KLASIFIKASI RISIKO KEMATIAN PASIEN BERDASARKAN PENYAKIT PENYERTA DAN USIA PASIEN MENGGUNAKAN METODE C4.5

Fery Bayu Aji[1], Fajri Rakhmat Umbara[2], Fatan Kasyidi[3]

[1,2,3] Program Studi Teknik Informatika, Universitas Jenderal Achmad

Jl. Terusan Jend. Sudirman, Cibeber, Kec. Cimahi Sel., Kota Cimahi, Jawa Barat 40531
[1] ferybayuaji18@if.unjani.ac.id, [2] fajri.rakhmat@lecture.unjani.ac.id, [3] fatan.kasyidi@lecture.unjani.ac.id

Abstract

The high mortality rate in people suffering from certain diseases, so we conducted this study. To find out patients with high or low risk of death. So the risk classification of patients' mortality based on age and comorbidities was carried out using the decision tree algorithm c4.5. The C4.5 algorithm is used to find patterns to determine which class the data belongs to by using training data and test data used for evaluation. The results of the accuracy test using the Confusion Matrix with split data, 70% training data 30% test data produces 72% while with split data, 80% training data 20% test data produces 70% accuracy. By using the split data model, 70% training data of 30% test data is superior to 80% training data of 20% test data. While the accuracy produced by the tree cutting or pruning method can reduce accuracy with an accuracy result of 66%. And based on the resulting decision tree without pruning it produces a tree depth of 11 levels

Keywords : *Classification, C4.5, Decision Tree*

Abstrak

Tingginya angka kematian pada orang yang menderita penyakit tertentu, maka kami melakukan penelitian ini. Untuk mengetahui pasien dengan risiko kematian tinggi atau rendah. Maka dilakukannya klasifikasi risiko kematian pasien berdasarkan usia dan penyakit penyerta pasien dengan menggunakan decision tree algoritma c4.5. Algoritma C4.5 digunakan untuk menemukan pola menentukan data tersebut masuk kedalam kelas mana dengan menggunakan data latih dan data uji digunakan untuk evaluasi. Hasil pengujian akurasi menggunakan Confusion Matrix dengan pembagian data, data latih 70% data data uji 30% mengghasilkan 72% sementara dengan pembagian data, data latih 80% data data uji 20% menghasilkan akurasi sebesar 70%. Dengan menggunakan model pembagian data, data latih 70% data data uji 30% lebih unggul dibanding data latih 80% data data uji 20%. Semantara akurasi yang dihasilkan dengan metode pemotongan pohon atau pruning dapat menurunkan akurasi dengan hasil akurasi sebesar 66%. Dan berdasarkan pohon keputusan yang dihasilkan tanpa pruning menghasilkan kedalaman pohon sebanyak 11 level.

Kata kunci : *Klasifikasi, C4.5, Decision Tree*

1. PENDAHULUAN

Noncommunicable diseases (NCDs) atau Penyakit tidak menular yaitu penyebab utama kematian seluruh dunia. Setiap tahunnya dari 70% dari semua kematian di seluruh dunia orang meninggal karena Stroke, Kanker, Serangan Jantung, Penyakit pernapasan kronis, Diabetes atau Gangguan mental[1]. Penyakit kardiovaskular menyumbang 30% jumlah semua kematian di Amerika, Penyakit Kanker menyumbang 22% dari semua kematian di Amerika, Penyakit Pernapasan kronis menyumbang 9% dari semua kematian dan Diabetes menyumbang 3% dari semua kematian di Amerika 24% lainnya bukan karena Penyakit Tidak Menular(PTM) dan 75% dari kematian terjadi pada mereka berusia 30 sampai 69 tahun[2].

