OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Penerapan Metode *Decision Tree* Menggunakan Algoritma *Iterative Dichotomiser* 3 (ID3) Untuk Klasifikasi Resiko Penyakit Jantung

Eva Fauziah[1*], Ahmad Fikri Zulfikar[1]

[1]Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Jl. Raya Puspiptek No. 46, Kel. Buaran, Kec. Serpong, Kota Tangerang Selatan. Banten 15310, Indonesia
Email: [1*]evafauziah112234@gmail.com, [2]dosen00386@unpam.ac.id

Abstrak−Penyakit jantung atau kardiovaskular adalah suatu kondisi yang disebabkan adanya penyempitan dan penyumbatan pembuluh darah, dimana menjadi salah satu penyakit yang mematikan paling banyak diderita disetiap negara. Resiko penyakit jantung menjadi suatu peristiwa yang tidak dapat dihindari karena kurangnya perhatian terhadap kesehatan jantung yang mana tidak diterapkannya pola hidup sehat dan pola makan sehat. Untuk itu dibutuhkan analisis terhadap resiko penyakit jantung. Klasifikasi merupakan salah satu metode data mining yang banyak digunakan dalam menentukan suatu keputusan yang diprediksi berdasarkan data terdahulu yang diproses menggunakan algoritma klasifikasi. Algoritma klasifikasi yang digunakan adalah iterative dichotomiser 3 (ID3) dengan menggunakan dataset yang diambil dari UCI Machine Learning Repository, bersumber dari V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Dataset terdiri dari 14 atribut diantaranya: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num (atribut prediksi). Metode evaluasi yang digunakan adalah confussion matrix dengan hasil perhitungan akurasi sebesar 85, 71%, presisi sebesar 84, 62% dan recal sebesar 84, 62%.

Kata Kunci: *Resiko Penyakit Jantung, Decision Tree Algoritma ID3, Klasifikasi*

Abstract−*Heart or cardiovascular disease is a condition caused by narrowing and blockage of blood vessels, which is one of the most common deadly diseases in every country. The risk of heart disease becomes an event that cannot be avoided because of a lack of attention to heart health where a healthy lifestyle and healthy eating patterns are not implemented. For this reason, an analysis of the risk of heart disease is needed. Classification is a data mining method that is widely used in determining a predictable decision based on previous data that is processed using a classification algorithm. The classification algorithm used is iterative dichotomizer 3 (ID3) using a dataset taken from the UCI Machine Learning Repository, sourced from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The dataset consists of 14 attributes including: age, sex, cp, tresbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num (predictive attribute). The evaluation method used is the confusion matrix with the results of calculating an accuracy of 85.71%, a precision of 84.62% and a recal of 84.62%.*

Keywords*: Heart Disease Risk, ID3 Algorithm Decision Tree, Classification*

1. PENDAHULUAN

Jantung adalah organ vital yang berperan dalam keberlangsungan hidup manusia. Jantung merupakan organ penyedia oksigen dan mengalirkannya melalui aliran darah keseluruh tubuh. Kemampuan fungsi jantung berturut-turut akan menurun karena selama manusia tersebut masih hidup maka jantung akan terus-menerus bekerja.

Penyakit jantung (kardiovaskular) merupakan berbagai keadaan ataupun kondisi terjadinya penyempitan dan penyumbatan pembuluh darah yang menjadi sebab terjadinya serangan jantung, nyeri dada atau angina dan stroke. Penyakit jantung menjadi salah satu penyakit yang paling banyak diderita (Bianto, Kusrini, & Sudarmawan, 2019).Penyakit jantung dapat dikategorikan sebagai penyakit mematikan yang mana penyakit nomor satu yang mendunia. Menurut data World Health Organization (WHO) pada tahun 2012 terdapat 17,5 juta orang meninggal karena penyakit jantung, setara dengan 31% dari 56,5 juta kematian di seluruh dunia. Penyebab resiko penyakit berawal tidak diterapkannya pola hidup sehat dan pola makan sehat.

2. METODOLOGI PENELITIAN

2.1 Metodologi Penelitian

Metode deskriptif adalah metode yang mendeskripsikan fenomena yang diteliti (menjelaskan objek penelitian). Adapun proses-proses yang dilakukan penulis dalam penelitian ini yaitu:

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

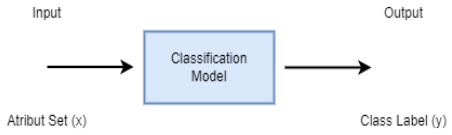
a. Identifikasi masalah
Tahapan yang dilakukan untuk menemukan permasalahan dan metode yang sesuai, yaitu dengan mengamati objek dari fenomena objek penelitian (observasi) dan studi literature guna mendapatkan informasi dari berbagai sumber seperti: buku, jurnal, prociding, karya ilmiah, website, artikel yang berhubungan dengan penelitian.
b. Pengumpulan data
Dalam pengumpulan data terdapat dua jenis data yaitu data primer dan data sekunder. Data primer merupakan data yang didapat dari sumber secara langsung sedangkan data sekunder merupakan data yang didapat dari sumber secara tidak langsung. Penulis dalam mengumpulkan data dari data riset UCI Machine Learning Repository.
c. Preprocessing data
Tahapan yang dilakukan yaitu memfilter data untuk tujuan memperoleh data yang optimal dan sesuai untuk bahan penelitian. Tahapan yang dilakukan berupa proses cleaning data, proses diskretisasi nilai dan pembagian data.
d. Penerapan metode klasifikasi
Tahapan yang dilakukan untuk membentuk model yang efektif dalam menganalisis data kasus penyakit jantung dengan pohon keputusan.
e. Evaluasi model
Tahapan yang dilakukan untuk menguji algoritma dengan menganalisis tingkat akurasi menggunakan teknik Confussion Matrix.

