Volume 4. No.1, Juni 2024 P-ISSN2715-6842 E-ISSN 2809-0454
JURMIK (Jurnal Rekam Medis dan Manajemen Informasi Kesehatan)

Feature Selection Information Gain pada Klasifikasi Pasien Penyakit Jantung (*Heart Disease*)

Siska Narulita[1*], Priyo Nugroho Adi[2]

[1]Universitas Nasional Karangturi
Email: [1]siskanarulita84@gmail.com
[2]Institut Teknologi dan Bisnis Semarang
Email: [2]priyo.adi.edu@itbsemarang.ac.id

ABSTRACT

Heart disease, also known as cardiovascular disease, is a condition where there is a blockage or narrowing of blood vessels that can lead to heart attack, chest pain, or stroke. It needs appropriate medical treatment because this disease can be the cause of death. Data mining methods are helpful in diagnosing and treating heart disease. Data mining methods can play a major role in the process of improving the quality of care for heart disease patients, providing valuable information for informed decision-making regarding prevention and treatment. The data analysis process uses classification algorithms, namely Decision Tree (C4.5), Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) combined with feature selection information gain method. The results show that data mining methods are very useful in diagnosing and treating heart disease. The highest percentage of correct classifications for both models before and after the implementation of feature selection information gain was obtained by the RF algorithm, which amounted to 95.71%. However, the implementation of the feature selection information gain method in this study did not contribute significantly to improving the classification quality of each algorithm used.

Keywords*: data mining; feature selection; information gain*

.

ABSTRAK

Heart disease atau penyakit jantung, atau juga dikenal dengan sebutan penyakit kardiovaskular merupakan suatu keadaan dimana terjadi adanya penyumbatan atau penyempitan pembuluh darah yang dapat mengakibatkan serangan jantung (*heart attack*), nyeri pada dada, ataupun stroke. Perlu adanya penanganan medis yang sesuai karena penyakit ini dapat menjadi penyebab kematian. Metode *data mining* sangat membantu dalam mendiagnosis dan pengobatan penyakit jantung. Metode *data mining* dapat memainkan peran utama dalam proses peningkatan kualitas perawatan untuk pasien penyakit jantung, memberikan informasi yang berharga untuk pembuatan keputusan yang tepat terkait pencegahan dan pengobatannya. Proses analisis data menggunakan algoritma klasifikasi, yaitu *Decision Tree* (C4.5), *Naive Bayes* (NB), *Support Vector Machine* (SVM), dan *Random Forest* (RF) yang dikombinasikan dengan metode *feature selection information gain*. Hasil penelitian menunjukkan bahwa metode *data mining* sangat bermanfaat dalam mendiagnosis dan pengobatan penyakit jantung. Persentase pengklasifikasian dengan benar tertinggi baik untuk model sebelum dan sesudah implementasi *feature selection information gain* diperoleh algoritma RF, yaitu sebesar 95,71%. Namun, implementasi dari metode *feature selection information gain* pada penelitian ini tidak memberikan kontribusi yang signifikan terhadap peningkatan kualitas klasifikasi dari setiap algoritma yang digunakan.

Kata kunci: *data mining; feature selection; information gain*

PENDAHULUAN

Heart disease atau penyakit jantung, atau juga dikenal dengan sebutan penyakit kardiovaskular (1) merupakan suatu keadaan dimana terjadi adanya penyumbatan atau penyempitan pembuluh darah yang dapat mengakibatkan serangan jantung (*heart attack*), nyeri pada dada (*angina*), ataupun stroke (2). Jantung merupakan organ vital yang mempunyai fungsi untuk memompa darah ke seluruh tubuh. Jika terjadi permasalahan pada jantung, maka proses peredaran darah di dalam tubuh dapat mengalami gangguan (2). Perlu adanya penanganan medis yang sesuai karena penyakit ini dapat menjadi penyebab kematian. Penyakit jantung itu sendiri masuk ke dalam golongan penyakit yang tidak menular (PTM) (3). Sekitar 71% penyebab kematian di dunia adalah PTM pada tahun 2016, 35% diantaranya disebabkan disebabkan penyakit jantung dan pembuluh darah (4). Perlunya upaya deteksi dini agar penanganan dapat segera dilakukan (5).