Karena tingginya angka kematian pada orang yang menderita penyakit tertentu, maka kami melakukan penelitian ini. Untuk membantu tim medis dalam melakukan perawatan intensif dalam penanganannya. Untuk mengetahui

pasien dengan risiko kematian tinggi atau rendah. Maka dilakukannya prediksi menentukan pasien dengan risiko kematian tinggi atau rendah berdasarkan penyakit dan usia pasien tersebut. Maka tujuan umum dari penelitian ini untuk mengetahui akurasi pada prediksi risiko kematian pada pasien.[3]

Adapun penelitian sebelumnya "Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units" memprediksi kematian pasien di *ICU* menggunakan metode *Random Forest* membantu untuk memprediksi kematian pasien *ICU*. Namun keterbatasan penelitian tersebut karena kurang banyaknya data pasien [4]. Perbedaan metode pada penelitian ini menggunakan metode klasifikasi C4.5 sementara penelitian sebelumnya menggunakan *Random Forest.*

Teknik yang digunakan dalam penelitian ini memprediksi menggunakan *Decision tree* dengan algoritma C4.5. Karena pada penelitian sebelumnya belum ada yang menggunakan metode C4.5 menggunakan dataset tersebut. Selain ini algoritma C4.5 ini memiliki kelebihan jika di bandingkan dengan metode ID3 dan CART yaitu memiliki kemampuan yang tidak membatasi cabang dalam bentuk data biner[5] dan algoritma c4.5 ini juga memiliki kelebihan dapat menghasilkan pohon keputusan yang mudah diimplementasikan dan efisian dalam menangani atribut[6].

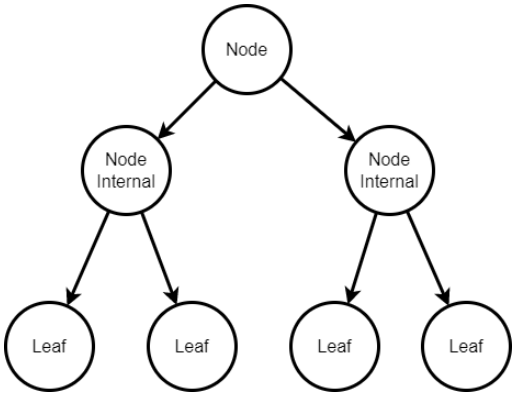
Secara umum metode ini merupakan metode yang baik untuk memprediksi memberikan akurasi tinggi. *Decision tree* c4.5 ini menghasilkan akurasi yang lebih baik jika dibandingkan dengan model klasifikasi yang lainnya [7].

Penelitian ini telah dibangun sebuah sistem yang dapat mengklasifikasikan Risiko kematian pasien berdasarkan penyakit penyerta dan usia pasien menggunakan algoritma C4.5. Yang menghasilkan kelas risiko kematian tinggi dan rendah.

Pada penelitian sebelumnya memprediksi risiko kematian pasien stroke menghasilkan akurasi dengan C4.5 yaitu 90,5% dan Logistic Regression 88.8%, Support Vector Machine86% dan Random Forest 89.6% dengan menggunakan dataset pasien penderita penyakit stoke namun pada penelitian ini keterbatasan data yang kurang banyaknya dan dataset yang digunakan terbatas pada prediksi risiko kematian pada pasien stroke sementara penelitian ini menggunakan data sebanyak 2034[7]. Sementara penelitian memprediksi kematian akibat penyakit gagal jantung dengan algoritma *K-Nearest Neighbord* menghasilkan nilai akurasi 94.92% dengan rapid miner dan 68% dengan mengunakan bahasa *python* dengan menggunakan dataset pasien penderita penyakit gagal jantung. Penelitian ini dibatasi oleh menggunakan dataset sebanyak 297 data dimana pada penelitian ini tidak menggunakan algoritma C4.5 sehingga dilakukan penelitian ini menggunakan metode Decision Tree C4.5 [8].

1.1. Decision Tree C4.5

Decision tree adalah struktur pohon seperti diagram alur yang memiliki *node* dan *leaf* , dimana setiap cabang menunjukkan pola pada nilai atribut, setiap cabang mewakili hasil tes, dan daun pohon tersebut diartikan sebagai kelas atau distribusi kelas. Sebagai contoh Algoritma *Decision tree* yang ada seperti ID3, C4.5 dan CART. Membangun struktur seperti diagram alur dimana setiap *node internal* (*nonleaf*) menunjukkan hasil pengujian pada atribut, masing-masing cabang sesuai dengan hasil pengujian, dan setiap simpul eksternal (*leaf*) menunjukkan prediksi kelas. [9]