2.2 Data Mining

Data mining berhubungan erat dengan analisis data, memiliki banyak kesamaan tetapi dengan ciri khas yang berbeda berdasarkan aspek-aspek tertentu yang menitikberatkan pada masalah dan solusi yang dibutuhkan. Data mining adalah ilmu dari kumpulan data (database) yang mengekstraksi informasi yang berguna (Senubekti & Dewi, 2022). Data mining memiliki metode, paradigma, mekanisme, dan algoritme yang dapat digunakan untuk menggali modalitas informasi dan pengetahuan yang bermanfaat (Mostafa & Mahmoud, 2022).

2.3 Klasifikasi

Klasifikasi merupakan proses fungsi yang menerangkan atau menyeleksi kelas pada data agar dapat memprediksikan kelas yang labelnya belum diketahui dari objek tersebut (Islamiati & Widiartha, 2015). Klasifikasi diartikan sebagai suatu metode pembelajaran pada fungsi target f yang memetakan setiap set atribut x ke dalam satu dari beberapa label kelas y. Berikut gambar diagram metode klasifikasi:



Gambar 1. Diagram Klasifikasi

2.4 *Decision Tree*

Pohon keputusan (Decision Tree) adalah struktur pohon seperti diagram alur, di mana setiap simpul internal (node non-daun) menunjukkan pengujian pada atribut, setiap cabang mewakili hasil pengujian, setiap simpul daun (simpul terminal) memegang label kelas dan simpul paling atas dalam sebuah pohon adalah simpul akar (Han, Kamber, & Pei, 2012).

2.5 *Iterative Dichotomiser* 3 (ID3)

ID3 diterapkan menggunakan fungsi rekursif (fungsi yang memanggil fungsi itu sendiri) yang terstruktur secara top-down (dari atas ke bawah dibentuk dari simpul akar ke daun). ID3 merupakan algoritma pembelajaran pohon yang paling dasar (Tyasti, Ispriyanti, & Hoyyi, 2015). Jadi ID3 atau Iterative Dichotomiser 3 merupakan algoritma induksi decision tree yang paling dasar dengan menggunakan aturan information gain dalam mempartisi suatu data secara rekursif dan top-down

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Tahapan algoritma ID3 sebagai berikut:

a. Menyiapkan data set
b. Menghitung nilai entropi
c. Menghitung nilai gain
d. Membuat node cabang dari nilai gain yang terbesar
e. Ulangi langkah (b) sampai dengan (d) hingga semua node terpartisi.

2.6 *Information Gain*

Information Gain didasarkan pada karya perintis oleh Claude Shannon tentang teori informasi yang mempelajari nilai atau isi informasi dari pesan. Information Gain adalah perolehan informasi yang digunakan untuk memilih atribut uji dengan ukuran information gain (Islamiati & Widiartha, 2015).

2.6.1 *Entropi*

Metode deskriptif adalah metode yang mendeskripsikan fenomena yang diteliti (menjelaskan objek penelitian). Adapun proses-proses yang dilakukan penulis dalam penelitian ini yaitu:

Menurut teori informasi, entropi adalah ukuran ketidakpastian tentang sumber pesan. Rumus entropi menurut (Han, Kamber, & Pei, 2012) adalah:

$$\boldsymbol{Info\ (D) = \sum_{i=1}^{m} - pi * log_2 pi}$$

Keterangan:
D = Himpunan (dataset) kasus
m = Banyaknya partisi D
pi = Proporsi kelas i (probabilitas yang didapat dari jumlah dibagi total kasus).

Fungsi informasi didefinisikan sebagai perbedaan antara kebutuhan informasi asli (hanya berdasarkan proporsi kelas) dan kebutuhan baru (diperoleh setelah partisi pada A) yaitu rumus gain:

$$\boldsymbol{Gain(D, A) \equiv Info(D) - \sum_{v \in Values(A)} \frac{|D_j|}{|D|} Info(D_j)}$$

Keterangan:
D = Himpunan kelas klasifikasi
A = Atribut
v = Menyatakan suatu nilai yang mungkin untuk atribut A
Values (A) = Himpunan nilai-nilai yang mungkin untuk atribut A
$|D_j|$ = Sub-himpunan kelas klasifikasi (jumlah sampel untuk v)
|D| = Jumlah himpunan dalam D (jumlah seluruh sampel data)
Info (D_j) = Entropi untuk sampel-sampel yang memiliki nilai v

2.7 *Confussion Matrix*

Confusion matrix merupakan penguji model klasifikasi dengan menghitung nilai akurasi (Ridho & Hendra, 2021). Confusion Matrix adalah ringkasan tabel jumlah prediksi yang benar dan salah yang dibuat oleh algoritma klasifikasi dimana terdiri dari matrik nilai prediksi dan nilai actual/ground truth. Bagian confusion matrix adalah nilai Accuracy, Recall dan Precision. Accuracy merupakan menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Precision (confidence) merupakan proporsi kasus yang diprediksi positif yang juga positif benar. Recall (sensitivity) merupakan proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar. Berikut Perhitungan akurasi dengan tabel confusion matrix:

Tabel 1. *Confussion Matrix*

	Actual	
Prediksi	TP	FP
	FN	TN

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Confusion matrix terdapat empat istilah yang perlu dipahami sebagai representasi hasil proses klasifikasi yang digunakan dalam menghitung banyak ukuran evaluasi:

a. True Positive (TP) merupakan data positif yang diprediksi benar.
b. True Negative (TN) merupakan data negatif yang diprediksi benar.
c. False Positive (FP) merupakan data negatif namun diprediksi sebagai data positif.
d. False Negative (FN) merupakan data positif namun diprediksi sebagai negatif.

