

# Classification of Coronary Heart Disease at Semen Padang Hospital using Algorithm Classification And Regression Trees (CART)

Defal Aditya Defran, Atus Amadi Putra\*, Dodi Vionanda, dan Tessy Octavia Mukhti

Departemen Statistika, Universitas Negeri Padang, Kota Padang, Negara Indonesia

\*Corresponding author: [atusamadiputra@fmipa.unp.ac.id](mailto:atusamadiputra@fmipa.unp.ac.id)

Submitted : 28 September 2023

Revised : 25 Oktober 2023

Accepted : 27 Oktober 2023

## ABSTRACT

*One of the heart diseases that is very popular today is coronary heart disease (CHD). The main factors that cause CHD include age, gender, hypertension, blood sugar and cholesterol. One method that can be used to group CHD is classification. Classification And Regression Trees (CART) is a decision tree that describes the relationship between a response variable and one or more predictor variables. The purpose of the CHD research is to determine the main defining variables and classification accuracy of CHD patient at Semen Padang Hospital. Based on the results of the optimal tree, the attribute that is the main characteristic in classifying CHD patients at Semen Padang Hospital is age. The determination of the classification results using the confusion matrix produced an accuracy value of 66.67%, a sensitivity of 56.52% for classifying CHD patients, and a specificity of 84.61% for classifying non-CHD patients.*

**Keywords:** CART, Classification, Coronary Heart Disease, Decision Tree.



This is an open access article under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2022 by author and Universitas Negeri Padang.

## I. PENDAHULUAN

Penyakit kardiovaskuler adalah suatu penyakit *degeneratif* yang disebabkan oleh menurunnya fungsi pembuluh darah dan jantung. Menurut WHO, penyakit kardiovaskuler menjadi penyebab utama kematian secara global. Pada tahun 2022 penyakit ini merenggut sekitar 17,9 juta. Salah satu penyakit jantung yang sangat populer pada saat ini ialah penyakit jantung koroner (PJK). PJK terjadi akibat adanya gangguan fungsi jantung akibat iskemia pada otot jantung yang mengakibatkan penyempitan pada pembuluh darah arteri (koroner), sehingga mengakibatkan kerusakan pada lapisan dinding pembuluh darah (*Aterosklerosis*) (Susanto dan Basuki, 2023).

Prevalensi PJK di Indonesia mengalami peningkatan yang signifikan. Data prevalensi PJK diperoleh dari Riset Kesehatan Dasar (Rikesdas) yang dilakukan setiap lima tahun sekali. Berdasarkan hasil Rikesdas 2018, angka prevalensi pasien jantung sebesar 1,5% (Kementerian Kesehatan, 2018). Angka ini mengalami peningkatan dibandingkan dengan hasil Rikesdas tahun 2013, dimana angka prevalensi pasien penyakit jantung pada tahun 2013 sebesar 0,5%. Sumatera Barat termasuk kedalam salah satu provinsi yang memiliki prevalensi pasien penyakit jantung lebih tinggi dibandingkan rata-rata nasional yaitu sebesar 1,6% (Kementerian Kesehatan, 2013).

Masalah PJK terus meningkat jumlahnya seiring berkembangnya gaya hidup yang tidak sehat di kalangan masyarakat. Faktor utama penyebab PJK, antara lain faktor usia, jenis kelamin, hipertensi, gula darah, dan kolesterol (Maspiyanti, 2015). PJK merupakan penyakit yang cukup berbahaya, karena dapat menyebabkan kematian mendadak. Namun sebnarnya dapat dicegah dengan cara sendiri maupun dengan dokter spesialis jantung (Bustan, 2007). Pada saat ini banyak masyarakat tidak menyadari bahwa diri mereka mempunyai risiko menderita PJK karena tidak mengetahui faktor-faktor PJK (Nurhidayat, 2011).

Klasifikasi dapat digunakan untuk mengkategorikan individu dalam kelompok risiko berdasarkan faktor-faktor risiko yang berhubungan dengan PJK. Klasifikasi merupakan pengelompokan terhadap sekumpulan data dalam bentuk label/kelas tertentu yang bersifat kategorik. Klasifikasi juga berfungsi untuk prediksi terhadap label/kelas data baru dengan model yang telah dibangun dan dapat memberikan hasil akhir berupa akurasi model yang dihasilkan. Klasifikasi pada PJK digunakan untuk mengkategorikan pasien ke dalam berbagai kategori berdasarkan faktor risiko

yang berkontribusi pada PJK, sehingga mendapatkan model yang dapat mengidentifikasi apakah pasien memiliki PJK atau tidak. Ada beberapa jenis metode klasifikasi, yaitu *Classification And Regression Tree* (CART), ID3, C4.5, *Chi-Square Automatic Detection* (CHAID), analisis regresi logistik, SVM dan lain sebagainya. Namun metode yang digunakan pada penelitian ini adalah klasifikasi dengan algoritma CART.