Metode *data mining* sangat membantu dalam mendiagnosis dan pengobatan penyakit jantung. Dengan menganalisis data medis dan menemukan pola atau hubungan yang relevan, metode *data mining* dapat membantu mengidentifikasi faktor risiko, menilai keberhasilan pengobatan, dan mencegah terjadinya komplikasi di kemudian hari. Selain itu, metode *data mining* juga dapat membantu studi epidemiologi, yaitu cabang dari ilmu kesehatan yang menganalisis sifat dan penyebaran berbagai permasalahan kesehatan dalam suatu penduduk tertentu, serta mempelajari penyebab timbulnya permasalahan (6) dengan cara menganalisis pola penyebaran penyakit dan mengidentifikasi kelompok yang beresiko tinggi, serta membantu mengoptimalkan proses perawatan kesehatan dan mengidentifikasi alternatif pengobatan yang paling efektif. Metode *data mining* dapat memainkan peran utama dalam proses peningkatan kualitas perawatan untuk pasien penyakit jantung, memberikan informasi yang berharga untuk pembuatan keputusan yang tepat terkait pencegahan dan pengobatannya.

Penelitian ini menunjukkan bahwa penggunaan metode *data mining* dapat membantu proses diagnosis dan pengobatan pada penyakit jantung. Pada penelitian ini dilakukan pengujian pada beberapa algoritma klasifikasi. Penelitian ini juga menunjukkan hasil yang lebih baik setelah menerapkan metode *feature selection*. Hasil penelitian yang dilakukan dapat membantu meningkatkan diagnosis dan mendukung pemilihan algoritma klasifikasi yang paling efektif untuk permasalahan diagnosis penyakit ini.

METODE

Dalam penelitian ini menggunakan *dataset heart disease prediction*, yang diperoleh dari UCI *Machine Learning Repository*. *Dataset* ini merangkum berbagai metrik kesehatan dari pasien jantung, seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, fitur elektrokardiografi (EKG), detak jantung, dan lain sebagainya. Tujuan penelitian ini adalah untuk mengembangkan model prediksi yang mampu mengidentifikasi pasien dengan penyakit jantung secara akurat. Setiap pasien dideskripsikan dengan 14 atribut, yaitu usia pasien (*age*), jenis kelamin pasien (*sex*), jenis nyeri dada (cp), tekanan darah saat istirahat (trestbps), kolesterol serum (chol), gula darah saat puasa (fbs), hasil elektrokardiografi saat istirahat (restecg), denyut jantung maksimum tercapai (thalach), angina yang disebabkan olahraga (exang), depresi ST yang disebabkan oleh olahraga relatif terhadap istirahat (*oldpeak*), kemiringan ST/LT maksimal pada uji latih jantung (*slope*), ca, thal, dan target.
Pasien dibedakan menjadi 2 kategori yaitu pasien yang diprediksi tidak mempunyai penyakit jantung (0) dan pasien yang diprediksi mempunyai penyakit jantung (1). Pada dataset, jumlah pasien yang dikategorikan tidak terindikasi penyakit jantung sebanyak 138 pasien, sedangkan pasien yang terindikasi penyakit jantung sebanyak 165 pasien.

Perhitungan dilakukan menggunakan *software* RapidMiner Studio Version 9.10. Tahap awal penelitian, *dataset* diklasifikasi menggunakan algoritma berikut ini:

1. *Decision Tree* (C4.5) (7)
   Algoritma C4.5 merupakan perbaikan dari ID3. Algoritma C4.5 menangani fitur atau atribut bertipe numerik, dapat melakukan *pruning decision tree* (pemotongan), dan *deriving rule set* (penurunan). Algoritma ini menggunakan kriteria *gain* dalam penentuan fitur sebagai pemecah node pada pohon yang diinduksi. Pada algoritma C4.5 dalam membangun decision tree, yang dilakukan pertama kali adalah memilih fitur sebagai akar. Dari setiap nilai pada akar tersebut dibuat cabang. Selanjutnya, melakukan pembagian kasus dalam cabang. Ulangi proses pada setiap cabang, sampai semua kasus pada cabang mempunyai kelas yang sama.

2. *Naive Bayes* (NB) (8)
   Naive Bayes merupakan pengklasifikasi probabilistik yang didasarkan pada teorema *Bayes* dan asumsi independensi fitur. Singkatnya, *Naive Bayes* membuat model probabilistik yang menetapkan kelas untuk objek berdasarkan karakteristik objek tersebut. Algoritma *Naive Bayes* bekerja dengan membandingkan probabilitas bahwa suatu objek termasuk ke dalam kelas yang berbeda, berdasarkan fitur-fitur yang ada pada objek tersebut. *Naive Bayes* melakukan perhitungan probabilitas bersyarat, yaitu probabilitas dari setiap fitur pada setiap kelas menggunakan *data training*. Sedangkan untuk menetapkan kelas ke objek baru, *Naive Bayes* melakukan perhitungan nilai fungsi probabilitas untuk setiap kelas dan kemudian menetapkan objek baru ke kelas dengan nilai tertinggi dari fungsi ini. Algoritma *Naive Bayes* dianggap sebagai algoritma klasifikasi yang sederhana dan efisien yang bekerja dengan baik pada *dataset* yang besar. Salah satu keunggulan utama algoritma ini adalah kecepatan dan kebutuhan komputasinya yang rendah, sehingga ideal untuk aplikasi *real-time*.