2.8 UCI *Machine Learning Repository*

UCI *Machine Learning Repository* merupakan repository yang mempunyai banyak arsip dengan informasi data set yang dibuat oleh David Aha dan sesama mahasiswa pascasarjana di UC Irvine. Data yang diambil dari UCI Machine Learning adalah data sekunder yang digunakan dalam penelitian (Pusporani, Qomariyah, & Irhamah, 2019). Jadi UCI Machine Learning Repository adalah situs web yang menyimpan kumpulan database yang digunakan oleh komunitas machine learning untuk analisis empiris algoritma pembelajaran mesin dan digunakan sebagai data sekunder untuk ide riset dalam penelitian.

2.9 *Software Rapid Miner*

Rapid Miner adalah perangkat lunak yang bersifat open source yang dibuat menggunakan Bahasa java dibawah lisensi GNU Public License untuk data mining yang terdiri dari operator untuk input, output, visualisasi, dan data preprocessing (Fatmawati, 2016). Jadi RapidMiner adalah software open source untuk menganalisa data mining, text mining dan analisis prediksi yang dapat bekerja di semua sistem operasi. Tampilan pada RapidMiner dikenal dengan istilah perspective dan terdapat 3 perspective dalam RapidMiner diantaranya welcome perspective, design perspective, dan result perspective.

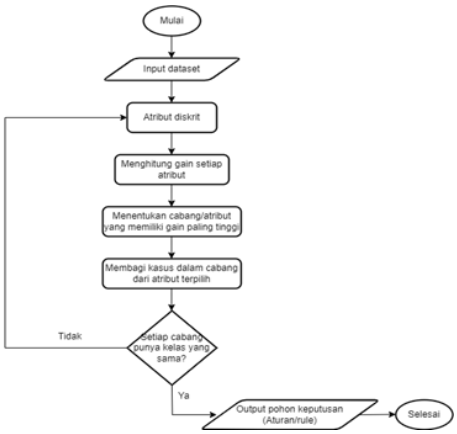
3. ANALISA DAN PEMBAHASAN

3.1 Analisa

Analisa merupakan pengamatan terhadap penelitian pada suatu masalah untuk mencari struktur masalah secara mendetail hingga didapat suatu penarikan kesimpulan. Resiko penyakit jantung menjadi fenomena yang tidak dapat dihindari, karena pada umumnya seseorang kurang memperhatikan kesehatan jantung dengan tidak menerapkan pola hidup sehat dan pola makan sehat. Penyakit jantung menjadi bagian dari penyebab kematian tertinggi. Dari pengamatan fenomena resiko penyakit jantung, terdapat faktor-faktor dan gejala yang timbul. Faktor dan gejala tersebut mempengaruhi terjadinya penyakit jantung, seperti sakit dada, tekanan darah tinggi, kolesterol, kadar gula darah, hasil EKG dan jumlah denyut jantung (Aulia, 2018).

Berdasarkan analisa sistem berjalan maka sistem yang diusulkan oleh penulis yaitu menerapkan metode decision tree menggunakan algoritma iterative dichotomiser 3 (ID3).

a. Berikut *Flowchart* Proses Membangun Model Algoritma ID3:



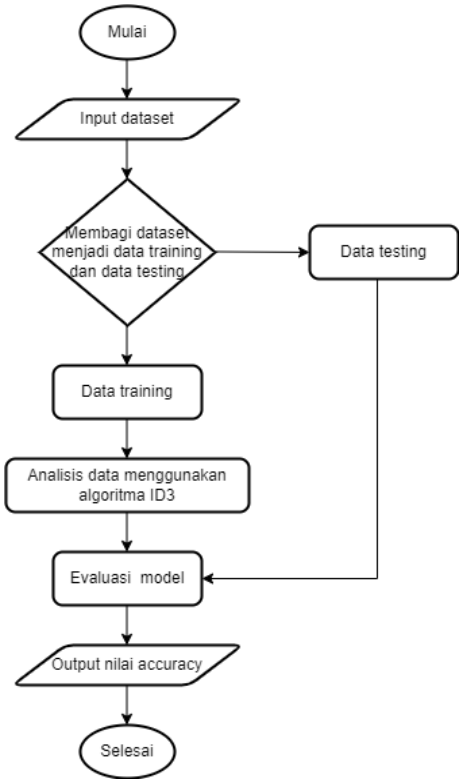
Gambar 2. *Flowchart* ID3

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Keterangan:

1. Peneliti menyiapkan dataset
2. Memproses data yang terdiri dari atribut diskrit
3. Mencari informasi gain dengan menghitung gain pada setiap atribut
4. Peneliti menentukan cabang atau atribut yang memiliki gain paling tinggi
5. Memproses split dengan membagi kasus dalam cabang dari atribut terpilih
6. Melakukan pengkondisian kasus, apabila setiap cabang mempunyai kelas yang homogen maka proses berhenti dengan hasil berupa kesimpulan (keputusan), apabila tidak homogen maka proses berulang dari proses (b) hingga proses (f) atau proses berulang untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

b. Berikut *Flowchart* Proses Pengujian Algoritma ID3:



Gambar 3. *Flowchart* Pengujian Model

Metode yang digunakan pada pengumpulan data dalam program aplikasi ini adalah sebagai berikut:

Keterangan:

1. Peneliti menyiapkan dataset
2. Membagi data menjadi 2, data training untuk model dan data testing untuk menguji
3. Data training dianalisis menggunakan algoritma ID3
4. Memasukan data testing ke dalam proses evaluasi model algoritma ID3
5. Hasil berupa nilai akurasi.

3.2 Pengumpulan Data

Data yang digunakan berasal dari Cleveland Heart Disease Dataset merupakan data sekunder yang dipublikasikan UCI Machine Learning Repository. Berdasarkan eksperimen yang dipublikasikan terdapat 14 atribut yang digunakan dengan 303 record, konklusinya nilai nol untuk tipe label yang tidak terkena penyakit jantung dan lebih dari nol untuk tipe label yang terkena penyakit jantung pada atribut target, terdapat beberapa missing value (data kosong) pada dataset dan karakteristik atribut terdiri dari kategorikal, integer dan real.