Gambar 1. Decision Tree

Algoritma C4.5 termasuk kedalam *Decision tree* yaitu membuat pohon keputusan dengan mempunyai percabangan sampai aturannya tepenuhi. Pohon keputusan dapat diartikan sebagai suatu cara untuk membagi sekumpulan data menjadi himpunan-himpunan yang lebih kecil dengan menerapkan serangkaian *rule* atau aturan keputusan. Pohon keputusan dibangun dengan membagi data secara rekursif sehingga setiap bagianya terdiri dari data yang berasal dari kelas yang sama. Algoritma C4.5 dapat menangani data numerik baik yang bersifat diskrit maupun kontinyu. Jika suatu data set mempunyai beberapa nilai pengamatan *missing value*, jika

jumlah pengamatan terbatas maka atribut *missing value* ini dapat diganti dengan nilai rata-rata dari variabel yang bersangkutan.[10]

Tentukan akar pohon dengan menghitung nilai *gain* tertinggi setiap atribut atau berdasarkan nilai index entropy terendah. [11]

Menghitung nilai *index* dari *entropy*:

$$Entropy(s) = \sum_{i=1}^{n} -pi\ .\ Log_2 pi \quad (1)$$

Keterangan :
s = himpunan Kasus
n = jumlah partisi atribut A
Pi = proposi Si terhadap S

Menghitung nilai gain dengan rumus berikut:

$$Gain(S,A) = Entropy(s) \sum_{i=1}^{n} \frac{si}{s}.Emtropy(s) \quad (2)$$

Keterangan :
S = himpunan Kasus
A = Atribut
n = jumlah partisi atribut A
Si = jumlah kasus pada partisi ke-i

Selanjutnya menghitung *Split Information*

$$SplitInformation = -\sum_{i=1}^{n} \frac{S1}{s} Log2 \frac{S1}{S} \quad (3)$$

Keterangan :
s = himpunan Kasus
n = jumlah partisi atribut A
Si = jumlah kasus pada partisi ke-i

Kemudian menghitung *gain ratio*

$$\text{Gainratio(S, A)} = \frac{Gail(S,A)}{SplitInformation(S,A)} \quad (4)$$

Keterangan :
S = himpunan Kasus
A = Atribut

1.2. Simple Random Sampling

Simple random Sampling adalah Teknik untuk menduplikasi data sesuai label dari data minoritas. Data tersebut direplika secara acak atau bisa disebut data sintetis. *Simple random Sampling* digunakan untuk menyeimbangkan dataset yang mengalami *imbalance data.* Hal ini berguna untuk memperbaiki kualitas klasifikasi. [12] Pemilihan metode *Simple random Sampling* karena *Simple random Sampling* ini dapat menangani dataset yang *imbalance data.* Kelebihan metode *Simple random Sampling* ini metode *Simple random Sampling* ini dapat menangani dataset yang bias dan cocok digunakan untuk metode decision tree. Teknik yang digunakan *Simple random sampling*adalah dengan cara menduplikat label minoritas sehingga sama dengan label mayoritas. Proses menduplikasi dataset ini dilakukan dengan cara acak berdasarkan label yang akan diduplikasi sehingga dataset tidak *overfitting*. [13].

1.3. Confusion Matrix

Pengujian model digunakan untuk mengukur akurasi sistem dalam melakukan klasifikasi dengan hasil yang dikeluarkan berdasarkan dengan yang diinginkan. Berdasarkan karakteristik hasil klasifikasi c4.5. Pengujian model disini menggunakan *Confusion matrix*. *Confusion matrix* digunakan untuk mengevaluasi model klasifikasi untuk mengukur objek yang benar atau salah pada hasil yang yang dihasilkan[14] Contoh perhitungan *confusion matrix* ditunjukan pada tabel 1[15]

TABEL 1. CONFUSION MATRIX

Confusion Matrix		Prediksi	
		1	0
Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan:
True positive (TP) = jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1
True Negative (TN)= jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0
False Positive (FP) = jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1
False Negative (FN) = jumlah dokumen dari kelas 1yang salah diklasifikasikan sebagai kelas 0

Rumus perhitungan accuracy, precision dan recall adalah sebagai berikut:

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

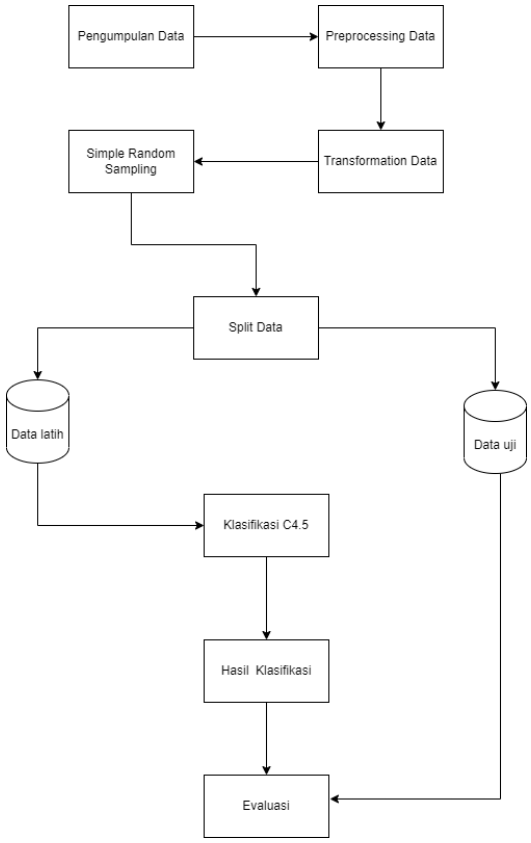
$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = 2\frac{\text{precision x recall}}{\text{precision+recall}} \quad (4)$$

2. METODOLOGI PENELITIAN

3.1. Skema Alur Penelitian

Metode Penelitian pada penelitian ini berdasarkan gambar 1yaitu terdiri dari Pengumpulan data, Preprocessing Data, Transformasi Data, Penanganan data yang tidak seimbang dengan Simple Random Sampling, Melakukan pembagianmenjadi data uji dan data latih, klasifikasi C4.5, Hasil Klasifikasi, dan terakhir melakukan evaluasi menggunakan Confusion Matrix



Gambar 2. Metode Penelitian

3.2. Pengumpulan Data

Pengumpulan data diambil dari www.kaggle.com, data berasal dari Rumah Sakit ICU Beth Israel Deaconess Medical Center, Boston, AS. Data diunggah pada September 2021. Dalam data tersebut terdapat atribut Outcome, Age, Hypertensive, Atrialfibrillation, CHD, Diabetes , Deficiencyanemias, Depression, Hyperlipemia, Renal failure, COPD. Atribut tersebut digunakan sebagai dataset yang selanjutnya dimasukan kedalam Preprocessing data.

3.3. Preprocessing

Preprocessing ini memilih atribut yang akan digunakan dalam penelitian ini. Atribut itu adalah atribut Outcome(hasil keluaran), Age(Umur pasien), Hypertensive (Tekanan darah tinggi), Atrialfibrillation (Denyut jantung tidak teratur), CHD (Coronary heart disease atau Penyakit Jantung Korone), Diabetes (Meningkat kadar gula darah), Deficiencyanemias (Kekurangan zat besi),

TABEL 2. ATRIBUT

No	Nama Atribut	Keterangan
1	Outcome	Hasil keluaran Pasien
2	Age	Umur pasien
3	Hypertensive	tekanan darah terlalu tinggi.
4	Atrialfibrillation	Denyut jantung tidak beraturan dan lebih cepat di atas normal
5	CHD with no MI	Penyakit Jantung Koroner.
6	diabetes	Kelebihan kadar gula darah
7	deficiencyanemias	kekurangan zat besi.
8	depression	Gangguan mental
9	Hyperlipemia	ketidakseimbangan lemak dalam darah.
10	Renal failure	Gagal ginjal.
11	COPD	Penyakit yang dapat menghalangi aliran udara pada paru-paru

3.4. Transformasi Data

Transformasi data mengubah atribut Age dari numerik menjadi nominal berdasarkan kategori umur yang dikeluarkan oleh United States Sensus yaitu anak (*children*): 0-17 tahun,