3. *Support Vector Machine* (SVM) (8)
   Tujuan utama dari algoritma SVM adalah menemukan *hyperplane* dengan *margin* maksimum untuk memisahkan *data input* yang berasal dari kelas berbeda. *Margin* merupakan jarak antara *hyperplane* dengan titik-titik terdekat dari *data training* yang disebut vektor pendukung (*support vectors*). Algoritma digunakan diberbagai bidang, seperti pengenalan gambar (*image recognition*), analisis teks (*text analysis*), klasifikasi bioinformatika, keuangan dan teknik mesin. Terdapat dua jenis SVM, yaitu linier dan non-linier, perbedaannya dalam hal bagaimana pendefinisian hyperplane. SVM rentan terhadap permasalahan ketinggian angka (*height number problem*), hal ini berarti membutuhkan pemrosesan dan pemilihan fitur input yang tepat untuk mencapai hasil terbaik. Salah satu kelebihan dari SVM adalah kebal terhadap *overfitting*, yaitu mempunyai risiko *overfitting* yang rendah, terutama ketika digunakan pada proses normalisasi yang tepat. Sedangkan kelemahan dari SVM adalah memakan waktu yang lama untuk dipelajari, terutama pada dataset yang berukuran sangat besar dan untuk model non-linier yang lebih kompleks.

4. *Random Forest* (RF) (8)
   RF merupakan algoritma *machine learning* yang banyak digunakan dalam klasifikasi, regresi, dan deteksi anomali. RF didasarkan pada konsep pohon keputusan, yaitu kumpulan dari banyak pohon keputusan yang bekerja bersama sebagai pengklasifikasi tunggal. Dalam RF, setiap pohon keputusan dilatih pada subset data dan fitur yang berbeda untuk mencegah *overfitting*. Setiap pohon membuat prediksi berdasarkan pohon keputusannya, kemudian hasilnya digabungkan untuk membuat prediksi akhir. Kelebihan dari RF adalah akurasi

prediksi yang tinggi, ketahanan terhadap *overfitting*, efisiensi ketika bekerja dengan *dataset* yang besar, serta kemudahan interpretasi hasil dan kecepatan eksekusi.

Tahap kedua pada penelitian ini dilakukan seleksi fitur (*feature selection*) untuk mengidentifikasi fitur atau atribut yang mempunyai dampak paling signifikan terhadap hasil prediksi. Dimensi *dataset* direduksi menggunakan teknik atau metode *feature selection information gain*. Metode *feature selection* ini digunakan untuk menentukan seberapa besar kontribusi suatu fitur atau atribut dalam meningkatkan akurasi prediksi dalam proses analisis data. *Information gain* menggunakan entropi untuk menentukan banyaknya data yang tidak terstruktur dan kompleksitas informasinya. *Information gain* digunakan untuk mengurangi dimensi data melalui pemilihan fitur atau atribut terpenting yang mempunyai dampak paling besar pada hasil prediksi. Metode ini sangat berguna untuk *dataset* berukuran besar, dimana dengan menemukan fitur atau atribut terpenting, dapat mempercepat proses analisis data dan meningkatkan efisiensi prediksi secara signifikan (8).

Selanjutnya *dataset* diklasifikasi menggunakan algoritma yang sama pada tahap awal. Metrik pengukuran yang digunakan untuk mengevaluasi algoritma klasifikasi tersebut adalah:

1. TP *Rate*, merupakan metrik kinerja yang digunakan untuk mengevaluasi efektivitas model klasifikasi dalam *machine learning*. TP *Rate* merupakan perbandingan prediksi benar positif terhadap keseluruhan prediksi positif (9).
2. *Precision*, menggambarkan tingkat akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model yang terbentuk. Dengan kata lain, *precision* adalah rasio prediksi benar positif dibanding keseluruhan hasil prediksi positif (10).
3. *Recall*, menggambarkan kesuksesan model dalam menemukan kembali sebuah informasi. *Recall* merupakan rasio prediksi benar positif dibanding keseluruhan data benar positif (10).
4. F-*Measure*, merupakan indikator pengukuran kualitas model.
5. ROC *Area*, merupakan area yang menunjukkan kinerja model klasifikasi pada semua ambang batas (*thresholds*) klasifikasi.

