.

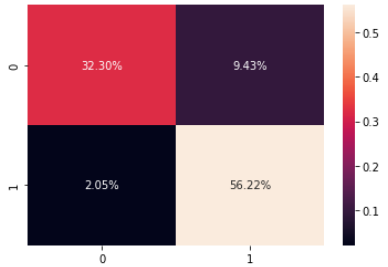


3.1 Skenario Dataset Rasio 60% dan 40%

Skenario pengujian model pertama yaitu *split data* untuk data latih 60% dan data uji 40% dengan pembagian 4898 data latih dan 3266 data uji. Hasil *confusion matrix* untuk data latih dan data uji dapat dilihat pada gambar 3 dan 4 dan untuk perhitungan akurasi dapat dilihat pada Tabel 5.



Gambar 3. *Confusion Matrix* Data Latih



Gambar 4. *Confusion Matrix* Data Uji

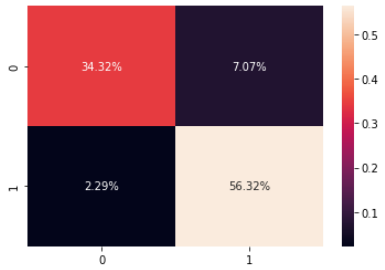
Tabel 5. Skenario Pengujian Model 1

Skenario Split Data 60% dan 40%				
Nilai	**Data Latih**	**Data Uji**	**Data Latih *Pruning***	**Data Uji *Pruning***
Accuracy	90,52%	88,51%	75,41%	73,86%
Recall	96,20%	96,47%	85,41%	84,33%
Precision	88,72%	85,63%	76,01%	74,42%
F1 Score	92,31%	90,73%	80,43%	79,07%

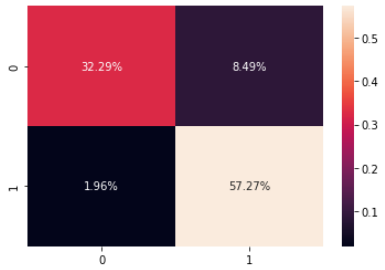
Dari pengujian model didapatkan hasil data uji sebesar 88,51% dengan hasil *confusion matrix* pada nilai *True Positive* sebesar 1064 data dan pada nilai *False Negative* sebesar 1856 data. Dimana nilai tersebut belum terlalu baik, karena pada skenario yang lain akurasi pada data uji lebih tinggi dibandingkan dengan skenario pertama dikarenakan data latih yang digunakan lebih sedikit.

3.2. Skenario Dataset Rasio 70% dan 30%

Skenario pengujian model kedua yaitu *split data* untuk data latih 70% dan data uji 30% dengan pembagian 5714 data latih dan 2450 data uji. Hasil *confusion matrix* untuk data latih dan data uji dapat dilihat pada gambar 5 dan 6 dan untuk perhitungan akurasi dapat dilihat pada Tabel 6.



Gambar 5. *Confusion Matrix* Data Latih



Gambar 6. *Confusion Matrix* Data Uji

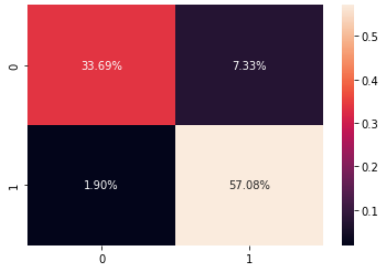
Tabel 6. Skenario Pengujian Model 2

Skenario Split Data 70% dan 30%				
Nilai	**Data Latih**	**Data Uji**	**Data Latih *Pruning***	**Data Uji *Pruning***
Accuracy	90,63%	89,18%	75,41%	74,73%
Recall	96,08%	96,24%	85,41%	78,36%
Precision	88,84%	86,76%	76,01%	79,17%
F1 Score	92,32%	91,25%	80,43%	78,76%

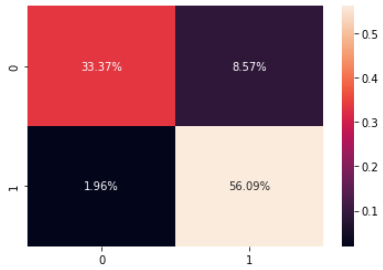
Pada pengujian model skenario kedua ini memiliki nilai akurasi yang cukup baik yaitu sebesar 89,81% dengan hasil *confusion matrix* dengan nilai *True Positive* sebesar 789 data dan nilai *False Negative* sebesar 1402 data. Tetapi, nilai akurasi dari skenario ini belum lebih baik dari akurasi skenario berikutnya.

3.3 Skenario Dataset Rasio 80% dan 20%

Skenario pengujian model pertama yaitu *split data* untuk data latih 80% dan data uji 20% dengan pembagian 6531 data latih dan 1633 data uji. Hasil *confusion matrix* untuk data latih dan data uji dapat dilihat pada gambar 7 dan 8 dan untuk perhitungan akurasi dapat dilihat pada Tabel 7.



Gambar 7. *Confusion Matrix* Data Latih



Gambar 8. *Confusion Matrix* Data Uji

Tabel 7. Skenario Pengujian Model 3

Skenario Split Data 80% dan 20%				
Nilai	**Data Latih**	**Data Uji**	**Data Latih *Pruning***	**Data Uji *Pruning***
Accuracy	90,76%	89,83%	75,07%	73,95%
Recall	92,51%	96,62%	84,97%	84,71%
Precision	88,61%	86,74%	75,98%	74,39%
F1 Score	92,31%	91,41%	80,23%	79,22%

Pengujian model diatas didapatkan nilai akurasi terbesar dari semua skenario dengan nilai akurasi 89,83% dengan hasil nilai *confusion matrix* pada nilai *True Positive* sebesar 541 data dan nilai *False Negative* sebesar 926 data. Nilai akurasi pada skenario ini merupakan nilai akurasi terbesar dari skenario-skenario sebelumnya, karena pada skenario ini menggunakan data latih yang lebih besar dari skenario sebelumnya.