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

3.3 *Preprocessing Data*

Pada penelitian ini, pengolahan data sebagai bagian dari preprocessing guna mengurangi ukuran data yaitu:

3.3.1 *Cleaning Data*

Cleaning data yaitu proses pembersihan data yang dapat dilakukan dengan mengeleminasi data atau memperbaiki data yang rusak.

3.3.2 Diskretisasi Nilai

Diskretisasi adalah suatu proses transformasi data dari atribut yang berbentuk kontinu menjadi atribut kategori.

a. Age (umr): Umur dalam tahun (<45=umr-A, 45-60=umr-B, >60=umr-C)
b. Sex (jk): Jenis Kelamin (0=Perempuan, 1=Laki-laki)
c. CP (cp): Jenis Nyeri Dada (1=Typical Angina, 2=Atypical Angina, 3=Non Anginal, 4=Asymptomatic
d. Trestbps (trestbps): Tekanan darah istirahat (<129=tr-A, 129-163=tr-B, >163= tr-C)
e. Chol (chol): Kolestoral serum dalam mg/dl (<272=chol-A, 272-417=chol-B, >417= chol-C)
f. Fbs (fbs): gula darah puasa > 120 mg/dl (0=T, 1=Y)
g. Restecg (restecg): Hasil elektrokardiografi istirahat (0=Normal, 1= ST-T Abnormal, 2=Hipertropy)
h. Thalach (thalach): Denyut jantung maksimum (<114=th-A, 114-157=th-B, >157=th-C)
i. Exang (exang): Angina yang diinduksi olahraga (0=T, 1=Y)
j. Oldpeak (oldpeak) (<2,06= op A, 2,06-3,13= op B, >3,13= op C)
k. Slope (slope): Kemiringan segmen ST latihan puncak (1=Up-sloping, 2=Flat, 3=Down-sloping)
l. Ca (ca): Pembuluh darah yang diwarnai oleh flouroscopy (0=Normal, 1=Arteri-Koroner, 2= Anerisme, 3=Arteri-Perifer)
m. Thal (thal) (3=Normal, 6=Fixed Defect, 7=Reversible Defect)
n. Num (num): Atribut target/label (0=Negatip, 1=Positip)

3.3.3 Pembagian Data

Dataset merupakan sekumpulan data yang memiliki sifat sebagai himpunan data yang bersumber dari macam-macam informasi pada masa sebelumnya yang kemudian dikelola untuk menjadi sebuah informasi baru. Dataset yang digunakan dalam penelitian menjadi 279 record, terdapat 13 atribut input dan 1 atribut sebagai atribut target/label dengan 2 tipe kelas. Dari dataset tersebut, 90 % digunakan sebagai data training dan 10% digunakan untuk data testing. Data training yang digunakan dalam penelitian berjumlah 251 data kasus penyakit jantung. Data testing yang digunakan berjumlah 28 data kasus penyakit jantung.

3.4 Penerapan Metode Klasifikasi

Pembahasan Entropi:

Entropi Total:

$$Info(Total) = \left(-\frac{121}{251} * log_2\left(\frac{121}{251}\right)\right) + \left(-\frac{130}{251} * log_2\left(\frac{130}{251}\right)\right)$$

$$Info(Total) = (-0{,}48207171) * (-1{,}05268031) + (-0{,}51792828) * (-0{,}94917574)$$
$$= 0{,}50746739 + 0{,}49160495 = 0{,}9990724$$

Entropi umr:

$$Info(umr\ A) = \left(-\frac{12}{40} * log_2\left(\frac{12}{40}\right)\right) + \left(-\frac{28}{40} * log_2\left(\frac{28}{40}\right)\right) = 0{,}8812909$$

$$Info(umr\ B) = \left(-\frac{74}{146} * log_2\left(\frac{74}{146}\right)\right) + \left(-\frac{72}{146} * log_2\left(\frac{72}{146}\right)\right) = 0{,}9998646$$

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

$$Info(umr\ C) = \left(-\frac{35}{65} * log_2\left(\frac{35}{65}\right)\right) + \left(-\frac{30}{65} * log_2\left(\frac{30}{65}\right)\right) = 0{,}9957275$$

Entropi jk:

$$Info(L) = \left(-\frac{98}{170} * log_2\left(\frac{98}{170}\right)\right) + \left(-\frac{72}{170} * log_2\left(\frac{72}{170}\right)\right) = 0{,}9830606$$

$$Info(P) = \left(-\frac{23}{81} * log_2\left(\frac{23}{81}\right)\right) + \left(-\frac{58}{81} * log_2\left(\frac{58}{81}\right)\right) = 0{,}8607781$$

Entropi cp:

$$Info(Typical\ Angina) = \left(-\frac{7}{20} * log_2\left(\frac{7}{20}\right)\right) + \left(-\frac{13}{20} * log_2\left(\frac{13}{20}\right)\right) = 0{,}9340681$$

$$Info(Atypical\ Angina) = \left(-\frac{8}{39} * log_2\left(\frac{8}{39}\right)\right) + \left(-\frac{31}{39} * log_2\left(\frac{31}{39}\right)\right) = 0{,}7320667$$

$$Info(Non\ Anginal) = \left(-\frac{15}{70} * log_2\left(\frac{15}{70}\right)\right) + \left(-\frac{55}{70} * log_2\left(\frac{55}{70}\right)\right) = 0{,}7495953$$

$$Info(Asymptomatic) = \left(-\frac{91}{122} * log_2\left(\frac{91}{122}\right)\right) + \left(-\frac{31}{122} * log_2\left(\frac{31}{122}\right)\right) = 0{,}8177095$$