HASIL

Pada tahap awal penelitian, model dievaluasi menggunakan semua fitur atau atribut yang tersedia. Tabel 1 menunjukkan hasil prediksi yang dilakukan oleh algoritma C4.5.

Tabel 1. Parameter Model - Algoritma C4.5

Metrik	Nilai
TP *Rate*	0,99
Precision	0,88
Recall	0,99
F-*Measure*	0,93
ROC *Area*	0,96

Metrik kinerja algoritma NB ditunjukkan pada tabel 2 berikut:

Tabel 2. Parameter Model - Algoritma *Naive Bayes*

Metrik	Nilai
TP *Rate*	0,88
Precision	0,83
Recall	0,88
F-*Measure*	0,86
ROC *Area*	0,84

Sedangkan metrik kinerja untuk algoritma SVM dan RF sebelum penerapan *feature selection* ditunjukkan pada tabel 3 dan 4.

Tabel 3. Parameter Model - Algoritma SVM

Metrik	Nilai
TP *Rate*	0,95
Precision	0,80
Recall	0,95
F-*Measure*	0,86
ROC *Area*	0,84

Tabel 4. Parameter Model - Algoritma RF

Metrik	Nilai

TP *Rate*	0,99
Precision	0,94
Recall	0,99
F-*Measure*	0,96
ROC *Area*	0,92

Setelah dilakukan *training* pada *dataset* menggunakan beberapa algoritma klasifikasi, dilakukan implementasi metode *feature selection information gain* pada *dataset* sebelum proses implementasi pada algoritma klasifikasi. Metrik pengukuran yang digunakan untuk mengevaluasi algoritma tersebut sama dengan metrik pengukuran sebelum digunakan *metode feature selection information gain*. Nilai metrik pengukuran hasil dari implementasi feature selection information gain pada dataset yang kemudian diimplementasikan pada beberapa algoritma klasifikasi ditunjukkan pada tabel-tabel berikut ini:

Tabel 5. Parameter Model - IG + C4.5

Metrik	Nilai
TP *Rate*	0,99
Precision	0,88
Recall	0,99
F-*Measure*	0,93
ROC *Area*	0,96

Dari tabel 5 di atas, semua nilai metrik pengukuran pada perbandingan implementasi algoritma C4.5 terhadap implementasi *feature selection information gain* dan algoritma C4.5 adalah sama.

Selanjutnya dilakukan pengukuran metrik kinerja dari implementasi *feature selection information gain* dengan algoritma NB ditunjukkan pada tabel 6.

Tabel 6. Parameter Model - IG + NB

Metrik	Nilai
TP *Rate*	0,88
Precision	0,83
Recall	0,88
F-*Measure*	0,86
ROC *Area*	0,91

Pada tabel 6 nampak bahwa nilai metrik kinerja pada implementasi algoritma NB sama ketika menggunakan *feature selection information gain*, kecuali pada ROC *area*, dimana pada implementasi algoritma NB mempunyai nilai ROC *area* yang lebih kecil, yaitu 0,84. Sedangkan ketika mengimplementasikan *feature selection information gain* nilai ROC *area*nya lebih besar, yaitu 0,91.

Tabel 7. Parameter Model - IG + SVM

Metrik	Nilai
TP *Rate*	0,95
Precision	0,80
Recall	0,95
F-*Measure*	0,86
ROC *Area*	0,84

Sama halnya pada perbandingan nilai metrik kinerja algoritma NB baik sebelum dan sesudah penggunaan *feature selection information gain*, perbandingan nilai metrik kinerja algoritma SVM sebelum dan sesudah penggunaan metode *feature selection information gain* diperoleh nilai yang sama, namun pada implementasi algoritma SVM nilai ROC *area* juga sama, dimana sebelum implementasi *feature selection information gain* nilai ROC *area* sebesar 0,84, sedangkan setelah implementasi juga mempunyai nilai 0,84, ditunjukkan pada tabel 7.

Tabel 8. Parameter Model - IG + RF

Metrik	Nilai
TP *Rate*	0,99
Precision	0,94
Recall	0,99
F-*Measure*	0,96
ROC *Area*	0,96

Pada tabel 6 ditunjukkan bahwa sebelum dan sesudah implementasi *feature selection information gain* pada algoritma RF nilai metrik pengukuran semua sama, kecuali nilai ROC *area*. Sebelum implementasi *feature selection information gain*, nilai ROC *area* sebesar 0,92, sedangkan setelah penerapannya diperoleh nilai 0,96.

