

Implementasi Algoritma C4.5 pada Analisis Faktor Risiko Penyakit Jantung Koroner

Elsa Adinda Yestina^{1*}, Bakhtiyar Hadi Prakoso, Erna Selviyanti, Gandu Eko Julianto Suyoso

Jurusan Kesehatan, Manajemen Informasi Kesehatan, Politeknik Negeri Jember, Jember, Indonesia

Email: ^{1*}elsaystn@gmail.com, ²bakhtiyar.hp@polije.ac.id, ³ernaselviyanti@polije.ac.id, ⁴gandu.eko.js@polije.ac.id

Email Penulis Korespondensi: elsaystn@gmail.com

Submitted 07-08-2025; Accepted 23-09-2025; Published 31-10-2025

Abstrak

Penyakit jantung koroner merupakan penyakit tidak menular akibat proses aterosklerosis sehingga terjadi penyumbatan atau penyempitan pada pembuluh darah koroner yang menyebabkan pasokan oksigen ke jantung berkurang. Pada data morbiditas dan mortalitas di RSD dr. Soebandi Jember tahun 2020-2024 mengalami fluktuasi. Ditinjau dari segi usia, mayoritas penderitanya berusia produktif antara 15-64 tahun. Dampak yang telah terjadi yaitu peningkatan kasus penyakit jantung koroner yang semakin tinggi dan menurunnya produktivitas pada usia produktif. Tujuan penelitian ini untuk menganalisis faktor risiko penyakit jantung koroner berdasarkan rekam medis pasien rawat inap menggunakan algoritma C4.5 di RSD dr. Soebandi Jember. Penelitian ini menggunakan metode algoritma C4.5 dengan *tools* RapidMiner dan mengimplementasikan *k-fold cross validation*. Pemilihan algoritma C4.5 dikarenakan menghasilkan klasifikasi berdasarkan nilai *gain ratio* sehingga hasil klasifikasi dapat dianalisis atribut faktor risikonya. Penentuan faktor risiko diketahui pada hasil keputusan berupa *rules* klasifikasi. Implementasi *k-fold cross validation* menghasilkan akurasi tertinggi pada $k=8$, yaitu *accuracy* sebesar 86,09%, *precision* sebesar 82,63%, dan *recall* sebesar 91,39%. Berdasarkan hasil penelitian, diabetes melitus merupakan faktor risiko paling berpengaruh terhadap penyakit jantung koroner karena memperoleh nilai *gain ratio* tertinggi. Faktor risiko lain adalah inaktivitas fisik, jenis kelamin, obesitas, tekanan darah tinggi, usia, dan merokok. Saran untuk RSD dr. Soebandi yaitu meningkatkan Komunikasi, Informasi, Edukasi (KIE) khususnya kepada pasien Diabetes Melitus karena berpengaruh terhadap penyakit jantung koroner.

Kata Kunci: Penyakit Jantung Koroner; Faktor Risiko; Algoritma C4.5; RapidMiner; K-Fold Cross Validation

Abstract

Coronary heart disease is a non-communicable disease due to the process of atherosclerosis resulting in blockage or narrowing of the coronary blood vessels which causes reduced oxygen supply to the heart. In the morbidity and mortality data at RSD dr. Soebandi Jember from 2020-2024, there were fluctuations. In terms of age, the majority of sufferers are productive aged between 15-64 years. The impact that has occurred is an increase in cases of coronary heart disease which is getting higher and decreasing productivity at a productive age. The purpose of this study was to analyze risk factors for coronary heart disease based on medical records of inpatients using the C4.5 algorithm at RSD dr. Soebandi Jember. This research uses the C4.5 algorithm method with RapidMiner tools and implements *k-fold cross validation*. The selection of the C4.5 algorithm is because it produces classification based on the *gain ratio* value so that the classification results can be analyzed for risk factor attributes. Determination of risk factors is known in the decision results in the form of classification rules. The implementation *k-fold cross validation* produces the highest accuracy at $k=8$, namely accuracy of 86,09%, precision of 82,63%, and recall of 91,39%. Based on the results, diabetes mellitus is the most influential risk factor for coronary heart disease because it has the highest *gain ratio* value. Other risk factors are physical inactivity, gender, obesity, high blood pressure, age, and smoking. Suggestions for dr. Soebandi Hospital are to improve Communication, Information, Education (IEC), especially for diabetes mellitus patients because it affects coronary heart disease.

Keywords: Coronary Heart Disease, Risk Factors, C4.5 Algorithm, RapidMiner, K-Fold Cross Validation

1. PENDAHULUAN

Penyakit jantung koroner merupakan penyakit yang terjadi karena penyumbatan atau penyempitan pada pembuluh darah koroner sehingga pasokan oksigen ke jantung berkurang [1]. Penyakit jantung koroner di seluruh dunia menjadi pembunuh pertama sebesar 36% dari seluruh kematian tahun 2020. Di Indonesia, penyakit jantung koroner merupakan penyebab utama dari seluruh kematian sebesar 26,4%. Berdasarkan data pelaporan Dinas Kesehatan Kabupaten Jember tahun 2024, RSD dr. Soebandi merupakan peringkat pertama kasus penyakit jantung koroner di wilayah Jember. Pada data morbiditas tahun 2020 -2024 mengalami fluktuasi dengan kasus tertinggi tahun 2023 sebesar 365 kasus kesakitan. Pada data mortalitas juga mengalami fluktuasi dengan kasus tertinggi tahun 2020 sebesar 16 kasus kematian. Ditinjau dari segi usia, penderita penyakit jantung koroner di RSD dr. Soebandi tahun 2020-2024 didominasi oleh usia produktif (15-64 tahun). Dampak ketika permasalahan tersebut tidak segera diatasi yaitu mengganggu produktivitas kerja khususnya pada usia produktif [2]. Oleh sebab itu, untuk menangani permasalahan tersebut dilakukan analisis faktor risiko sebagai bahan untuk pengendalian kasus penyakit jantung koroner.

Faktor risiko merupakan tanda atau sekumpulan gejala pada penyakit yang diderita seseorang yang berhubungan dengan peningkatan kejadian penyakit baru berikutnya. Faktor risiko penyakit jantung koroner yang dapat dihindari yaitu stres, kurang aktivitas fisik, merokok, makanan dan kolesterol, diabetes melitus, hipertensi, dan obesitas sedangkan faktor risiko yang tidak dapat dihindari adalah usia, jenis kelamin, dan riwayat keluarga [3]. Kompleksitas antar faktor risiko dan perbedaan data medis membutuhkan metode analisis yang mampu mengungkap pola tersembunyi dan hubungan antar variabel. Pendekatan konvensional seringkali kurang efektif dalam mengolah data besar dan kompleks. Oleh karena itu, metode data *mining* dapat menjadi solusi potensial untuk mengklasifikasikan faktor risiko penyakit jantung koroner.

Data *mining* merupakan sekumpulan prosedur dan pendekatan yang digunakan oleh perangkat lunak untuk mengetahui pola tersembunyi dari kumpulan data. Selain itu, data *mining* juga mengekstraksi pengetahuan, analitik data, dan pengumpulan informasi[4]. Selanjutnya digunakan untuk membuat keputusan sebagai hasil klasifikasi. Penelitian ini menggunakan teknik klasifikasi algoritma C4.5. Algoritma C4.5 merupakan versi perbaikan dari ID3 digunakan untuk mengklasifikasikan suatu kejadian dengan pembentukan pohon keputusan. Keunggulan yang membedakan algoritma C4.5 dari ID3 adalah dapat menangani fitur dengan tipe numerik, melakukan pemotongan (*pruning*) *decision tree*, dan penurunan (*deriving*) *rule set*[5]. Hal pertama yang dilakukan dalam membangun pohon keputusan adalah menentukan *node* akar. *Node* akar diperoleh dari hasil perhitungan nilai *gain ratio* tertinggi[6]. Hasil pohon keputusan berupa *rules* digunakan sebagai bahan pertimbangan dalam pengambilan keputusan[7]. Dengan demikian, algoritma C4.5 mampu mengidentifikasi faktor risiko paling berpengaruh terhadap penyakit jantung koroner melalui pohon keputusan tersebut. Model yang dihasilkan mampu mendeteksi faktor risiko pada seseorang mengenai penyakit jantung koroner.