Dewasa (*adult*): 18-64 tahun, Lanjut usia (*elderly*): di atas 65 tahun

TABEL 3. TRANSFORMASI DATA

No	Umur	Kategori
1	0-17	1
2	18-64	2
3	65++	3

3.5. Simple Random Sampling

Simple random Sampling digunakan setelah menambahkan dataset. Kemudian dataset dihitung berapa jumlah label positif atau meninggal dan jumlah negatif atau tidak meninggal. Metode ini yaitu Teknik random sederhana dengan penambahan data dengan data sintetik secara random yang selanjutnya akan digunakan klasifikasi algoritma dengan C4.5. Metode Simple random Sampling ini data ditambahkan berdasarkan label yang tidak seimbang dengan cara diacak menghasilkan data sintetik baru. Setelah dihitung dengan hasil tabel seperti di bawah ini :

Keterangan Label :
1 = Positif meninggal
0 = Negatif meninggal

TABEL 4. JUMLAH LABEL DATASET

Label	Jumlah data
0	1017
1	160

Kemudian dilakukan penyeimbangan dataset dengan hasil sebagai berikut

TABEL 5. PENYEIMBANGAN DATASET

Label	Jumlah data
0	1017
1	1017

Penyeimbangan dataset ini digunakan untuk memperbaiki model sehingga tidak *overfitting*.

3.6. Split Data

Split data berasal dari dataset yang dimasukan kepada model. Kemudian data tersebut mengalami pembagian data atau split data salah satu cara evaluasi membagi data latih dan data uji. Data latih digunakan untuk melatih algoritma dalam mencari pola dan untuk membuat model pada klasifikasi c4.5 yang kemudian di uji menggunakan data uji. Data uji ini digunakan untuk mengukur atau mengetahui performa yang didapat dari tahapan uji dan untuk mengukur tingkat akurasi dari algoritmadan dataset tersebut menggunakan model ini. Pada tahap data ini dataset dibagi 70%-30% dengan pembagian Data latih 70% sebanyak sebanyak 1420 data dan data uji 30 sebanyak sebanyak 614 data

3.7. Klarifikasi C4.5

Klasifikasi disini menggunakan algoritma c4.5 dimana meggunakan sampel data sebanyak 15 data

TABEL 6. SAMPEL DATASET

No	Outcome	Categories Age	...	Renal failure	COPD
1	0	2	...	0	1
2	0	2	...	1	0
3	1	2	...	1	0
4	0	2	...	1	0
5	1	2	...	0	0
6	0	2	...	0	0
7	0	2	...	0	0
8	1	3	...	1	1
9	1	3	...	1	0
10	0	3	...	0	0
11	0	3	...	1	0
12	0	3	...	1	0
13	0	3	...	0	0
14	0	3	...	1	1
15	1	3	...	1	0

Keterangan:
1 = Positif
0 = Negatif

Menghasilkan perhitungan Entropy setiap atribut di tabel 7. Dimana Outcome sebagai Entropy total.

TABEL 7. PERHITUNGAN ENTROPY

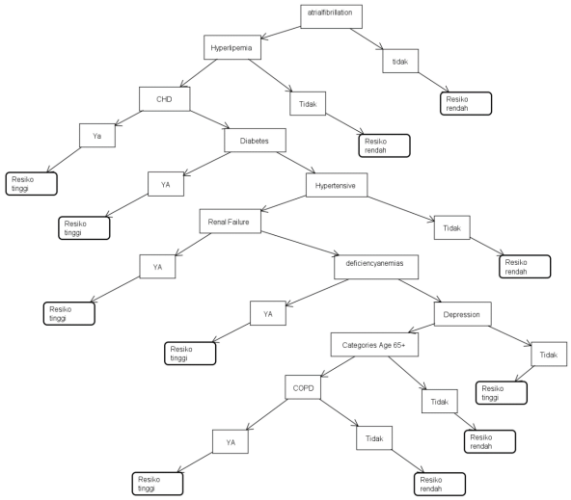
Atribut	Entropy
Outcome	0,918295834
Categories Age	
1(umur 0-17)	0
2(Umur 18-64)	0,863120569
3(Umur 65++)	0,954434003
Hypertensive	
ya (1)	0,89049164
tidak(0)	1
Atrialfibrillation	
ya (1)	0,918295834
tidak(0)	0,503258335
CHD with no MI	
ya (1)	0,970950594
tidak(0)	0,881290899
Diabetes	
ya (1)	0,918295834
tidak(0)	0,918295834
Deficiencyanemias	
ya (1)	0,811278124
tidak(0)	0,985228136
Depression	
ya (1)	0,811278124
tidak(0)	0,945660305
Hyperlipemia	
ya (1)	0,918295834
tidak(0)	0,918295834
Renal failure	
ya (1)	0,99107606
tidak(0)	0,650022422
COPD	
ya (1)	0,918295834
tidak(0)	0,839598958

Menghasilkan perhitunganGain, Split Info dan Gain Ratio bisa dilihat di tabel 8.