Entropi trestbps:

$$Info(tr\ A) = \left(-\frac{48}{107} * log_2\left(\frac{48}{107}\right)\right) + \left(-\frac{59}{107} * log_2\left(\frac{59}{107}\right)\right) = 0{,}9923629$$

$$Info(tr\ B) = \left(-\frac{64}{131} * log_2\left(\frac{64}{131}\right)\right) + \left(-\frac{67}{131} * log_2\left(\frac{67}{131}\right)\right) = 0{,}9996217$$

$$Info(tr\ C) = \left(-\frac{9}{13} * log_2\left(\frac{9}{13}\right)\right) + \left(-\frac{4}{13} * log_2\left(\frac{4}{13}\right)\right) = 0{,}8904916$$

Entropi chol:

$$Info(chol\ A) = \left(-\frac{77}{175} * log_2\left(\frac{77}{175}\right)\right) + \left(-\frac{98}{175} * log_2\left(\frac{98}{175}\right)\right) = 0{,}9895875$$

$$Info(chol\ B) = \left(-\frac{44}{75} * log_2\left(\frac{44}{75}\right)\right) + \left(-\frac{31}{75} * log_2\left(\frac{31}{75}\right)\right) = 0{,}9782177$$

$$Info(chol\ C) = \left(-\frac{0}{1} * log_2\left(\frac{0}{1}\right)\right) + \left(-\frac{1}{1} * log_2\left(\frac{1}{1}\right)\right) = 0$$

Entropi fbs:

$$Info(Y) = \left(-\frac{19}{40} * log_2\left(\frac{19}{40}\right)\right) + \left(-\frac{21}{40} * log_2\left(\frac{21}{40}\right)\right) = 0{,}9981959$$

$$Info(T) = \left(-\frac{102}{211} * log_2\left(\frac{102}{211}\right)\right) + \left(-\frac{109}{211} * log_2\left(\frac{109}{211}\right)\right) = 0{,}9992059$$

Entropi restecg:

$$Info(Normal) = \left(-\frac{47}{122} * log_2\left(\frac{47}{122}\right)\right) + \left(-\frac{75}{122} * log_2\left(\frac{75}{122}\right)\right) = 0{,}9616629$$

$$Info(ST.T\ Abnormal) = \left(-\frac{3}{4} * log_2\left(\frac{3}{4}\right)\right) + \left(-\frac{1}{4} * log_2\left(\frac{1}{4}\right)\right) = 0{,}8112781$$

$$Info(Hipertropi) = \left(-\frac{71}{125} * log_2\left(\frac{71}{125}\right)\right) + \left(-\frac{54}{125} * log_2\left(\frac{54}{125}\right)\right) = 0{,}9866165$$

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Entropi thalach:

$$Info(th\ A) = \left(-\frac{20}{23} * log_2\left(\frac{20}{23}\right)\right) + \left(-\frac{3}{23} * log_2\left(\frac{3}{23}\right)\right) = 0{,}5586294$$

$$Info(th\ B) = \left(-\frac{73}{131} * log_2\left(\frac{73}{131}\right)\right) + \left(-\frac{58}{131} * log_2\left(\frac{58}{131}\right)\right) = 0{,}9905215$$

$$Info(th\ C) = \left(-\frac{28}{97} * log_2\left(\frac{28}{97}\right)\right) + \left(-\frac{69}{97} * log_2\left(\frac{69}{97}\right)\right) = 0{,}8669837$$

Entropi exang:

$$Info(Y) = \left(-\frac{66}{82} * log_2\left(\frac{66}{82}\right)\right) + \left(-\frac{66}{82} * log_2\left(\frac{66}{82}\right)\right) = 0{,}7120641$$

$$Info(T) = \left(-\frac{55}{169} * log_2\left(\frac{55}{169}\right)\right) + \left(-\frac{114}{169} * log_2\left(\frac{114}{169}\right)\right) = 0{,}9102034$$

Entropi oldpeak:

$$Info(op\ A) = \left(-\frac{85}{209} * log_2\left(\frac{85}{209}\right)\right) + \left(-\frac{124}{209} * log_2\left(\frac{124}{209}\right)\right) = 0{,}9747344$$

$$Info(op\ B) = \left(-\frac{32}{37} * log_2\left(\frac{32}{37}\right)\right) + \left(-\frac{5}{37} * log_2\left(\frac{5}{37}\right)\right) = 0{,}571355$$

$$Info(op\ C) = \left(-\frac{4}{5} * log_2\left(\frac{4}{5}\right)\right) + \left(-\frac{1}{5} * log_2\left(\frac{1}{5}\right)\right) = 0{,}7219281$$

Entropi slope:

$$Info(Upsloping) = \left(-\frac{28}{111} * log_2\left(\frac{28}{111}\right)\right) + \left(-\frac{83}{111} * log_2\left(\frac{83}{111}\right)\right) = 0{,}8148284$$

$$Info(Flat) = \left(-\frac{83}{122} * log_2\left(\frac{83}{122}\right)\right) + \left(-\frac{39}{122} * log_2\left(\frac{39}{122}\right)\right) = 0{,}9040246$$

$$Info(Downsloping) = \left(-\frac{10}{18} * log_2\left(\frac{10}{18}\right)\right) + \left(-\frac{8}{18} * log_2\left(\frac{8}{18}\right)\right) = 0{,}9910761$$

Entropi ca:

$$Info(Normal) = \left(-\frac{40}{141} * log_2\left(\frac{40}{141}\right)\right) + \left(-\frac{101}{141} * log_2\left(\frac{101}{141}\right)\right) = 0{,}8604274$$

$$Info(Arteri\ Koroner) = \left(-\frac{37}{57} * log_2\left(\frac{37}{57}\right)\right) + \left(-\frac{20}{57} * log_2\left(\frac{20}{57}\right)\right) = 0{,}934849$$

$$Info(Anerisme) = \left(-\frac{27}{33} * log_2\left(\frac{27}{33}\right)\right) + \left(-\frac{6}{33} * log_2\left(\frac{6}{33}\right)\right) = 0{,}6840384$$

$$Info(Arteri\ Perifer) = \left(-\frac{17}{20} * log_2\left(\frac{17}{20}\right)\right) + \left(-\frac{3}{20} * log_2\left(\frac{3}{20}\right)\right) = 0{,}6098403$$