Berdasarkan penelitian sebelumnya, belum ada yang membahas faktor risiko penyakit jantung koroner menggunakan data *mining* khususnya algoritma C4.5, jadi artikel ini akan membahas faktor risiko penyakit jantung koroner dalam implementasi data *mining* menggunakan algoritma C4.5. Penelitian sebelumnya tahun 2023 bertujuan untuk mengklasifikasikan penyakit jantung koroner menggunakan algoritma *naïve bayes* menghasilkan akurasi 79,10%[8]. Pada penelitian lain tahun 2022, algoritma *naïve bayes* diimplementasikan dalam pembuatan sistem deteksi dini penyakit jantung koroner dengan atribut gejala menghasilkan akurasi 83,1%[9]. Penerapan data *mining* pada penelitian sebelumnya tahun 2024 membandingkan kinerja algoritma *naïve bayes* dan C4.5 pada klasifikasi diagnosa penyakit arteri koroner yang menghasilkan algoritma terbaik adalah C4.5 sebesar 46,9%[10]. Selain itu, terdapat penelitian tahun 2024 yang mengimplementasikan *decision tree* dalam pengklasifikasian penyakit jantung menghasilkan akurasi sebesar 87,5%[11]. Penelitian-penelitian tersebut hanya mengklasifikasikan penyakit jantung koroner tanpa melakukan analisis terhadap atribut-atribut yang digunakan. Penelitian lain tahun 2023 membahas faktor risiko penyakit jantung koroner namun tidak melakukan proses klasifikasi melainkan membahas pengaruh kejadian penyakit jantung koroner dengan pengumpulan data melalui kuisioner menghasilkan sebesar 70,5% berisiko pada laki-laki dan hipertensi sebesar 72,7% [12]. Dengan demikian, keterbaruan pada penelitian ini mencakup proses klasifikasi, kemudian hasil klasifikasi akan dilakukan analisis pada setiap atributnya.

Tujuan dari penelitian ini adalah menganalisis faktor risiko penyakit jantung koroner menggunakan algoritma C4.5. Hasil analisis dapat digunakan sebagai upaya dalam meningkatkan Komunikasi, Informasi, Edukasi (KIE) khususnya kepada pasien yang berisiko. Hal ini sesuai dengan Keputusan Menteri Kesehatan Nomor 854 tahun 2009 tentang Pedoman Pengendalian Penyakit Jantung dan Pembuluh Darah, yaitu kegiatan pokok yang dilakukan untuk program pencegahan dan pemberantasan penyakit dalam hal menurunkan angka kesakitan adalah peningkatan komunikasi, informasi, edukasi (KIE) dan pencegahan melalui kegiatan pemberdayaan dalam pencegahan serta penanggulangan faktor risiko penyakit jantung dan pembuluh darah[13]. Selanjutnya, hasil dari pengolahan C4.5 menghasilkan *rules* yang dapat dianalisis atribut faktor risiko penyakit jantung koroner. Penelitian ini menggunakan 9 atribut faktor risiko yaitu usia, jenis kelamin, status merokok, tekanan darah tinggi, dislipidemia, diabetes melitus (DM), obesitas, inaktivitas fisik, dan riwayat keluarga [3][14].

2. METODOLOGI PENELITIAN

2.1 Jenis Penelitian

Jenis penelitian ini adalah penelitian kuantitatif. Instrumen penelitian digunakan sebagai pengumpulan data dan data dianalisis secara kuantitatif dengan tujuan menguji hipotesis dan menganalisis hubungan sebab akibat [15]. Metode ini bertujuan untuk mengidentifikasi faktor-faktor yang berpengaruh terhadap fenomena tersebut serta menemukan pola tersembunyi dalam data. Hasil penelitian dapat dimanfaatkan untuk mendukung proses pengambilan keputusan [16].

2.2 Metode Pengumpulan Data

Data penelitian ini bersumber dari data sekunder yaitu rekam medis pasien rawat inap penyakit jantung koroner dan *non* penyakit jantung koroner di RSD dr. Soebandi Jember. Jumlah populasi sebesar 337 rekam medis pasien rawat inap yang terdiagnosis Penyakit Jantung Koroner tahun 2024. Selanjutnya dilakukan perhitungan sampel menggunakan rumus Isaac dan Michael karena jumlah populasi sudah diketahui:

$$S = \frac{\lambda^2 \cdot N \cdot P \cdot Q}{d^2 \cdot (N-1) \cdot \lambda^2 \cdot P \cdot Q} \quad (1)$$

Keterangan:

- S : Jumlah sampel yang diperlukan
- λ^2 : Nilai Chi-kuadrat tabel (dengan df = 1 dan tingkat signifikansinya 10%) yaitu 2,706
- N : Jumlah populasi
- P : Proporsi populasi dengan nilai ketetapan 0,5
- Q : (Q = 1 – P = 0,5)
- D : 0,05

Perhitungan sampel pada penelitian ini yaitu sebagai berikut:

$$S = \frac{2,706.337.0,5,0,5}{0,05^2 \cdot (337-1) + 2,706.0,5,0,5}$$

$$S = \frac{227,98}{1,52}$$

$$S = 150,3333$$

$$S = 151$$

Hasil perhitungan sampel menggunakan rumus Isaac dan Michael berjumlah 151. Perbandingan antara penyakit jantung koroner dan *non* penyakit jantung koroner adalah 1:1, sehingga diperoleh sampel 151 untuk penyakit jantung koroner dan sampel 151 untuk *non* penyakit jantung koroner. Sampel keseluruhan berjumlah 302 data. Teknik pengambilan sampel menggunakan teknik *Purposive Sampling* yaitu pengambilan sampel dari populasi yang diambil dengan mempertimbangkan kriteria tertentu [15]. Adapun kriteria sampel yaitu:

- a. Kriteria Inklusi
 1. Rekam medis pasien rawat inap dengan diagnosis utama penyakit jantung koroner tahun 2024 di RSD dr. Soebandi Jember.
 2. Rekam medis pasien rawat inap dengan diagnosis utama penyakit kardiovaskular *non* penyakit jantung koroner tahun 2024 di RSD dr. Soebandi Jember.
 3. Rekam medis yang terisi dengan maksimal ketidaklengkapan pada bagian atribut yaitu 2 atribut
- b. Kriteria Eksklusi

Rekam medis yang rusak dan tulisannya tidak terbaca.

Instrumen pengumpulan data pada penelitian ini adalah lembar observasi dan lembar *checklist* terhadap rekam medis pasien rawat inap. Lembar observasi digunakan untuk proses validasi mengenai atribut yang akan diteliti pada rekam medis. Lalu, lembar *checklist* digunakan untuk mengumpulkan data dari rekam medis pasien rawat inap supaya pencatatan atribut penelitian lebih mudah.