TABEL 8. GAIN, SPLIT INFO DAN GAIN RATIO

Atribut	Gain	Split info	Gain Ratio
Categories Age	0.006474	0,99679	0,006495
Hypertensive	0,013203	0,56650	0,023306
Atrialfibrillation	0,249022	0,97095	0,256472
CHD with no MI	0,007118	0,82135	0,008666
diabetes	0	0,97095	0
deficiencyanemias	0,025841	0,99679	0,025924
depression	0,008470	0,836640	0,010124
Hyperlipemia	0	0,721928	0
Renal failure	0,063641	0,97095	0,06554
COPD	0,062957	0,72192	0,08720

Perhitungan entropy, gain, split info dan gain ratio menghasilkan Atribut atrialfibrillation menjadi yang terbesar maka digunakanan sebagai node 1.1 begitu selanjutnya. Sehingga menghasilkan pohon keputusan seperti gambar 3



Gambar 3. Pohon Keputusan Sampel Data

3. HASIL DAN PEMBAHASAN

Evaluasi model ini menggunakan confusion matrix. Confusion matrix adalah salah satu cara untuk melihat akurasi yang dihasilkan oleh model klasifikasi. Pengujian akurasi model ini untuk mengetahui accuracy, precision, recall dan F1-Score. Pengujian ini menggunakan 10 kali percobaan dengan hasil pengujian model ini

dibagi menjadi 2 yaitu pengujian dengan split data 70-30 dan 80-20 menghasilkan evaluasi dengan akuarasi tertinggi sebagai berikut:

3.1 Evaluasi dengan jumlah split data 70-30

Pengujian dengan jumlah split data 70-30 menggunakan confusion matrix menghasilkan confusion matrix seperti di tabel sebagai berikut dengan *confusion matrix*dapat menghasilkan *Accuracy precision recall* dan *f1-Score* dengan masing masing nilai sebagai berikut

TABEL 9. CONFUSION MATRIX SPLIT DATA 70-30

	Predicted : positif	Predicted : positif
Actual : positif	193	130
Actual : positif	40	248

Dengan tabel *confusion matrix* seperti di tabel 9 makan menghasilkan *accuracy, precision* dan *recall* sebagai berikut

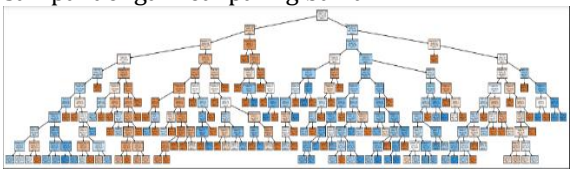
$$\text{Accuracy} = \frac{TP+TN}{Total} = \frac{193+248}{193+130+40+248} = 0.72$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{193}{193+130} = 0.59$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{193}{193+40} = 0.82$$

$$\text{F1 Score} = 2\,\frac{\text{recall x precision}}{\text{recall+precision}} = 2\,\frac{0.82 \times 0{,}59}{0.82+0.59} = 0.68$$

Dan menghasilkan pohon keputusan seperti berikut dimana menghasilkan kedalaman model mencapai level 11. Kedalaman model ini dihitung dari banyaknya level node dari root sampai dengan leaf paling bawah



Gambar 4. Pohon Keputusan Split Data 70-30

3.2 Evaluasi dengan jumlah split data 80-20

Pengujian dengan jumlah split data 80-20 menggunakan *confusion matrix* menghasilkan *confusion matrix* seperti di tabel 10sebagai berikut dengan *confusion matrix* dapat menghasilkan *Accuracy precision recall* dan *f1-Score* dengan masing masing nilai sebagai berikut

TABEL 10. CONFUSION MATRIX SPLIT DATA 80-20

	Predicted : positif	Predicted : positif
Actual : positif	174	122
Actual : positif	87	228

Dengan tabel confusion matrix seperti di tabel 9 maka menghasilkan accuracy, precision dan recall sebagai berikut