Entropi thal:

$$Info(Normal) = \left(-\frac{31}{131} * log_2\left(\frac{31}{131}\right)\right) + \left(-\frac{100}{131} * log_2\left(\frac{100}{131}\right)\right) = 0{,}78941$$

$$Info(Fixed\ Defect) = \left(-\frac{11}{17} * log_2\left(\frac{11}{17}\right)\right) + \left(-\frac{6}{17} * log_2\left(\frac{6}{17}\right)\right) = 0{,}9366674$$

$$Info(Reversable\ Defect) = \left(-\frac{79}{103} * log_2\left(\frac{79}{103}\right)\right) + \left(-\frac{24}{103} * log_2\left(\frac{24}{103}\right)\right) = 0{,}7832211$$

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Pembahasan Gain:

Gain umr:

$$Gain(Total\ umr) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|umr_j|}{|Total|} * Entropi\ umr_j$$

$$Gain(Total\ umr) = 0{,}9990724 - \left(\left(\frac{40}{251} * 0{,}8812909\right) + \left(\frac{146}{251} * 0{,}9998646\right) + \left(\frac{65}{251} * 0{,}9957275\right)\right)$$

$$= 0{,}01917533$$

Gain jk:

$$Gain(Total\ jk) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|jk_j|}{|Total|} * Entropi\ jk_j$$

$$Gain(Total\ jk) = 0{,}9990724 - \left(\left(\frac{170}{251} * 0{,}9830606\right) + \left(\frac{81}{251} * 0{,}8607781\right)\right) = 0{,}055473477$$

Gain cp:

$$Gain(Total\ cp) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|cp_j|}{|Total|} * Entropi\ cp_j$$

$$Gain(Total\ cp) = 0{,}9990724 - \left(\left(\frac{20}{251} * 0{,}9340681\right) + \left(\frac{39}{251} * 0{,}7320667\right) + \left(\frac{70}{251} * 0{,}7495953\right) + \left(\frac{122}{251} * 0{,}8177095\right)\right) = 0{,}204394332$$

Gain trestbps:

$$Gain(Total\ trestbps) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|trestbps_j|}{|Total|} * Entropi\ trestbps_j$$

$$Gain(Total\ trestbps) = 0{,}9990724 - \left(\left(\frac{107}{251} * 0{,}9923629\right) + \left(\frac{131}{251} * 0{,}9996217\right) + \left(\frac{13}{251} * 0{,}8904916\right)\right)$$

$$= 0{,}008197244$$

Gain chol:

$$Gain(Total\ chol) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|chol_j|}{|Total|} * Entropi\ chol_j$$

$$Gain(Total\ chol) = 0{,}9990724 - \left(\left(\frac{175}{251} * 0{,}9895875\right) + \left(\frac{75}{251} * 0{,}9782177\right) + \left(\frac{1}{251} * 0\right)\right)$$

$$= 0{,}016824795$$

Gain fbs:

$$Gain(Total\ fbs) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|fbs_j|}{|Total|} * Entropi\ fbs_j$$

$$Gain(Total\ fbs) = 0{,}9990724 - \left(\left(\frac{40}{251} * 0{,}9981959\right) + \left(\frac{211}{251} * 0{,}9992059\right)\right) = 2{,}73982E-05$$

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Gain restecg:

$$Gain(Total\ restecg) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|restecg_j|}{|Total|} * Entropi\ restecg_j$$

$$Gain(Total\ restecg) = 0{,}9990724 - \left(\left(\frac{122}{251} * 0{,}9616629\right) + \left(\frac{4}{251} * 0{,}8112781\right) + \left(\frac{125}{251} * 0{,}9866165\right)\right)$$

$$= 0{,}027378923$$

Gain thalach:

$$Gain(Total\ thalach) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|thalach_j|}{|Total|} * Entropi\ thalach_j$$

$$Gain(Total\ thalach) = 0{,}9990724 - \left(\left(\frac{23}{251} * 0{,}5586294\right) + \left(\frac{131}{251} * 0{,}9905215\right) + \left(\frac{97}{251} * 0{,}8669837\right)\right)$$

$$= 0{,}095868311$$

Gain exang:

$$Gain(Total\ exang) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|exang_j|}{|Total|} * Entropi\ exang_j$$

$$Gain(Total\ exang) = 0{,}9990724 - \left(\left(\frac{82}{251} * 0{,}7120641\right) + \left(\frac{169}{251} * 0{,}9102034\right)\right) = 0{,}153599734$$

Gain oldpeak:

$$Gain(Total\ oldpeak) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|oldpeak_j|}{|Total|} * Entropi\ oldpeak_j$$

$$Gain(Total\ oldpeak) = 0{,}9990724 - \left(\left(\frac{209}{251} * 0{,}9747344\right) + \left(\frac{37}{251} * 0{,}571355\right) + \left(\frac{5}{251} * 0{,}7219281\right)\right)$$

$$= 0{,}088836269$$

Gain slope:

$$Gain(Total\ slope) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|slope_j|}{|Total|} * Entropi\ slope_j$$

$$Gain(Total\ slope) = 0{,}9990724 - \left(\left(\frac{111}{251} * 0{,}8148284\right) + \left(\frac{122}{251} * 0{,}9040246\right) + \left(\frac{18}{251} * 0{,}9910761\right)\right)$$

$$= 0{,}128250395$$

Gain ca:

$$Gain(Total\ ca) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|ca_j|}{|Total|} * Entropi\ ca_j$$

$$Gain(Total\ ca) = 0{,}9990724 - \left(\left(\frac{141}{251} * 0{,}8604274\right) + \left(\frac{57}{251} * 0{,}934849\right) + \left(\frac{33}{251} * 0{,}6840384\right) + \left(\frac{20}{251} * 0{,}6098403\right)\right) = 0{,}164902146$$