2.2 Data Mining

Data *mining* berfungsi untuk membuat atau mengambil keputusan masa depan berdasarkan informasi dari masa lalu. Berikut merupakan tahapan data *mining*:

- a. Data *selection*

Pada tahapan ini, data dipilih dari kumpulan data operasional [17]. Pengumpulan data dari rekam medis sesuai dengan lembar *checklist*.
- b. Data *cleaning/pre-processing*

Pada tahapan ini digunakan untuk menghilangkan kebisingan data yaitu data yang tidak konsisten, tidak relevan, tidak lengkap (*missing value*), data yang memiliki nilai berbeda dari nilai normal (*outlier*), dan data duplikasi [4]. Data *cleaning* merupakan proses menghilangkan data. Data yang tidak relevan sebaiknya dibuang karena dapat mengakibatkan *missing values* [18]. Penelitian ini dilakukan *pre-processing* data untuk menangani duplikasi data, *missing value*, dan inskonsisten data.
- c. Data *integration*

Pada tahapan ini digunakan untuk menggabungkan beberapa sumber data menjadi satu pandangan yang konsisten [4]. Tahapan ini hanya digunakan ketika data berasal dari sumber yang berbeda karena data dari sumber yang berbeda memiliki format, struktur, atau konteks yang berbeda pula [16]. Pada penelitian ini tidak dilakukan karena data berasal dari 1 sumber yaitu rekam medis pasien di RSD dr. Soebandi Jember.
- d. Data *transformation*

Pada tahapan ini data diubah ke dalam bentuk yang sesuai dari format asli menjadi format yang lebih sesuai untuk analisis dan data *mining*. Tujuannya untuk mempersiapkan data supaya dapat digunakan ketika mengekstrak informasi yang berguna dan mengetahui pola yang ada dalam data [4]. Data yang digunakan akan diubah dari tipe data *text* menjadi numerik dengan mengelompokkan nilai dari setiap atribut. Pengelompokan dapat dilakukan menggunakan format kode 1 dan 2 berdasarkan nilai normal dari setiap atribut [16].
- e. Data *mining*

Tahap ini merupakan proses pemilihan algoritma yang tepat yang akan digunakan dalam proses ini. Pemilihan algoritma sangat tergantung pada keseluruhan proses dan tujuan dari KDD [17]. Tahap ini algoritma C4.5 dipilih sebagai proses klasifikasi.
- f. *Model evaluation* dan *knowledge presentation*

Tahap ini merupakan interpretasi pola yang berasal dari prosedur data *mining*. Pola yang diidentifikasi tersebut akan disajikan dengan tampilan yang mudah dipahami [4].

2.3 Algoritma C4.5

Algoritma C4.5 adalah metode klasifikasi yang digunakan untuk membangun sebuah model pohon keputusan [10]. Proses klasifikasi dapat dilakukan dengan bantuan *tools* ataupun manual menggunakan rumus perhitungan. Berikut merupakan proses algoritma C4.5 menggunakan rumus perhitungan:

- a. Melakukan persiapan terhadap data latih (*training*)
- b. Melakukan perhitungan nilai *entropy*. Rumus yang digunakan untuk menghitung nilai *entropy* pada persamaan (1)

$$Entropy(\text{total}) = (-\left(\frac{B}{A}\right) \log_2\left(\frac{B}{A}\right)) + (-\left(\frac{C}{A}\right) \log_2\left(\frac{C}{A}\right)) \quad (1)$$

- c. Melakukan perhitungan nilai *gain*. Rumus yang digunakan untuk menghitung nilai *gain* pada persamaan (2)

$$Gain(E) = Entropi(D) - \left(\left(\frac{|H|}{|A|}\right) * Entropi(K) + \left(\frac{|O|}{|A|}\right) * Entropi(R) \right) \quad (2)$$

- d. Melakukan perhitungan nilai *split information*. Rumus yang digunakan untuk menghitung nilai *split information* pada persamaan (3)

$$Split\ Information\ (F) = \left(-\frac{|H|}{|A|} \log_2 \frac{|H|}{|A|} + \left(-\frac{|O|}{|A|} \log_2 \frac{|O|}{|A|} \right) \right) \quad (3)$$

- e. Menentukan nilai *gain ratio*. Nilai *gain ratio* yang tertinggi akan dijadikan atribut akar (*root*), sehingga pohon keputusan dapat terbentuk sebagai node 1. Rumus yang digunakan untuk menghitung nilai *gain ratio* pada persamaan (4)

$$Gain\ Ratio\ (G) = \frac{E}{F} \quad (4)$$

Oleh sebab itu, untuk mempermudah proses perhitungan digunakan tabel bantu perhitungan. Berikut merupakan tabel bantu perhitungan algoritma C4.5:

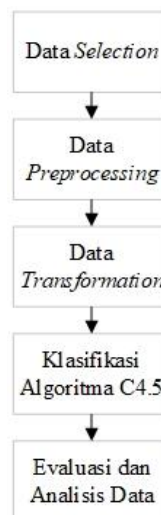
Tabel 1. Tabel Bantu Perhitungan Algoritma C4.5

	Jumlah	PJK	Non PJK	Entropy	Gain	Split Info	Gain Ratio
Total	A	B	C	D			
Atribut					E	F	G
Ya	H	I	J	K	L	M	N
Tidak	O	P	Q	R	S	T	U

Tabel 1 merupakan tabel bantu yang digunakan sebagai bentuk implementasi dari proses algoritma C4.5, yaitu mempersiapkan data *training* dan *testing*, menghitung *entropy*, *gain*, *split info*, dan *gain ratio*. Sebelum proses perhitungan algoritma C4.5, antara data *training* dan *testing* dibagi sesuai dengan nilai k yang digunakan. Setelah itu, menghitung tingkat *accuracy*, *precision*, dan *recall* menggunakan *confusion matrix*. Hasil akurasi tertinggi dari akan dibuatkan pohon keputusan untuk dianalisis.

2.3 Alur Penelitian

Alur penelitian berguna untuk memudahkan peneliti saat melakukan penelitian, berikut merupakan alur penelitian yang akan dilakukan peneliti untuk menyelesaikan permasalahan:



Gambar 1. Alur Penelitian

Berdasarkan gambar 1 tersebut, yang dilakukan adalah data *selection* atau peneliti mengumpulkan data dari rekam medis menggunakan lembar *checklist*. Lalu *pre-processing* data untuk melakukan pemfilteran duplikasi data, *missing value*, dan inkonsistensi data.

- a. Duplikasi Data

Duplikasi data merupakan data kembar atau sama. Tahapan ini bertujuan untuk memeriksa data pada dataset yang lebih dari satu. Jika ada data yang teridentifikasi duplikasi, maka data duplikasi tersebut akan dihapus dan hanya disisakan satu sebagai data tunggal yang tidak duplikasi. Penghapusan data duplikasi berguna untuk mencegah perhitungan ganda yang dapat mengganggu hasil analisis[19].

b. Missing Value

Missing value merupakan nilai dari suatu data yang hilang. Pada *missing value* ini, data yang hilang dapat ditangani dengan mengabaikan *tuple* yang hilang. Namun, pengabaian *tuple* dapat menghilangkan kesempatan untuk menggunakan nilai-nilai dari atribut lain dalam analisis. Apabila satu atau dua atribut hilang, informasi dari atribut lainnya tetap dapat berkontribusi pada pemahaman dataset secara keseluruhan. Pada tahap ini, dapat dilakukan pengisian pada data yang hilang namun membutuhkan waktu yang lama apabila kumpulan data cukup besar. Nilai data yang hilang tersebut dapat diganti dengan kecenderungan sentral atau nilai yang sering muncul pada atribut tersebut [4]. Selain itu, penanganan *missing value* juga dapat dengan data *cleaning*. Data *cleaning* merupakan proses menghilangkan data. Data yang tidak relevan sebaiknya dibuang karena dapat mengakibatkan *missing values* [18].

c. Inkonsistensi Data

Inkonsistensi data merupakan ketidakselarasan atau ketidaksesuaian dalam suatu data. Hal ini terjadi ketika terjadi pengambilan keputusan dalam situasi yang berbeda namun seharusnya sama [20]. Inkonsistensi data dapat dideteksi melalui pemeriksaan format data dan validasi nilai dalam pengisian data. Dalam hal ini, pengisian format pada data harus konsisten dari awal hingga akhir.