Algoritma CART merupakan suatu pohon keputusan yang digunakan untuk melihat hubungan suatu variabel terikat dengan satu atau lebih variabel bebas. Tujuan dari algoritma CART adalah untuk mengelompokkan dan memisahkan kelompok data dengan akurat. Kelebihan dari algoritma CART yaitu tidak adanya asumsi distribusi variabel bebas yang harus dipenuhi karena merupakan metode nonparametrik, CART memudahkan dalam memperoleh keputusan pada data yang bersifat multivariabel, dan hasil akhir CART memiliki bentuk sederhana dan mudah untuk diinterpretasikan. Berdasarkan latar belakang tersebut, dilakukan penelitian pasien PJK di Semen Padang Hospital yang bertujuan untuk menentukan variabel yang menjadi penciri utama dan ketetapan klasifikasi pasien PJK di Semen Padang Hospital.

## II. METODE PENELITIAN

### A. Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder, dimana data diambil dari data rekam medis pasien rawat inap Spesialis Jantung Tahun 2022 di Semen Padang Hospital. Variabel yang digunakan pada penelitian adalah umur ( $X_1$ ), jenis kelamin ( $X_2$ ), trombosit ( $X_3$ ), hipertensi ( $X_4$ ), merokok ( $X_5$ ), kadar gula darah ( $X_6$ ), kolesterol ( $X_7$ ) dan diagnosa penyakit jantung ( $Y$ ).

### B. Teknik Analisis Data

Teknik analisis data yang digunakan dalam penelitian ini adalah analisis CART menggunakan bantuan *software R Studio*. Adapun tahapan dalam analisis CART sebagai berikut.

#### 1. Analisis statistika deskriptif pada variabel penelitian

Statistika deskriptif adalah rangkuman informasi yang terkandung dalam kumpulan data penelitian agar lebih mudah dipahami dan dapat memberikan gambaran jelas tentang karakteristik data tersebut

#### 2. Data *Training* dan *Testing*

Menurut Musu dkk (2021), data latih adalah kumpulan data yang memiliki atribut/label yang digunakan untuk mengenal karakteristik kumpulan data sehingga dapat menghasilkan model data. Sedangkan data uji merupakan kumpulan data yang digunakan untuk memprediksi model dalam mengklasifikasikan data.

#### 3. Pembentukan pohon klasifikasi dengan langkah-langkah sebagai berikut.

##### a. Pemilahan variabel pemisah

Proses pemilihan variabel pemisah pada analisis CART dimulai dengan membagi data menjadi dua kelompok berdasarkan kriteria tertentu. Kriteria ini biasanya menggunakan *Indeks Gini*. Persamaan (1) menyajikan fungsi *Indeks Gini*.

$$i(t) = 1 - \sum_{j=1} P^2(j|t) \quad (1)$$

Tahap selanjutnya adalah menentukan kriteria *goodness of split* untuk menilai pemilihan terbaik dari seluruh simpul  $t$ . Menurut Breiman *et al* (1984) cara menentukan *goodness of split* disajikan pada Persamaan (2).

$$\phi(s, t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (2)$$

Pemilahan yang menghasilkan nilai *goodness of split* paling tinggi adalah pemilah yang terbaik.

##### b. Menentukan simpul terminal

Simpul  $t$  menjadi simpul terminal, apabila terdapat batasan minimum  $n$  seperti terdapat satu pengamatan pada setiap simpul anak. Jumlah minimum kasus suatu terminal pada pengamatan umumnya berjumlah kurang atau sama dengan 5 ( $n \leq 5$ ) atau sudah tercapainya batas kedalaman pohon maksimal.