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

Gain thal:

$$Gain(Total\ thal) = Entropi(Total) - \sum_{v \in Values(A)} \frac{|thal_j|}{|Total|} * Entropi\ thal_j$$

$$Gain(Total\ thal) = 0{,}9990724 - \left(\left(\frac{131}{251} * 0{,}78941\right) + \left(\frac{17}{251} * 0{,}9366674\right) + \left(\frac{103}{251} * 0{,}7832211\right)\right)$$

$$= 0{,}202228428$$

4. IMPLEMENTASI

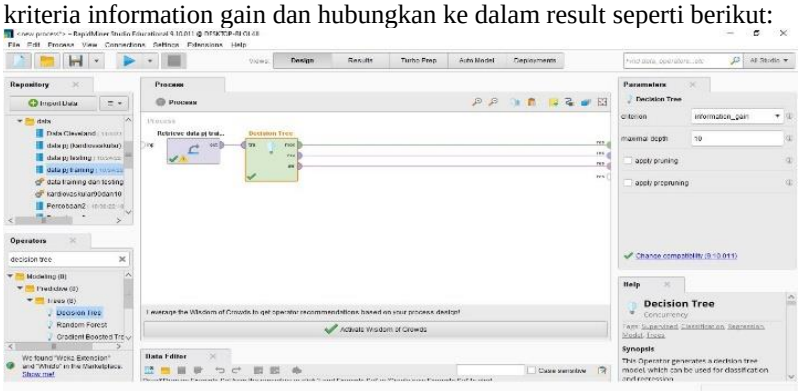
4.1 Implementasi

Implementasi adalah suatu proses yang mengubah rencana menjadi tindakan dengan tujuan mempraktekan dan mewujudkan susunan rencana kedalam suatu bentuk nyata.

4.1.1 Implementasi ID3

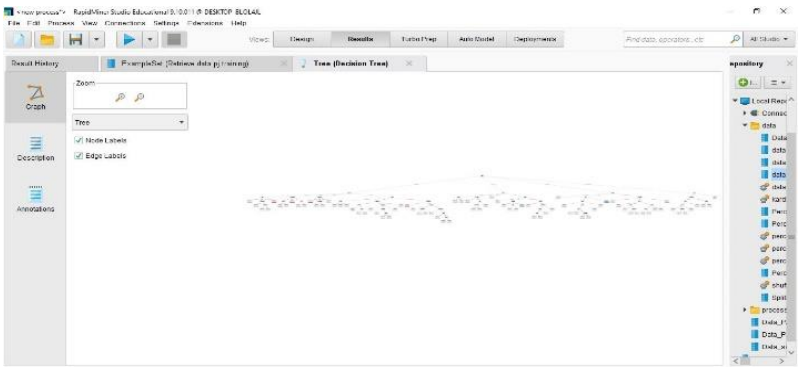
Berikut adalah proses klasifikasi resiko penyakit jantung menggunakan algoritma ID3 dengan software RapidMiner:

a. Tahap awal membuka software RapidMiner versi 9.10 yang telah terinstal di computer/pc. proses loading setelah itu muncul menu utama, setelah tampilan menu utama klik blank process untuk ke halaman proses pengolahan data.
b. Tahap kedua proses input dan format data, klik import data pada repository untuk proses input data, kemudian memformat data dengan mengatur tipe dan role setiap atribut.
c. Tahap ketiga proses analisa data training menggunakan algorita ID3, drag data yang tersimpan di repository. Selanjutnya menambahkan operator decision tree dengan kriteria information gain dan hubungkan ke dalam result seperti berikut:



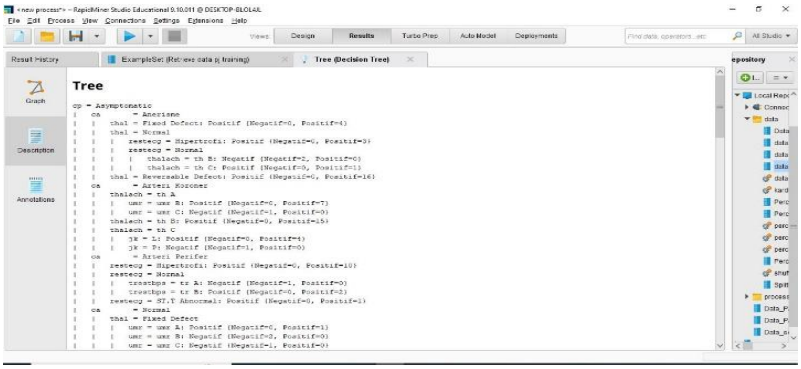
Gambar 4. Proses Algoritma

Klik play untuk mendapatkan hasil model graph dan rule seperti berikut:



Gambar 5. Graph ID3

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219



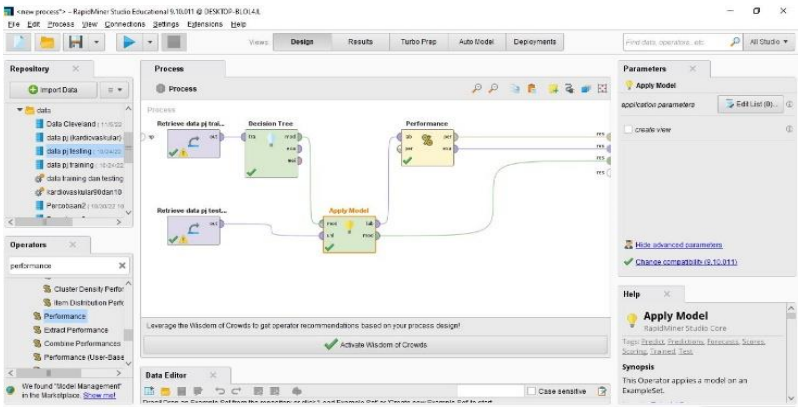
Gambar 6. Deskripsi ID3

4.1.2 Evaluasi Model

Berikut adalah proses pengujian model dengan software RapidMiner:

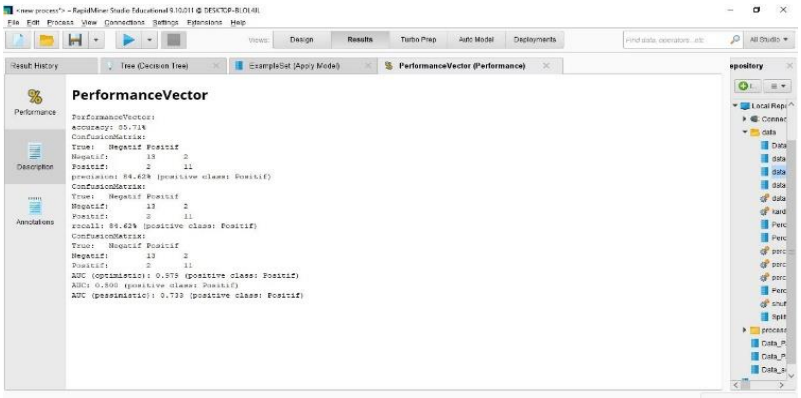
a. Tahap Terakhir,

setelah menganalisis data training menjadi model decision tree maka menguji model data training dengan menambahkan data testing (menggunakan operator apply model untuk mengaplikasikan model pada data testing dan operator performance untuk menampilkan hasil performa model terhadap data testing) di halaman proses pada software RapidMiner seperti berikut:



Gambar 7. Proses Pengujian

Pada operator performance hasil dari compussion table berupa nilai accuracy, recall dan precision dan operator Apply Model berupa hasil prediksi pengujian model terhadap data testing, seperti berikut:



Gambar 8. Hasil Evaluasi

OKTAL : Jurnal Ilmu Komputer dan Science
Volume 2, No. 4, April 2023
ISSN 2828-2442 (media online)
Hal 1207-1219

5. KESIMPULAN

5.1 Kesimpulan

Penulis telah melakukan penelitian dan menganalisis resiko penyakit jantung dengan metode decision tree menggunakan algoritma ID3, untuk itu penulis dapat mengambil kesimpulan seperti berikut:

a. Dengan penerapan metode decision tree menggunakan algoritma ID3 dapat mempercepat pengambilan keputusan dalam mengklasifikasikan resiko penyakit jantung.
b. Dari hasil penelitian, keefektivan dalam mengklasifikasikan resiko penyakit jantung dengan metode decision tree menggunakan algoritma ID3 mendapat akurasi sebesar 85, 71%, presisi sebesar 84, 62% dan recal sebesar 84, 62%.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan, maka penulis dapat memberikan saran seperti berikut:

a. Mengoptimalkan data dengan teknik preprocessing lain misalnya seperti mereduksi dimensi data, melakukan diskretisasi nilai dengan teknik entropy-based discretization dan yang lainnya. Agar kategori data menjadi lebih spesifik dan lebih cepat dalam pengambilan keputusan.
b. Untuk mendapatkan perbandingan tingkat efektivitas yang paling baik dalam menganalisis resiko penyakit jantung, perlu adanya penelitian lebih lanjut dengan menguji menggunakan metode lain ataupun komparasi C.45, CART, naïve bayes dan lain sebagainya.

REFERENCES

Aulia, W. (2018). Sistem Pakar Diagnosa Penyakit Jantung Koroner dengan Metode Probabilistic Fuzzy Decision Tree. *Jurnal Sains dan Informatika, IV*(2), 106-117.

Bianto, M. A., Kusrini, & Sudarmawan. (2019, Januari). Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes. *Citec Journal, VI*(1), 75-83.

Fatmawati. (2016, Maret). Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 dan Naive Bayes untuk Prediksi Penyakit Diabetes. *Jurnal Techno Nusa Mandiri, XIII*(1), 50-59.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques (Third Edition ed.).* Waltham, USA: Elsevier Inc.

Islamiati, S., & Widiartha, I. M. (2015, Oktober). Klasifikasi Penyakit Jantung Menggunakan Metode Decision Tree dengan Penerapan Algoritma C5.0. *Jurusan Ilmu Komputer*, 308-316.

Mostafa, A. A., & Mahmoud, H. E. (2022). Review of Data Mining Concept and its Techniques. *International Journal of Academic Research in Business and Social Sciences, XII*(6), 611 – 619.

Pusporani, E., Qomariyah, S., & Irhamah. (2019). Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning. Inferensi, *II(1)*, 25-32.

Ridho, R., & Hendra. (2021, Mei). Klasifikasi Diagnosis Penyakit COVID-19 Menggunakan Metode Decision Tree. *jurnal.umj.ac.id, XI*(3), 69 – 75.

Senubekti, M. A., & Dewi, L. A. (2022, Juli). Prinsip Klasifikasi dan Data Mining dengan Algoritma C4.5. *Jurnal Nuansa Informatika, XVI*(2), 87-93.

Tyasti, A. E., Ispriyanti, D., & Hoyyi, A. (2015). Algoritma Iterative Dichotomiser 3 (ID3) untuk Mengidentifikasi Data Rekam Medis (Studi Kasus Penyakit Diabetes Mellitus Di Balai Kesehatan Kementerian Perindustrian, Jakarta). *Jurnal Gaussian, IV*(2), 237 - 246.