Tahapan selanjutnya adalah data *transformation* yang dilakukan dengan mengubah tipe data supaya mudah untuk diproses. Proses pengubahan dari kategori nominal “Ya” menjadi nilai numerik “1” dan “Tidak” menjadi “2”. Setelah itu, proses klasifikasi Algoritma C4.5. Saat melakukan klasifikasi algoritma C4.5, menggunakan teknik *k-fold cross validation* untuk membagi data menjadi k subset. Percobaan dilakukan 5 kali untuk mendapatkan akurasi tertinggi. Proses pengolahan data ini menggunakan *tools* RapidMiner. Proses klasifikasi algoritma C4.5 akan menghasilkan sebuah pohon keputusan. Data *training* digunakan untuk menghasilkan pohon keputusan dan menghasilkan *rule* sedangkan data *testing* digunakan untuk mengevaluasi kinerja model. Kemudian melakukan evaluasi dari percobaan nilai k untuk menemukan nilai akurasi dengan cara melakukan pengukuran tingkat *accuracy*, *precision*, dan *recall* menggunakan *Confusion Matrix*. Lalu, peneliti akan menganalisis hasil klasifikasi algoritma C4.5 yang nantinya menghasilkan pohon keputusan untuk menemukan hubungan atau pola tersembunyi antar atribut berdasarkan pohon keputusan tersebut. Setiap skenario akan diproses menggunakan algoritma C4.5, lalu pohon keputusan yang dibentuk adalah pohon keputusan berdasarkan nilai k yang memiliki tingkat akurasi tertinggi. Kemudian peneliti menuangkan hasil pengolahan data analisisnya berdasarkan tujuan khusus penelitian..

2.3 Atribut Faktor Risiko

Pada penelitian ini menggunakan atribut faktor risiko penyakit jantung koroner yaitu usia, jenis kelamin, status merokok, tekanan darah tinggi, dislipidemia, diabetes melitus, obesitas, inaktivitas fisik, dan riwayat keluarga. Setiap atribut dikategorikan menjadi 2 kategori, yaitu risiko tinggi dan risiko rendah. Kategori ini ditetapkan berdasarkan nilai dari indikator. Berikut merupakan pengkategorian atribut faktor risiko penyakit jantung koroner:

Tabel 2. Pengkategorian Atribut

Atribut	Indikator	Kategori
Usia	1. ≥ 40 tahun pada laki-laki; ≥ 50 tahun pada perempuan	1. Risiko tinggi
	2. < 40 tahun pada laki-laki; < 50 tahun pada perempuan	2. Risiko rendah
Jenis kelamin	1. Laki-laki	1. Risiko tinggi
	2. Perempuan	2. Risiko rendah
Status merokok	1. Ya merokok	1. Risiko tinggi
	2. Tidak merokok	2. Risiko rendah
Tekanan darah tinggi	1. $\geq 140/90$ mmHg	1. Risiko tinggi
	2. $< 140/90$ mmHg	2. Risiko rendah
Dislipidemia	1. LDL ≥ 100 mg/dL, HDL < 40 mg/dL atau > 60 mg/dL, Kolesterol total ≥ 220 mg/dL, dan Triglisierida ≥ 150 mg/dL	1. Risiko tinggi
	2. LDL < 100 mg/dL, HDL 40-60 mg/dL, Kolesterol total < 220 mg/dL, dan Triglisierida < 150 mg/dL	2. Risiko rendah
Diabetes melitus	1. Ada riwayat personal DM	1. Risiko tinggi
	2. Tidak ada riwayat personal DM	2. Risiko rendah
Obesitas	1. IMT ≥ 25	1. Risiko tinggi
	2. IMT < 25	2. Risiko rendah
Inaktivitas fisik	1. Tidak olahraga	1. Risiko tinggi
	2. Olahraga	2. Risiko rendah
Riwayat keluarga	1. Ada riwayat keluarga PJK	1. Risiko tinggi
	2. Tidak ada riwayat keluarga PJK	2. Risiko rendah

Tabel 2 merupakan pengkategorian atribut faktor risiko yang bertujuan untuk membuat pengelompokan sehingga mempermudah dalam proses analisis untuk pengambilan keputusan. Pengkategorian ini dilakukan sebelum proses data *mining*.

3. HASIL DAN PEMBAHASAN

3.1 Proses Klasifikasi

Penelitian ini dimulai dengan tahap data *selection* yang merupakan pengumpulan rekam medis berjumlah 302. Berikut merupakan data yang digunakan:

Tabel 3. Data Rekam Medis PJK dan *non* PJK

No RM	Usia	Jenis Kelamin	Status Merokok	Tekanan Darah Tinggi	Dislipidemia	DM	Obesitas	Inaktivitas Fisik	Riwayat Keluarga	Risiko PJK
37202	Ya	Ya	Ya	Tidak	Ya	Tidak	Ya	Ya	-	Ya
02404	Ya	Ya	Ya	Tidak	-	Tidak	Ya	Ya	Tidak	Ya
32610	Tidak	Ya	Ya	Tidak	-	Tidak	Ya	Ya	-	Ya
05012	Tidak	Ya	Ya	Ya	-	Tidak	Ya	Ya	-	Ya
....										
51128	Ya	Tidak	Tidak	Tidak	-	Ya	Tidak	Tidak	-	Tidak

Tabel 3 di atas digunakan untuk mencari *rules* keputusan terhadap faktor risiko penyakit jantung koroner pada algoritma C4.5. Setelah data terkumpul dilakukan tahapan *pre-processing* data menggunakan *tools* RapidMiner. Filter yang digunakan pada tahap *pre-processing* adalah duplikasi data, *missing value*, inkonsistensi data.

a. Duplikasi data

Duplikasi data merupakan data yang kembar atau sama. Penghapusan data teridentifikasi duplikasi dapat mencegah perhitungan ganda yang mengganggu hasil analisis dan menghasilkan analisis yang akurat[19]. Jika ada data yang teridentifikasi duplikasi, maka data duplikasi tersebut akan dihapus dan hanya disisakan satu sebagai data tunggal yang tidak duplikasi. Penghapusan data duplikasi berguna untuk mencegah perhitungan ganda yang dapat mengganggu hasil analisis[19]. Pada penelitian ini, dilakukan pemfilteran untuk menangani data yang sama yaitu nomor rekam medis. Hasil *running* tidak menunjukkan adanya duplikasi pada nomor rekam medis.

b. Missing value

Missing value merupakan data yang hilang/kosong. Penanganan *missing value* dengan cara menghapus atribut, mencari modus apabila atribut karakter dan menggunakan *mean* apabila atribut numerik[21]. Apabila satu atau dua atribut hilang, informasi dari atribut lainnya tetap dapat berkontribusi pada pemahaman dataset secara keseluruhan. Pada tahap ini, dapat dilakukan pengisian pada data yang hilang namun membutuhkan waktu yang lama apabila kumpulan data cukup besar. Nilai data yang hilang tersebut dapat diganti dengan kecenderungan sentral atau nilai yang sering muncul pada atribut tersebut[4]. Pada penelitian ini, data yang mengalami *missing value* terdapat pada atribut dislipidemia dan riwayat keluarga. *Missing value* pada kedua atribut ini sangat tinggi yaitu atribut dislipidemia sebesar 280 dari 302 dan atribut riwayat keluarga sebesar 299 dari 302. Oleh karena itu, dilakukan data *cleaning* pada atribut tersebut. Pembersihan ini dilakukan supaya mendapatkan data yang relevan karena apabila tidak dilakukan *cleaning* data dapat mengurangi akurasi [22].

c. Inkonsistensi data

Inkonsistensi data digunakan untuk menangani data yang tidak konsisten. Inkonsistensi data merupakan ketidakselarasan atau ketidaksesuaian dalam suatu data. Hal ini terjadi ketika terjadi pengambilan keputusan dalam situasi yang berbeda namun seharusnya sama[20]. Inkonsistensi data dideteksi melalui pemeriksaan format data dan validasi nilai dalam pengisian data, artinya data dari awal sampai akhir harus memiliki format data dan penulisan yang sama. Pada penelitian ini tidak terdapat inkonsistensi data.