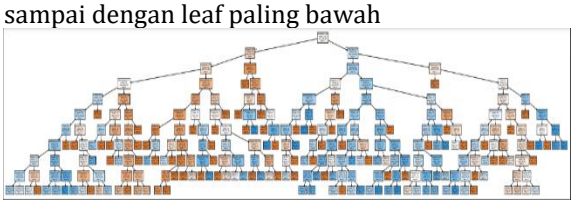
$$\text{Accuracy} = \frac{TP+TN}{Total} = \frac{126+159}{126+75+48+159} = 0.70$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{126}{126+75} = 0.62$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{126}{126+48} = 0.72$$

$$\text{F1 Score} = 2\,\frac{\text{recall x precision}}{\text{recall+precision}} = 2\,\frac{0.62 \times 0.72}{0.62+0.72} = 0{,}666$$

Dan menghasilkan pohon keputusan sebagai berikut menghasilkan kedalaman model mencapai level 11. Kedalaman model ini dihitung dari banyaknya level node dari root sampai dengan leaf paling bawah



Gambar 5. Pohon Keputusan Split Data 80-20

3.3 Evaluasi Confusion Matrix dengan pruning

Pengujian dengan jumlah split data 80-20 menggunakan confusion matrix menghasilkan confusion matrix seperti di tabel 5 sebagai berikut dengan confusion matrix dapat menghasilkan *Accuracy precision recall* dan *f1-Score* dengan masing masing nilai sebagai berikut

Evaluasi dengan metode pruning menggunakan *confusion matrix* menghasilkan *confusion matrix* seperti di tabel sebagai berikut dengan *confusion matrix* dapat menghasilkan *Accuracy precision recall* dan *f1-Score* dengan masing masing nilai sebagai berikut

TABEL 11. CONFUSION MATRIX PRUNING

	Predicted : positif	Predicted : positif
Actual : positif	174	122
Actual : positif	87	228

Dengan tabel *confusion matrix* seperti di tabel 11 makan menghasilkan *accuracy, precision* dan *recall* sebagai berikut

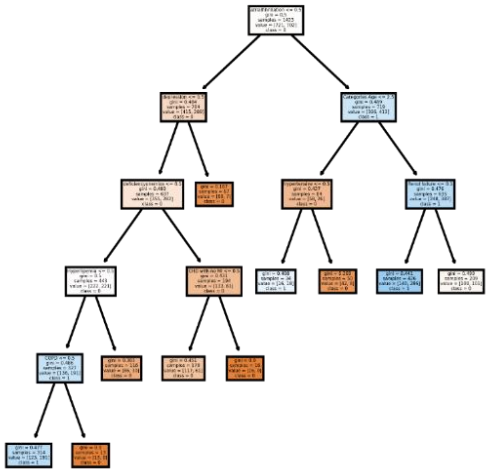
$$\text{Accuracy} = \frac{TP+TN}{Total} = \frac{174+228}{174+122+87+228} = 0.66$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{174}{174+122} = 0.58$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{174}{174+87} = 0.66$$

$$\text{F1 Score} = 2\,\frac{\text{recall x precision}}{\text{recall+precision}} = 2\,\frac{0.66 \text{ x } 0.58}{0.66+0.58} = 0{,}61$$

Menghasilkan pohon keputusan sebagai berikut. Menghasilkan kedalaman model sampai dengan level 6 dimana kedalaman model ini dihitung dari banyaknya level node dari root sampai dengan leaf paling bawah