Untuk *pruning* pada semua skenario yang telah dilakukan, tidak ada pengaruh baik pada peningkatan akurasi. Karena pada saat dilakukan *pruning* data yang digunakan menjadi lebih sedikit, maka pengaruh pada akurasi menjadi tidak lebih baik.

4. KESIMPULAN

Berdasarkan hasil eksperimen yang telah dilakukan dengan menggunakan tiga skenario *split data* yang dilakukan dapat disimpulkan bahwa Prediksi Awal Penyakit Stroke Berdasarkan Rekam Medis Menggunakan algoritma *Classification and Regression Tree* (CART) ini menghasilkan akurasi tertinggi yaitu sebesar 89,83% pada skenario *split data* untuk data latih 80% dan data uji 20%. Setelah dilakukan analisis terhadap eksperimen yang telah dilakukan didapatkan hasil analisis, lebih besar data latih maka akurasi yang didapat akan lebih besar, karena nantinya pada evaluasi yang dilakukan oleh *confusion matrix* nilai *true positive* dan nilai *true negative* akan lebih besar pada skenario dataset yang lebih besar. Hal tersebut akan mempengaruhi pada nilai akurasi karena nilai *true positive* adalah nilai dari prediksi positif dan itu benar dan nilai *true negative* adalah nilai prediksi negatif yang salah, maka dari itu nilai akurasi terbesar ada pada skenario dataset yang terbesar juga.

UCAPAN TERIMA KASIH

Penulis ingin mengucapkan terima kasih kepada Author dari situs web Kaggle.com yaitu Federico Soriano Palacios yang telah membagikan dataset yang penulis gunakan pada penelitian ini. Semoga penelitian ini dapat memberikan kontribusi pada perkembangan ilmu pengetahuan khususnya pada bidang teknologi.

DAFTAR RUJUKAN

Anggraeni, H. D., Saputra, R., & Noranita, B. (2013). Aplikasi Data Mining Analisis Data Transaksi Penjualan Obat Menggunakan Algoritma Apriori. *Journal of Informatics and Technoligy*, *2*(2), 22–28. https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

Aribowo, A., Kuswandhie, R., & Primadasa, Y. (2021). Penerapan dan Implementasi Algoritma CART Dalam Penentuan Kelayakan Penerima Bantuan PKH Di Desa Ngadirejo. *CogITo Smart Journal*, *7*(1), 40. https://doi.org/10.31154/cogito.v7i1.293.40-51

Arieska, P. K., & Herdiani, N. (2018). Pemilihan Teknik Sampling Berdasarkan Perhitungan Efisiensi Relatif. *Jurnal Statistika*, *6*(2), 166–171.

Bima, S. A., Setiawan, A., & Mahatma, T. (2013). *Pembentukan Sampel Baru Yang Masih Memenuhi Syarat Valid Dan Reliabel Dengan Teknik Resampling*. *October*, 1-5.

Byna, A., & Basit, M. (2020). Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *9*(3), 407–411. https://doi.org/10.32736/sisfokom.v9i3.1023

Indah Prabawati, N., Widodo, & Ajie, H. (2019). Kinerja Algoritma Classification And Regression Tree (Cart) dalam Mengklasifikasikan Lama Masa Studi Mahasiswa yang Mengikuti Organisasi di Universitas Negeri Jakarta. *PINTER : Jurnal Pendidikan Teknik*

Informatika Dan Komputer, *3*(2), 139–145. https://doi.org/10.21009/pinter.3.2.9

Nafi'iyah, N. (2015). *Algoritma Cart Dalam Penentuan Pohon Keputusan*. *7*(2), 41-47.

Pratama, F. A., Narasati, R., & Amalia, D. R. (2019). *Pengaruh Kata Cashback Terhadap Peningkatan Penjualan Menggunakan Data Mining*. *3*(2), 1–5.

Pratiwi, F. E., & Zain, I. (2014). Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara. *Jurnal Sains Dan Seni Pomits*, *3*(1), D54–D59. http://www.ejurnal.its.ac.id/index.php/sains_seni/article/view/6129

Rahayu, E. S., Wahono, R. S., & Supriyanto, C. (2015). Penerapan Metode Average Gain, Threshold Pruning dan Cost. *Journal of Intelligent Systems*, *1*(2), 91–97.

Subarkah, P., Ikhsan, A. N., & Setyanto, A. (2018). The effect of the number of attributes on the selection of study program using classification and regression trees algorithms. *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, 1–5. https://doi.org/10.1109/ICITISEE.2018.8721030

Sulastri, H., & Gufroni, A. I. (2017). Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *3*(2), 299–305. https://doi.org/10.25077/teknosi.v3i2.2017.299-305

Tjandrawinata, R. (2016). *Industri 4.0: revolusi industri abad ini dan pengaruhnya pada bidang kesehatan dan bioteknologi*. *February*. https://doi.org/10.5281/zenodo.49404, 31-39.

Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064

Zhang, B., Wei, Z., Ren, J., Cheng, Y., & Zheng, Z. (2018). An Empirical Study on Predicting Blood Pressure Using Classification and Regression Trees. *IEEE Access*, *6*, 21758–21768. https://doi.org/10.1109/ACCESS.2017.2787980