Setelah data bersih, tahap selanjutnya adalah tahapan data *transformation*. Data *transformation* merupakan proses untuk mengubah suatu kategori dari format asli menjadi format yang lebih sesuai berdasarkan kategori yang telah ditentukan sebelumnya. Proses ini dilakukan dengan cara mengubah kategori nominal “Ya” menjadi nilai numerik “1” dan “Tidak” menjadi “2”. Namun, *tools* yang digunakan pada penelitian ini adalah RapidMiner, yang mana *tools* ini mampu mendukung file bertipe *excel* dengan tipe data nominal atau diskrit [23]. Tipe data nominal pada penelitian ini sudah sesuai sehingga tahap *transformation* tidak dilakukan. Tahapan berikutnya adalah data *mining*. Pada tahapan ini, proses klasifikasi menggunakan algoritma C4.5. Proses klasifikasi menggunakan algoritma C4.5 menggunakan *tools* RapidMiner dengan teknik *k-fold cross validation*. *K-fold cross validation* merupakan teknik evaluasi yang digunakan untuk mempartisi kumpulan data menjadi *subset training* dan *testing* [24]. Penelitian ini menggunakan nilai $k=4$, $k=5$, $k=7$, $k=8$, dan $k=10$ dalam pembagian data *training* dan *testing*. Data *training* digunakan untuk membentuk pohon keputusan. Pohon keputusan dibentuk dengan menerapkan teknik *pruning* dan *prepruning* pada operator *decision tree* serta tanpa menerapkan teknik *pruning* dan *prepruning*. Parameter *prepruning* dan *pruning* berfungsi untuk mencegah *overfitting*, mengoptimalkan struktur pohon keputusan yang terbentuk, dan menghasilkan model yang lebih

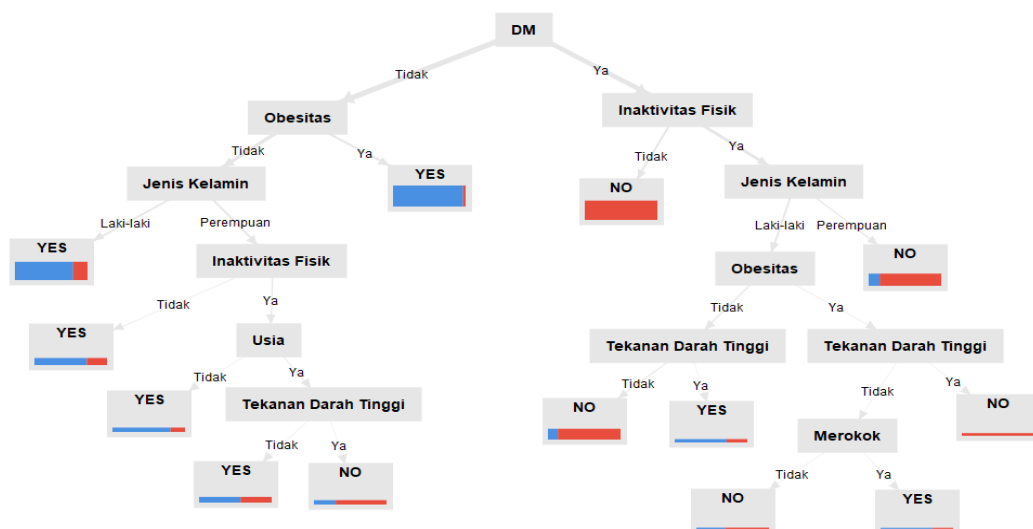
generalisasi[25]. Tujuan dari pendekatan ini untuk mengetahui perbedaan performa dan struktur pohon keputusan yang dihasilkan masing-masing teknik tersebut [26]. Hal ini tentu berbeda pada data *testing*, data *testing* digunakan untuk menghasilkan performa model. *Performance* ini menghasilkan tabel *confusion matrix* yang berisi *True Positive*, *False Positive*, *False Negative*, dan *True Negative*.

Penentuan *node* akar diperoleh berdasarkan perhitungan *gain ratio* tertinggi. Nilai *gain ratio* diperoleh dari proses perhitungan algoritma C4.5 seperti halnya contoh pada tabel 1. Proses klasifikasi melalui perhitungan manual menggunakan nilai $k=8$ untuk pembagian data *training* dan *testing*. Proses perhitungan manual hanya menggunakan data *training* karena perhitungan ini akan digunakan untuk membentuk pohon keputusan. Hasil perhitungan manual terdapat pada tabel berikut:

Tabel 4. Hasil Perhitungan Klasifikasi Algoritma C4.5

Atribut		Jumlah (S)	Ya (S)	Tidak (S)	Entropy	Gain	Split Info	Gain Ratio
		266	133	133	1			
Usia	Ya	225	117	108	0,9988	0,0064	0,6201	0,0103
	Tidak	41	16	25	0,9650			
Jenis kelamin	Ya	163	93	70	0,9856	0,0229	0,9630	0,0238
	Tidak	103	40	63	0,9637			
Obesitas	Ya	95	60	35	0,9495	0,0280	0,9403	0,0298
	Tidak	171	73	98	0,9845			
Tekanan darah tinggi	Ya	83	52	31	0,9533	0,0211	0,8955	0,0236
	Tidak	183	81	102	0,9905			
DM	Ya	121	18	103	0,6067	0,3231	0,9941	0,3250
	Tidak	145	115	30	0,7355			
Merokok	Ya	102	58	44	0,9864	0,0085	0,9604	0,0088
	Tidak	164	75	89	0,9947			
Inaktivitas Fisik	Ya	193	115	78	0,9733	0,0726	0,8478	0,0857
	Tidak	73	18	55	0,8058			

Berdasarkan hasil perhitungan pada tabel 4, diperoleh bahwa atribut diabetes melitus memiliki nilai *gain ratio* tertinggi yaitu sebesar 0,3250. Oleh karena itu, pada pembentukan pohon keputusan, atribut diabetes melitus akan menjadi *node* akar. Nilai *gain ratio* tertinggi setelah diabetes melitus adalah inaktivitas fisik sebesar 0,0857 dan obesitas sebesar 0,0298, sehingga *node* internal selanjutnya adalah inaktivitas fisik dan obesitas. Berikut merupakan hasil pohon keputusan dengan *pruning-prepruning*:



Gambar 2. Pohon Keputusan Faktor Risiko PJK dengan *pruning-prepruning*

Gambar 2 merupakan pohon keputusan yang telah terbentuk menggunakan *tools* RapidMiner yang mengaktifkan parameter *pruning* dan *prepruning* sehingga bentuk pohon lebih sederhana dan terkontrol. Hasilnya, pohon keputusan memiliki jumlah *node* dan kedalaman yang lebih sedikit sehingga tampak lebih ringkas dan mudah dipahami. Struktur pohon yang seperti ini mengurangi risiko *overfitting* dan meningkatkan kemampuan generalisasi model. Akar pohon berupa atribut DM membagi cabang pohon menjadi 2 bagian yaitu DM = Tidak dan DM = Ya. Pada masing-masing cabang tersebut akan membentuk suatu *rules* keputusan yang dapat dilihat pada tabel 4 berikut:

Tabel 5. *Rules* Jika DM = Tidak dengan *Pruning-Prepruning*

No	Rules	Keputusan
1	IF DM = Tidak AND Obesitas = Tidak AND Jenis Kelamin = Perempuan AND Inaktivitas Fisik = Tidak {YES=13, NO=5}	PJK
2	IF DM = Tidak AND Obesitas = Tidak AND Jenis Kelamin = Perempuan AND Inaktivitas Fisik = Ya AND Usia = Tidak {YES=8, NO=2}	PJK
3	IF DM = Tidak AND Obesitas = Tidak AND Jenis Kelamin = Perempuan AND Inaktivitas Fisik = Ya AND Usia = Ya AND Tekanan Darah Tinggi = Tidak {YES=8, NO=6}	PJK
4	IF DM = Tidak AND Obesitas = Tidak AND Jenis Kelamin = Perempuan AND Inaktivitas Fisik = Ya AND Usia = Ya AND Tekanan Darah Tinggi = Ya {YES=3, NO=7}	NON PJK
5	IF DM = Tidak AND Obesitas = Tidak AND Jenis Kelamin = Laki-laki {YES=41, NO=10}	PJK
6	IF DM = Tidak AND Obesitas = Ya {YES=57, NO=2}	PJK
7	IF DM = Ya AND Inaktivitas Fisik = Tidak {YES=0, NO=54}	NON PJK
8	IF DM = Ya AND Inaktivitas Fisik = Ya AND Jenis Kelamin = Perempuan {YES=5, NO=28}	NON PJK
9	IF DM = Ya AND Inaktivitas Fisik = Ya AND Jenis Kelamin = Laki-laki AND Obesitas = Tidak AND Tekanan Darah Tinggi = Tidak {YES=4, NO=25}	NON PJK
10	IF DM = Ya AND Inaktivitas Fisik = Ya AND Jenis Kelamin = Laki-laki AND Obesitas = Tidak AND Tekanan Darah Tinggi = Ya {YES=5, NO=2}	PJK
11	IF DM = Ya AND Inaktivitas Fisik = Ya AND Jenis Kelamin = Laki-laki AND Obesitas = Ya AND Tekanan Darah Tinggi = Tidak AND Merokok = Tidak {YES=2, NO=3}	NON PJK
12	IF DM = Ya AND Inaktivitas Fisik = Ya AND Jenis Kelamin = Laki-laki AND Obesitas = Ya AND Tekanan Darah Tinggi = Tidak AND Merokok = Ya {YES=5, NO=2}	PJK
13	IF DM = Ya AND Inaktivitas Fisik = Ya AND Jenis Kelamin = Laki-laki AND Obesitas = Ya AND Tekanan Darah Tinggi = Ya {YES=0, NO=5}	NON PJK

Berdasarkan tabel 5, diperoleh 13 *rules* dengan keputusan PJK sebanyak 7 *rules* sedangkan non PJK sebanyak 6 *rules*. Pada keputusan PJK berarti seseorang akan berisiko penyakit jantung koroner sedangkan pada keputusan non PJK berarti seseorang tidak berisiko terhadap penyakit jantung koroner. Secara keseluruhan *rules* keputusan lebih banyak menghasilkan *rules* dengan PJK dibandingkan dengan non PJK.

Tahap akhir dari data *mining* adalah evaluasi/pengujian. Pengujian dengan *confusion matrix* telah dilakukan sebanyak 5 kali baik dengan pengaktifan *pruning-prepruning* maupun tidak. Berikut merupakan hasil pengujian:

Tabel 6. Hasil Pengujian *Confusion Matrix*

K-fold cross validation	Pruning dan Prepruning Aktif			Pruning dan Prepruning Tidak Aktif		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
k=4	83,77	79,65	90,73	84,77%	81,44%	90,07%
k=5	81,46	79,5	84,77	83,11%	81,25%	86,09%
k=7	83,44	80,24	88,74	84,44%	82,10%	88,08%
k=8	86,09	82,63	91,39	84,77%	81,44%	90,07%
k=10	84,11	79,77	91,39	85,10%	81,93%	90,07%

Berdasarkan tabel 6, *accuracy* pada masing-masing pengujian menghasilkan nilai yang berbeda. *Accuracy* tertinggi implementasi *k-fold cross validation* yaitu menggunakan nilai k=8 pada pengaktifan *pruning* dan *prepruning* sebesar 86,09% sedangkan tanpa *pruning* dan *prepruning* *accuracy* tertinggi pada nilai k=10 yaitu sebesar 85,10%. Besar nilai *accuracy* menunjukkan bahwa algoritma C4.5 mampu mengklasifikasikan data dengan baik. Besar nilai *precision* memiliki arti seberapa tepat model dalam prediksi faktor risiko penyakit jantung koroner. Nilai *recall* menunjukkan kemampuan model dalam mendeteksi semua contoh data (berisiko dan tidak berisiko) yang benar-benar termasuk kelas positif dalam *dataset*.

3.2 Analisis Faktor Risiko

a. Diabetes Melitus

Atribut DM memberikan pengaruh terbesar dalam memisahkan label/kelas pada data dibuktikan dengan perolehan nilai *gain ratio* yang tinggi. Atribut dengan nilai *gain ratio* tertinggi merupakan atribut yang menunjukkan kemampuannya dalam membedakan label/kelas karena lebih relevan dalam proses klasifikasi [27]. Pasien dengan DM mempunyai risiko 16,9 kali untuk menderita penyakit jantung koroner [28]. DM identik dengan gula darah yang tinggi. Secara teori, gula darah yang tinggi dapat merusak dinding arteri sehingga mempercepat proses aterosklerosis [29]. Selain itu, kadar gula darah tinggi yang berlangsung lama akan berisiko terhadap meningkatnya kadar kolesterol, trigliserida sehingga trombus terbentuk dan mempercepat aterosklerosis [28]. Kolesterol yang tinggi dapat menempel pada pembuluh darah dan mengakibatkan terhambatnya aliran darah ke jantung. Selain itu, oksigen dalam darah juga akan terhambat sehingga jantung akan kekurangan suplai darah dan oksigen yang mengakibatkan kerja jantung lemah [30].

b. Inaktivitas Fisik

Inaktivitas fisik merupakan *node* internal pada pohon keputusan. Atribut inaktivitas fisik merupakan cabang hasil pemisahan data berdasarkan nilai atribut dari *node* akar atau DM. Inaktivitas fisik identik dengan terjadinya obesitas. Obesitas dapat menyebabkan otot jantung tidak bisa bergerak dengan baik sehingga meningkatkan risiko untuk terkena penyakit jantung koroner [3]. Aktivitas fisik dapat menurunkan tonus saraf simpatik, mendorong penurunan berat badan, dan meningkatkan metabolisme tubuh sehingga peredaran darah menjadi lebih lancar. Selain itu, dapat mengurangi peningkatan kadar kolesterol, hipertensi, berat badan, dan peningkatan HDL [31]. Namun, orang yang tidak melakukan aktivitas fisik atau inaktivitas fisik akan berisiko lebih besar untuk terkena PJK [32]. Inaktivitas fisik dapat menurunkan aliran darah dan fungsi endotel, sehingga mempercepat proses aterosklerosis. Endotel yang tidak sehat menyebabkan peningkatan adhesi sel inflamasi dan trombosit, sehingga mempercepat pembentukan plak dan risiko penyumbatan pada pembuluh darah [3].

c. Jenis Kelamin

Jenis kelamin merupakan *node* internal pada pohon keputusan yang membentuk 2 *node* internal. Berdasarkan jenis kelamin, baik laki-laki maupun perempuan memiliki faktor risiko terkena penyakit jantung koroner yang sama [33]. Laki-laki memiliki pola hidup yang tidak sehat seperti kebiasaan merokok dan mengonsumsi minuman beralkohol [12]. Kondisi ini dapat terjadi karena dalam rokok terdapat kandungan nikotin yang dapat merusak bagian endotel dinding pembuluh darah melalui pengeluaran katekolamin dan memudahkan terbentuknya trombus sehingga denyut jantung dan tekanan darah meningkat [34]. Perempuan juga berisiko terhadap PJK utamanya pada perempuan yang sudah menopause karena sudah tidak mendapatkan perlindungan dari hormon estrogen [12]. Pada perempuan hormon estrogen berfungsi untuk menjaga fleksibilitas aliran darah. Hormon ini bersifat melindungi sampai masa menopause tiba. Saat perempuan sudah memasuki masa menopause, hormon estrogen akan berkurang sehingga menyebabkan tingginya tingkat risiko penyakit jantung [33].