##### c. Penandaan label kelas

Menurut Pratiwi dan Zain (2104) proses penandaan label kelas pada setiap simpul terminal dilakukan dengan aturan jumlah terbanyak dengan menggunakan Persamaan (3).

$$P(j|t) = \max_j \frac{N_j(t)}{N(t)} \quad (3)$$

#### 4. Memangkas pohon klasifikasi

Pemangkasan pohon dilakukan untuk mencegah terjadinya pohon yang besar dan sangat kompleks. Ukuran pohon yang terlalu besar dapat mengakibatkan *overfitting*, namun jika pengamatan pohon dibatasi dengan ketepatan

tertentu akan menyebabkan *underfitting*. Untuk mendapatkan standar pohon yang sesuai, dilakukan pemangkasan pohon berdasarkan ukuran *cost complexity minimum* pada Persamaan (4).

$$g_m(t) = \frac{R(t) - R(T_k)}{|\tilde{T}_k| - 1} \quad (4)$$

5. Menentukan pohon klasifikasi optimal

Ukuran pohon yang terlalu besar akan menghasilkan nilai kompleksitas yang tinggi karena hasil dari pohon yang dihasilkan cenderung rumit untuk diinterpretasikan. Oleh karena itu, perlu dilakukan pemilihan pohon optimal yang sederhana namun tetap memberikan nilai akurasi yang tinggi.

6. Ukuran ketetapan klasifikasi

*Confusion matrix* adalah suatu metode yang biasa digunakan untuk menghitung ketepatan model dalam sebuah klasifikasi. Tabel 1 menyajikan *confusion matrix*

**Tabel 1.** *Confusion Matrix*

Classification		Predicted Class	
		Class = YES	Class = NO
Aktual Class	Class = YES	True Positif (TP)	False Positif (FP)
	Class = NO	False Negatif (FN)	True Negatif (TN)

(Sumber : Gorunescu, F. 2011)

Ada berbagai cara yang digunakan untuk mengukur ketepatan klasifikasi yaitu *sensitivity*, *specificity*, dan akurasi (Rokarch dan Maimon, 2015).

$$Sensitivity (\%) = \frac{TP}{TP+FN} \times 100\% \quad (5)$$

$$Specificity (\%) = \frac{TN}{FP+FN} \times 100\% \quad (6)$$

$$Accuracy (\%) = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (7)$$

### III. HASIL DAN PEMBAHASAN

#### A. Statistika Deskriptif

Gambaran karakteristik pasien penyakit jantung di Semen Padang Hospital dapat dilihat pada statistik deskriptif yang disajikan pada Tabel 2.

**Tabel 2.** Statistika Deskriptif

Variabel	Kategori	Jumlah
Umur ( $X_1$ )	<50	42
	$\geq$ 50	101
Jenis Kelamin ( $X_2$ )	Pria	85
	Wanita	58
Trombosit ( $X_3$ )	Normal	114
	Rendah	18
	Tinggi	11
Hipertensi ( $X_4$ )	Hipertensi 1	35
	Hipertensi 2	19
	Normal	39
	Pra-Hipertensi	50
Merokok ( $X_5$ )	Ya	67
	Tidak	76

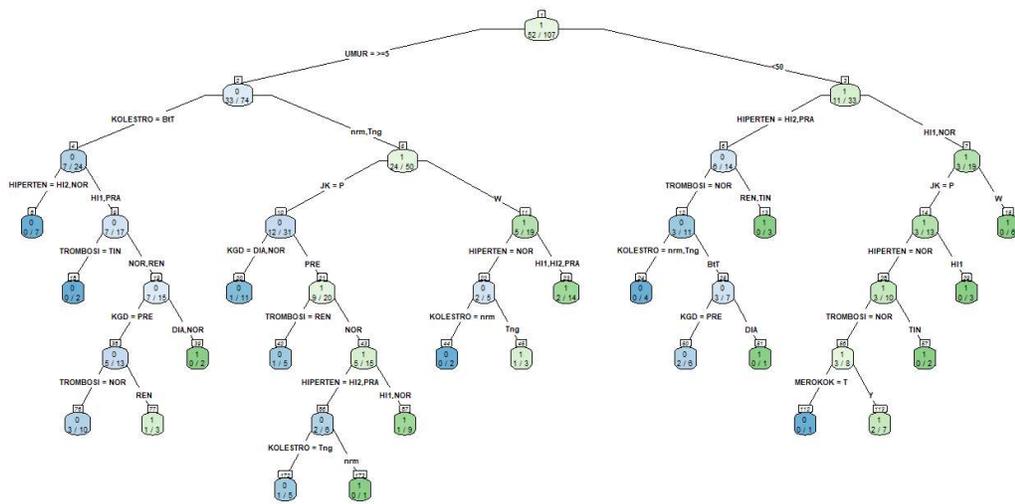
Variabel	Kategori	Jumlah
Kadar Gula Darah ( $X_6$ )	Diabetes	31
	Normal	16
	Pre-Diabetes	96
Kolesterol ( $X_7$ )	Batas Tinggi	53
	Normal	31
	Tinggi	59
Diagnosa Penyakit Jantung (Y)	PJK	67
	Bukan PJK	76