PEMBAHASAN

Tabel 9. Perbandingan Model Sebelum Penggunaan *Feature Selection*

Pengukuran	C4.5	NB	SVM	RF
Correctly Classified Instances	92,08 %	83,83 %	83,83 %	95,71%
Classification error	7,92 %	16,17 %	16,17 %	4,29 %
Kappa	0,84	0,67	0,67	0,91

Tabel 9 menyajikan perbandingan model sebelum adanya penggunaan *feature selection*.

Tabel 10. Perbandingan Model Setelah Penggunaan *Feature Selection*

Pengukuran	C4.5	NB	SVM	RF
Correctly Classified Instances	92,08 %	83,83 %	83,83 %	95,71%
Classification error	7,92 %	16,17 %	16,17 %	4,29 %
Kappa	0,84	0,67	0,67	0,91

Tabel 10 menyajikan perbandingan model setelah adanya penggunaan *feature selection*. Pada tabel 9 dan 10 disajikan perbandingan model-model yang terbentuk. Terlihat bahwa persentase pengklasifikasian dengan benar (ditunjukkan oleh nilai *correctly classified instances*) tertinggi baik untuk model sebelum dan sesudah implementasi *feature selection information gain* diperoleh algoritma RF, yaitu sebesar 95,71%. Dari tabel 9 dan 10 juga terlihat bahwa implementasi *feature selection information gain* tidak berkontribusi pada peningkatan kualitas klasifikasi dari setiap algoritma pada penelitian yang dilakukan.

SIMPULAN

Hasil penelitian menunjukkan bahwa metode *data mining* sangat bermanfaat dalam mendiagnosis dan pengobatan penyakit jantung. Dalam penelitian ini, dilakukan analisis terhadap 303 data pasien yang didiagnosis menderita penyakit jantung menggunakan beberapa algoritma klasifikasi dan metode *feature selection information gain*. Namun, implementasi dari metode *feature selection information gain* pada penelitian ini tidak memberikan kontribusi yang signifikan terhadap peningkatan kualitas klasifikasi dari setiap algoritma yang digunakan. Untuk penelitian selanjutnya dapat dilakukan menggunakan metode *feature selection* lainnya, seperti *chi-square*, *forward selection*, *backward selection*, atau metode *feature selection* lainnya.

DAFTAR PUSTAKA

1. Lazulfa I, F. RAJ. Analisis Faktor Prediksi Diagnosis Tingkat Keparahan Penyakit Jantung (Heart Disease) Menggunakan Metode Stepwise Binary Logistic Regression. Inov J Ilm Inov Teknol Inf. 2017;2(1):1–8.
2. Kementerian Kesehatan Republik Indonesia. Pathfinder: Kardiovaskular [Internet]. 30 Januari 2023. 2023. Available from: https://perpustakaan.kemkes.go.id
3. Kementrian Kesehatan Republik Indonesia. Mengenal Penyakit Tidak Menular dan Pencegahannya [Internet]. 25 Agustus 2022. 2022. Available from: https://ayosehat.kemkes.go.id
4. Direktorat Pencegahan dan Pengendalian Penyakit Tidak Menular. Buku Pedoman Manajemen Penyakit Tidak Menular [Internet]. Jakarta: Kementerian Kesehatan Republik Indonesia; 2019. Available from: https://p2ptm.kemkes.go.id
5. Widiastuti E, Saragih B, Fatchanuraliyah, Usman IKS, Hamzah A, Junita TV, et al. Lebih Awal Lebih Baik: Pencegahan dan Pengendalian Penyakit Tidak Menular (PTM) [Internet]. Jakarta: Kementerian Kesehatan Republik Indonesia; 2022. Available from: https://p2p.kemkes.go.id
6. Haryono, Rubaya AK, Husein A. Pengantar Epidemiologi. Yogyakarta: Poltekkes Jogja Press; 2021.
7. Muslim MA, Prasetiyo B, Mawarni ELH, Herowati AJ, Mirqotussa'adah, Rukmana SH, et al. Data Mining Algoritma C4.5. Semarang; 2019.
8. Zdrodowska M, Kasperczuk A, Dardzińska-Głębocka A. Selected Feature Selection Methods for Classifying Patients with Hepatitis C. In: 27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023). Poland: Procedia Computer Science; 2023. p. 3710–7.
9. Rahayu PW, Sudipa IGI, Suryani, Surachman A, Ridwan A, Darmawiguna

IGM, et al. Buku Ajar Data Mining. Bandung: PT Sonpedia Publishing Indonesia; 2024.
10. Tholib A. Implementasi Algoritma Machine Learning Berbasis Web dengan Framework Streamlit. Probolinggo: Pustaka Nurja; 2023.