d. Obesitas

Obesitas merupakan *node* internal pada pohon keputusan yang membentuk 2 *node* internal, artinya atribut ini signifikan dalam memisahkan data, tetapi tidak mampu untuk membuat keputusan akhir pada level pertama. Secara teori, orang yang obesitas memiliki lemak yang berlebih. Lemak yang berlebih akan menumpuk dan menghasilkan plak. Seiring berjalannya waktu, dapat membentuk aterosklerosis. Faktor risiko penyakit jantung koroner adalah obesitas karena obesitas dapat berdampak pada peningkatan pembentukan plak aterosklerosis [35]. Selain itu, pada orang obesitas akan membutuhkan oksigen dan makanan ke dalam tubuh lebih banyak sehingga kerja jantung akan meningkat dalam memompa darah ke jaringan tubuh. Hal ini menyebabkan hipertrofi ventrikel kiri dan menurunkan fungsi jantung, meningkatkan risiko gagal jantung dan PJK [36].

e. Tekanan Darah Tinggi

Tekanan darah tinggi merupakan *node* internal pada pohon keputusan. Pada atribut ini, *node* internal yang terbentuk sebanyak 3. Secara teori, tekanan darah tinggi memaksa jantung untuk bekerja lebih keras. Akibatnya, otot jantung kiri membesar sehingga pemompaan darah di jantung menjadi tidak efisien dan dapat merusak jantung. Tingginya tekanan darah mengakibatkan pembuluh darah yang memberi asupan darah ke jantung atau pembuluh darah koroner mengalami aterosklerosis. Kondisi ini mengakibatkan aliran darah terhambat, sehingga asupan oksigen yang menuju jantung akan mengalami penurunan [3].

f. Usia

Usia merupakan *node* internal pada pohon keputusan. Atribut usia merupakan cabang hasil pemisahan data berdasarkan nilai atribut dari node internal inaktivitas fisik. Dari segi usia, laki-laki akan berisiko lebih awal daripada perempuan. Laki-laki berisiko terkena penyakit jantung koroner timbul 10 tahun lebih awal dari pada perempuan karena pada laki-laki tidak memiliki perlindungan dari hormon estrogen. Estrogen berfungsi untuk menjaga fleksibilitas aliran darah yang melindungi sampai masa menopause tiba [33]. Pada perempuan berisiko terkena PJK dengan nilai lebih rendah sebelum menopause jika dibandingkan laki-laki. Akan tetapi, usia yang terus meningkat dan terjadi penurunan hormon estrogen juga berisiko terhadap PJK[37].

g. Merokok

Merokok merupakan *node* internal pada pohon keputusan. Atribut merokok hanya membentuk node internal tunggal yang terletak di paling akhir. Rokok memiliki kandungan yang berbahaya. Kandungan seperti nikotin dan karbon monoksida (CO) pada rokok sangat berbahaya karena dapat mempercepat proses aterosklerosis. Nikotin dapat merusak bagian endotel dinding pembuluh darah dan memudahkan terbentuknya trombus sedangkan karbon monoksida menimbulkan desaturasi hemoglobin yang dapat menurunkan persediaan oksigen[37].

4. KESIMPULAN

Pre-processing data dilakukan dengan melakukan pemfilteran yaitu duplikasi data tidak ditemukan nomor rekam medis ganda, filter *missing value* dilakukan pembersihan terhadap atribut dislipidemia dan riwayat keluarga, dan pada filter inkonsistensi data tidak ditemukan. Proses klasifikasi menggunakan algoritma C4.5 dengan *tools* RapidMiner. Lalu mengimplementasikan teknik *k-fold cross validation* dengan melakukan percobaan sebanyak 5 kali dengan nilai $k=4$, $k=5$, $k=7$, $k=8$, dan $k=10$. Pembentukan pohon keputusan dilakukan dengan *pruning-prepruning* dan tanpa *pruning-prepruning*. Performa model lebih tinggi ketika *pruning-prepruning* diaktifkan dengan performa tertinggi diperoleh dari nilai $k=8$ dengan nilai *accuracy* sebesar 86,09%, *precision* sebesar 82,63%, dan *recall* sebesar 91,39%. Berdasarkan hasil performa tersebut, algoritma C4.5 dianggap mampu untuk mengklasifikasikan data dengan baik. Pada penelitian ini diabetes melitus merupakan atribut paling berpengaruh terhadap penyakit jantung koroner karena memperoleh nilai *gain ratio* tertinggi. Oleh sebab itu, DM menjadi *node* akar di pohon keputusan. Faktor risiko lain adalah inaktivitas fisik, jenis kelamin, obesitas, tekanan darah tinggi, usia, dan merokok. Penentuan faktor risiko diketahui pada hasil keputusan berupa *rules* klasifikasi. Pada penelitian ini memiliki keterbatasan yang perlu diperhatikan bagi peneliti-peneliti selanjutnya dalam menyempurnakan penelitian. Beberapa keterbatasan dalam penelitian ini, yaitu jumlah sampel keseluruhan hanya 302 yang terdiri dari penyakit jantung koroner dan *non* penyakit jantung koroner yang tentunya masih kurang untuk dapat mengklasifikasikan faktor risiko penyakit jantung koroner dengan baik. Pada *dataset* penelitian ini terdapat ketidaknormalan pada label *non* penyakit jantung koroner sebanyak 4 data, sehingga pohon keputusan menghasilkan *rules* yang perlu ditinjau ulang pada keputusan “Non PJK” yaitu pada *rules* keputusan nomor 13. Selain itu, keterbatasan penelitian terdapat pada sumber data yang digunakan yaitu rekam medis, karena hanya berfokus pada terisi atau tidaknya atribut yang digunakan. Kekurangan dari rekam medis yaitu ketidaklengkapan pengisian sehingga terdapat atribut yang mengalami *missing value* tinggi yaitu dislipidemia dan riwayat keluarga sehingga atribut tersebut dihapus.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada pihak-pihak yang telah mendukung terlaksananya penelitian ini khususnya pegawai RSD dr. Soebandi Jember di bagian unit rekam medis dan pelayanan poli rawat jalan spesialis jantung.

REFERENCES

- [1] A. Rafidah, “Analisis Faktor Risiko Kejadian Penyakit Jantung Koroner di RSUD Sultan Imanuddin Pangkalan Bun,” 2020.
- [2] F. Nuraisyah, J. Srikanthia Purnama, Y. Nuryanti, R. Dika Agustin, R. Desriani, and M. Utami Putri, “Edukasi Pengetahuan Penyakit Tidak Menular dan Gernas Pada Usia Produktif di Dusun Karangbendo,” *J. Panrita Abdi*, vol. 6, no. 1, pp. 1–7, 2022.
- [3] Kartini *et al.*, *Epidemiologi Penyakit Tidak Menular*, Pertama. Purbalingga: Eureka Media Aksara, 2023.
- [4] J. H. P. H. Tong, *Data Mining Concepts and Techniques Fourth Edition*, 4th ed. America: Katey Birtcher, 2023.
- [5] A. Purwinarko, K. Budiman, A. Widiyatmoko, F. A. Sasi, and W. Hardyanto, “Integrating C4.5 and K-Nearest Neighbor Imputation with Relief Feature Selection for Enhancing Breast Cancer Diagnosis,” *Sci. J. Informatics*, vol. 12, no. 1, pp. 107–118, 2025.
- [6] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, “Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5,” *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.
- [7] A. S. Sitio and F. A. Sianturi, “Peningkatan Kinerja Decision Tree C4.5 Melalui Pendekatan Mean Dan Median Pada Data Numerik,” vol. 2, no. 1, pp. 24–32, 2025.
- [8] A. F. Riany and G. Testiana, “Penerapan Data Mining untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naïve Bayes,” *MDP Student Conf.*, vol. 2, no. 1, pp. 297–305, 2023.
- [9] D. Larassati, A. Zaidiah, and S. Afrizal, “Sistem Prediksi Penyakit Jantung Koroner Menggunakan Metode Naive Bayes,” *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 7, no. 2, pp. 533–546, 2022.
- [10] B. G. Sudarsono and E. Winarno, “Perbandingan Algoritma Klasifikasi Data Mining Dalam Diagnosa Penyakit Arteri Koroner,”