Gambar 6. Pohon Keputusan Dengan Pruning

4. Kesimpulan dan Saran

Kesimpulan pada penelitian ini yaitu hasil klasifikasi risiko kematian pasien menggunakan algoritma C4.5 menghasilkan 2 label yaitu risiko dengan kematian tinggi dan risiko kematian rendah. Pada penelitian ini akurasi yang dihasilkan, menghasilkan evaluasi dengan menggunakan *confusion matrix* dengan akurasi yang cukup baik yaitu dengan pembagian data sebanyak 70-30 menghasilkan akurasi sebesar 72% lebih baik dibandingan dengan pembagian data 80-20 menghasilkan akurasi sebesar 70%. Dan menggunakan metode pruning menghasilkan akurasi sebesar 66%. Dan dapat disimpulkan jika pembagian data sebesar 70 data latih dan 30 data uji menghasilkan akurasi yang lebih baik jika dibandingkan dengan pembagian data 80 data latih dan 20 data uji, perbedaan yang dihasilkan sebanyak 2% . Semantara akurasi yang dihasilkan dengan metode pemotongan pohon atau pruning dapat menurunkan akurasi dengan hasil akurasi sebesar 66%. Karena pruning yaitu salah satu teknik pemotongan pohon yang sebelum menemukan labelnya sehingga Teknik pruning dapat mengurasi akuasi dari model tersebut Dan berdasarkan pohon keputusan yang di hasilkan menghasilkan kedalaman pohon sebanyak 11 level.

Saran untuk penelitian selanjutnya, bahwa perlu menambah atribut yang digunakan yaitu jumlah penyakit yang berbeda atau rekam medis pasien yang lainnya dan menambah jumlah dataset yang digunakan sehingga diharapkan dapat menghasilkan model yang lebih baik. Dan menggunakan metode algoritma yang lain untuk membandingkan metode yang lebih baik menggunakan model dataset ini.

5. UCAPAN TERIMA KASIH

Ucapan terima kasih kepada website Kaggle yang telah menyediakan dataset serta ikut membantu dan mendukung penelitian ini.

Daftar Pustaka:

[1] E. Chigom, *Noncommunicable diseases progress monitor 2020*, no. Oct. World Health Organization, 2020.

[2] C. World Health Organization., *NONCOMMUNICABLE DISEASES COUNTRY PROFILES 2018*. World Health Organization, 2018.

[3] K. M. Sturgeon *et al.*, "A population-based study of cardiovascular disease mortality risk in US cancer patients," *Eur. Heart J.*, vol. 40, no. 48, pp. 3889–3897, 2019, doi: 10.1093/eurheartj/ehz766.

[4] M. H. Hsieh, M. J. Hsieh, C. Chen, and C. Hsieh, "Comparison of machine learning

models for the prediction of mortality of patients with unplanned extubation in intensive care units," *Sci. Rep.*, no. July, pp. 1–7, 2018, doi: 10.1038/s41598-018-35582-2.

[5] S. Supangat, A. R. Amna, and T. Rahmawati, "Implementasi Decision Tree C4.5 Untuk Menentukan Status Berat Badan dan Kebutuhan Energi Pada Anak Usia 7-12 Tahun," *Teknika*, vol. 7, no. 2, pp. 73–78, 2018, doi: 10.34148/teknika.v7i2.90.

[6] S. H. David Hartanto Kamagi, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," vol. Vol. VI, p. 1, 2014.

[7] Indarto, U. Ema, and R. Suwanto, "PREDIKSI RISIKO KEMATIAN PASIEN STROKE PERDARAHAN DENGAN MENGGUNAKAN TEKNIK KLASIFIKASI DATA MINING Indarto1," vol. 5, no. 2, 2020.

[8] D. Andri and M. Reza, "Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung," vol. III, no. 2020, pp. 105–112, 2022.

[9] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[10] I. Junaedi, N. Nuswantari, and V. Yasin, "Perancangan Dan Implementasi Algoritma C4 . 5 Untuk Data Mining," *J. Inf. Syst. Informatics Comput.*, vol. 3, no. 1, pp. 29–44, 2019, [Online]. Available: http://journal.stmikjayakarta.ac.id/index.php/jisicom/article/view/203%0Ahttp://journal.stmikjayakarta.ac.id/index.php/jisicom/article/download/203/158.

[11] P. Algoritma and C. Berbasis, "Penerapan algoritma c4.5 berbasis," vol. 13, pp. 13–19, 2017.

[12] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[13] M. Balamurugan and S. Kannan, "Performance analysis of cart and C5.0 using sampling techniques," *2016 IEEE Int. Conf. Adv. Comput. Appl. ICACA 2016*, pp. 72–75, 2017, doi: 10.1109/ICACA.2016.7887926.

[14] E. P. K. Orpa, E. F. Ripanti, and T. Tursina, "Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *J. Sist. dan Teknol. Inf.*, vol. 7, no. 4, p. 272, 2019, doi: 10.26418/justin.v7i4.33163.

[15] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.