Berdasarkan Tabel 2, memberikan informasi bahwa dari 143 pasien penyakit jantung terdapat 47% atau 67 pasien terkena PJK dengan 85 dari 143 pasien penyakit jantung memiliki jenis kelamin pria. Mayoritas pasien penyakit jantung memiliki trombosit normal yaitu sebesar 114 pasien dan sebanyak 50 pasien menderita prahipertensi dengan 76 dari 143 pasien tidak merokok. Umumnya kadar gula darah pasien penyakit jantung tergolong prediabetes yaitu sebanyak 67,1% atau 96 pasien dan mengalami kolesterol yang tinggi sebanyak 41,3% atau 59 pasien.

**B. Pembentukan Pohon Klasifikasi**

Pembagian data menjadi dua bagian yaitu untuk keperluan membangun model di butuhkan data latih sedangkan untuk keperluan pengujian model digunakan data uji. Dari 143 data akan dibagi dengan komposisi 75% data latih dan 25% lagi dijadikan sebagai data uji, yaitu sebanyak 107 data latih dan 36 data uji. Selanjutnya data tersebut klasifikasi menggunakan CART.

Pohon klasifikasi yang maksimal merupakan suatu pohon klasifikasi dimana terdapat banyak jumlah simpul terminal. Metode pemilahan pada algoritma CART menggunakan Indeks Gini. Kemudian hasil perhitungan Indeks Gini pada Persamaan (1) digunakan sebagai penentuan kriteria *goodness of split* yang paling tinggi. Gambar 1 menyajikan konstruksi pohon yang maksimal.



**Gambar 1.** Pohon Klasifikasi Maksimal

Berdasarkan Gambar 1, dapat diketahui pohon klasifikasi maksimal yang dilatih berukuran besar, yaitu dengan 1 simpul akar, 20 simpul non terminal, dan 22 simpul terminal. Untuk memperoleh pohon yang optimal dilakukan pemangkasan terhadap pohon klasifikasi maksimal yang berukuran besar. Pohon yang mempunyai nilai *complexity parameter minimum* akan dilakukan pemangkasan. Dengan menggunakan Persamaan (4) maka didapatkan nilai *complexity parameter minimum* sebesar 0,0048. Setelah melakukan pemangkasan maka pohon klasifikasi yang diperoleh telah optimal, sebagaimana disajikan pada Gambar 2.



## DAFTAR PUSTAKA

- Bustan, M.N. (2007). *Epidemiologi Penyakit Tidak Menular*. Jakarta: Rineka Cipta.
- Gorunescu, Florin. (2011). *Data Mining: Concepts, Models and technicques*. Verlag Berlin Heiderberg: Springer.
- Kementerian Kesehatan. (2013). *Laporan Riset Kesehatan Dasar (Riskesdas) Bidang Biomedis*. Jakarta: Badan Litbangkes, Kemenkes RI.
- Kementerian Kesehatan. (2018). *Laporan Riset Kesehatan Dasar (Riskesdas) Bidang Biomedis*. Jakarta: Badan Litbangkes, Kemenkes RI.
- Maspiyanti, F. (2015). "Diagnosa Penyakit Jantung pada Ponsel Menggunakan Pohon Keputusan". *Jurnal Teknologi Terpadu*, 1(1).
- Nurhidayat Saiful. (2011). *Asuhan Keperawatan Klien dengan Gangguan Sistem Kardiovaskuler*. Ponorogo: UMPO Press.
- Pratiwi, F. E., & Zain, I. (2014). "Klasifikasi Pengangguran Terbuka Menggunakan CART (Classification and Regression Tree) di Provinsi Sulawesi Utara". *Jurnal Sains dan Seni ITS*, 3(1), D54-D59.
- Rokarch, L., dan Maimon, O. (2015). *Data Mining With Decision Trees Theory And Application 2nd Ed*. Singapore: World Scientific.
- Susanto, F., & Basuki, S. P. H. (2023). "Sosialisasi Deteksi Resiko Penyakit Jantung Koroner dengan Pemeriksaan Radiologi CT Calsium Score". *Prosiding Deminar Nasional LPPM UMP*, 4, 41-44.