- Build. Informatics, Technol. Sci.*, vol. 6, no. 3, pp. 1595–1604, 2024.
- [11] L. Nafi'ah and Z. Fatah, "Implementasi Algoritma Decision Tree Untuk Pendeteksian Penyakit Jantung," vol. 3, no. 2, pp. 160–165, 2024.
 - [12] L. F. Tampubolon, A. Ginting, and F. E. Saragi Turnip, "Gambaran Faktor yang Mempengaruhi Kejadian Penyakit Jantung Koroner (PJK) di Pusat Jantung Terpadu (PJT)," *J. Ilm. Permas J. Ilm. STIKES Kendal*, vol. 13, no. 3, pp. 1043–1052, 2023.
 - [13] Kepmenkes RI, "Keputusan Menteri Kesehatan RI, Nomor 854/MENKES/SK/IX/2009 tentang Pedoman Pengendalian Penyakit Jantung dan Pembuluh Darah." Menteri Kesehatan Republik Indonesia, 2009.
 - [14] Kemenkes RI, *Petunjuk Teknis Penatalaksanaan Penyakit Kardiovaskular Untuk Dokter*. Kuningan, Jakarta: Kementerian Kesehatan Republik Indonesia, 2017.
 - [15] Sugiyono, *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*. 2023.
 - [16] A. B. Alfianah, "Analisis Perbandingan Kinerja Algoritma K-Nearest Neighbor (KNN) dan Algoritma C4.5 Untuk Klasifikasi Teknik Persalinan," *Jember*, 2024.
 - [17] E. Mutiara, "Algoritma Klasifikasi Naive Bayes Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Tuberculosis (TB)," *Swabumi*, vol. 8, no. 1, pp. 46–58, 2020.
 - [18] I. D. Lestari, "Klasifikasi Penyakit Pneumonia dengan Menggunakan Metode Algoritma C5.4 di Rumah Sakit Umum Kaliwates Jember," *Jember*, 2024.
 - [19] A. A. A. Daniswara and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 05, pp. 97–100, 2023.
 - [20] S. A. K. A. Upe, "Faktor-Faktor yang Memengaruhi Inkonsistensi Penerapan Reward and Punishment pada Aparatur Sipil Negara di Badan Kepegawaian dan Pengembangan Sumber Daya Manusia Kabupaten Konawe," *J. Adm. Pembang. dan Kebijak. Publik Vol. 15 No. 1 Februari 2024*, vol. 15, no. 1, pp. 41–51, 2024.
 - [21] A. Zulaikha, "Klasifikasi Penyakit Stroke Iskemik Menggunakan Metode K-Nearest Neighbor (KNN) di RS Bhayangkara Bondowoso Tahun 2023," 2024.
 - [22] B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 1, pp. 13–21, 2021.
 - [23] T. T. P. Fatchiah, "Klasifikasi Penyakit Tuberkulosis Paru Menggunakan Algoritma C4.5 di Rumah Sakit Paru Jember," *Jember*, 2024.
 - [24] P. I. Nainggolan, D. S. Prasvita, and D. S. Bukit, "Klasifikasi Informasi Kesehatan Pada Data Media Sosial Menggunakan Support Vector Machine dan K-Fold Cross Validation," *Malikussaleh J. Mech. Sci. Technol.*, vol. 5, no. 2, p. 34, 2021.
 - [25] F. M. Hana and W. C. Wahyudin, "Optimasi Parameter Algoritma Decision Tree C4.5 Pada Klasifikasi Blogger Profesional," vol. 4, no. 2, pp. 77–83, 2023.
 - [26] C. N. Syahputri and M. S. Hasibuan, "Optimasi Klasifikasi Decision Tree dengan Teknik Pruning untuk Mengurangi Overfitting," vol. 11, no. 2, pp. 87–96, 2024.
 - [27] T. I. P. Hati, Rahman, and A. Rizaldy, "Penggunaan Decision Tree dalam Penentuan Faktor yang Mempengaruhi Status Gizi Buruk Balita di Kelurahan Tamamaung," vol. 14, no. 2, pp. 208–216, 2024.
 - [28] I. Rahmawati, D. Dwiana, R. S. Ratiyun, and Yesi, "Hubungan Diabetes Melitus (DM) dengan Penyakit Jantung Koroner (PJK) pada Pasien yang Berobat di Poli Jantung," *J. Kesehat. dr. Soebandi*, vol. 8, no. 1, pp. 56–62, 2020.
 - [29] R. T. Elkurnia, Cokorde Istri Yuliandari Krisawardani K, Luh Oliva Saraswati Suastika, and Cyndiana Widia Dewi Sinardja, "Prevalensi dan faktor risiko penyakit jantung koroner pada pasien diabetes melitus tipe 2 Rumah Sakit Umum Pusat Prof. Dr. I.G.N.G. Ngoerah Denpasar," *Intisari Sains Medis*, vol. 14, no. 1, pp. 47–52, 2023.
 - [30] M. S. Putri, "Literature Review Kadar Kolesterol dan Kadar Gula Darah dengan Kejadian Penyakit Jantung Koroner," Yogyakarta, 2020.
 - [31] I. Pashar, Sultan, and L. Wendikbo, "Hubungan Pola Makan Dan Aktivitas Fisik Dengan Penyakit Jantung Koroner Di Rsud Labuang Baji Makassar," *J. Pubnursing Sci.*, vol. 2, no. 01, pp. 31–42, 2024.
 - [32] C. Rachmawati, S. Martini, and K. D. Artanti, "Analisis Faktor Risiko Modifikasi Penyakit Jantung Koroner Di RSU Haji Surabaya Tahun 2019 Modification Risk Factorsa Analysis in Coronary Heart Disease in Haji Hospital Surabaya in 2019," *Media Gizi Kesmas*, vol. 10, no. 1, pp. 47–55, 2021.
 - [33] L. Bachtiar, R. A. Gustaman, and S. Maywati, "Faktor Risiko yang Berhubungan dengan Kejadian Penyakit Jantung Koroner (Analisis Data Sekunder di Rumah Sakit Umum Daerah Kabupaten Subang)," *J. Kesehat. komunitas Indones.*, vol. 19, no. 1, p. 52, 2023.
 - [34] Aisyah, F. R. Hardy, T. Y. R. Pristya, and U. Q. Karima, "Kejadian Penyakit Jantung Koroner pada Pasien di RSUD Pasar Rebo," *HIGEIA (Journal Public Heal. Res. Dev.)*, vol. 6, no. 4, pp. 250–260, 2022.
 - [35] J. Haldy and L. M. Kurniawidjaja, "Coronary Heart Disease Risk and Associated Risk Factor Among Workers at PT.X in 2023," *Indonesia J. Heal. Promot.*, vol. 7, pp. 1–11, 2024.
 - [36] Diah Tika Anggraeni *et al.*, "Pemberdayaan Kader Kesehatan Melalui Program 'Oke Heart' Sebagai Upaya Deteksi Dini Pencegahan Gagal Jantung di Kecamatan Limo, Depok," vol. 8, p. 6, 2025.
 - [37] I. Johannis, I. A. Tedju Hinga, and A. B. Sir, "Faktor Risiko Hipertensi, Merokok dan Usia terhadap Kejadian Penyakit Jantung Koroner pada Pasien di RSUD Prof. Dr. W. Z. Johannes Kupang," *Media Kesehat. Masy.*, vol. 2, no. 1, pp. 33–40, 2020.