INSANtek – Jurnal Inovasi dan Sains Teknik Elektro
Volume 2 No. 1 Mei 2021
E-ISSN: 2722-547X

Optimasi Algoritma C4.5 dan Naïve Bayes Berbasis Particle Swarm Optimization Untuk Diagnosa Penyakit Peradangan Hati

Amrin[1], Omar Pahlevi[2], Irawan Satriadi[3]

[1,2,3]Universitas Bina Sarana Informatika
e-mail: amrin.ain@bsi.ac.id, omar.opi@bsi.ac.id, irawan.irs@bsi.ac.id

Abstrak - Peradangan hati merupakan salah satu penyakit menular yang menjadi masalah kesehatan masyarakat yang berpengaruh terhadap angka kesakitan, angka kematian, status kesehatan masyarakat, angka harapan hidup, dan dampak sosial ekonomi lainnya. Melakukan diagnosa dini pada penyakit ini adalah sesuatu yang sangat penting agar dapat secara cepat ditangani dan diobati. Pada penelitian ini penulis akan mengaplikasikan dan membandingkan beberapa metode klasifikasi data mining dan optimasi dengan *particle swarm optimization* (pso), diantaranya Algoritma C4.5, *Naïve Bayes*, C4.5 dengan pso, dan *Naïve Bayes* dengan pso untuk mendiagnosis penyakit peradangan hati, kemudian membandingkan mana dari beberapa metode tersebut yang paling akurat. Berdasarkan hasil penelitian, diketahui bahwa metode C4.5 dengan pso merupakan metode terbaik dengan akurasi 79,51% dan nilai under the curva (AUC) 0,950, kemudian metode *Naive Bayes* dengan pso memiliki akurasi 79,28% dan nilai AUC sebesar 0,739, kemudian metode C4.5 dengan tingkat akurasi sebesar 70,99% dan nilai AUC sebesar 0,950, selanjutnya metode *Naive Bayes* dengan tingkat akurasi sebesar 66,14%, dan nilai AUC sebesar 0,742. Hal ini membuktikan bahwa optimasi particle swarm optimization dapat meningkatkan kinerja metode klasifikasi yang digunakan.

Kata Kunci: *C4.5, Naive Bayes, particle swarm optimization, confusion matrix, ROC Curva*

PENDAHULUAN

Penyakit peradangan hati (liver) atau biasa juga disebut hepatitis merupakan salah satu penyakit menular yang menjadi masalah kesehatan masyarakat yang berpengaruh terhadap angka kesakitan, angka kematian, status kesehatan masyarakat, angka harapan hidup, dan dampak sosial ekonomi lainnya. Besaran masalah Hepatitis di Indonesia dapat diketahui dari berbagai studi, kajian, maupun kegiatan pengamatan penyakit. Hepatitis adalah peradangan hati yang bisa berkembang menjadi fibrosis (jaringan parut), sirosis atau kanker hati. Hepatitis disebabkan oleh berbagai faktor seperti infeksi virus, zat beracun (misalnya alkohol, obat-obatan tertentu), dan penyakit autoimun. Penyebab paling umum Hepatitis adalah yang disebabkan oleh Virus Hepatitis B dan C (Kemenkes, 2017).

Penyakit hati merupakan salah satu penyakit yang menjadi masalah nasional di semua negara, baik di negara-negara berkembang seperti Indonesia maupun di negara-negara maju. Penyakit ini dapat terjadi pada semua golongan umur, mulai dari anak-anak, remaja, dewasa sampai orang tua. Data statistik menunjukkan bahwa penyakit hati adalah salah satu penyebab kematian yang utama, baik di Amerika Serikat maupun di seluruh dunia. Penderita penyakit hati sulit untuk dideteksi, terutama pada tahap awal penyakit. Hal ini dikarenakan pasien tidak merasakan gejala penyakit dan seakan-akan hati berfungsi secara normal, padahal sebagian hati sudah mengalami kerusakan (Hannan et al., 2010).

Akhir akhir ini dalam bidang medis, diagnosa penyakit peradangan hati menjadi hal yang agak sulit dilakukan. Namun terdapat catatan rekam medis yang menyimpan gejala-gejala penyakit pasien. Hal demikian tentu sangat bermanfaat bagi para tenaga medis atau dokter. Mereka dapat memanfaatkan catatan rekam medis sebelumnya sebagai bahan untuk mengambil keputusan tentang diagnosis penyakit pasien.

Teknik analisa konvensional secara manual yang selama ini digunakan tidak lagi efektif digunakan untuk mendiagnosa. Seiring dengan perkembangan sistem berbasis pengetahuan medis, tuntutan akan adanya penggunaan sistem pengetahuan berbasis komputer sebagai teknik analisa dalam mendiagnosa penyakit menjadi semakin penting. oleh karenanya, saat inilah waktu yang tepat untuk mengembangkan sistem pengetahuan berbasis komputer yang modern, efektif dan efisien dalam mendiagnosa penyakit (Neshat & Yaghoobi, 2009).

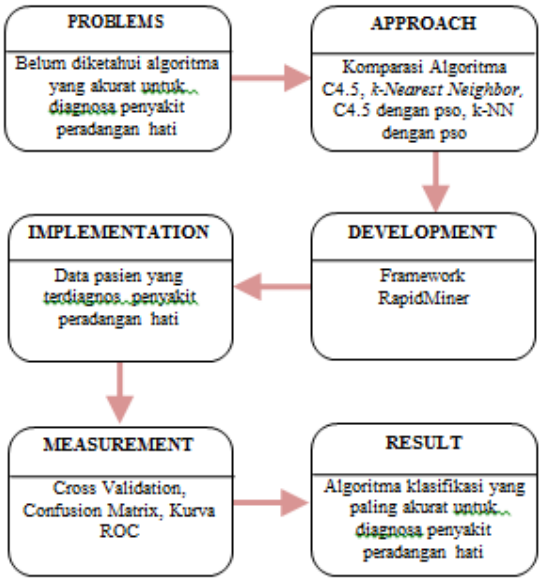
Pada penelitian ini, penulis akan menerapkan dan membandingkan metode klasifikasi data mining dan otimasi particle swarm optimization (pso), diantaranya yaitu Algoritma C4.5, *Naive Bayes*, C4.5 dengan pso, dan *Naive Bayes* dengan pso untuk mendiagnosa penyakit peradangan hati. Manakah metode tersebut yang paling akurat mendiagnosa penyakit peradangan hati. Penelitian ini diharapkan dapat membantu para tenaga kesehatan untuk mendiagnosa secara dini penyakit perdangan hati, dan secara luas dengan menggunakan aplikasi ini

diharapkan masyarakat secara umum dapat memprediksi apakah terdiagnosa penyakit peradangan hati atau tidak. Jikalau terdiagnosa, masyarakat dengan secepat mungkin berkonsultasi dan menghubungi dokter.

Sebagai bahan acuan dan perbandingan, peneliti melakukan study literature dengan beberapa penelitian terdahulu yang terkait dengan tema metode-metode yang pernah digunakan untuk menyelesaikan prediksi penyakit peradangan hati, diantaranya penelitian yang dilakukan oleh Varun Kumar (V. Kumar et al., 2012) dan kawan-kawan yang meneliti tentang Prediksi penyakit hepatitis dengan algoritma *Support Vector Machine (SVM)* dengan fitur seleksi.Pengukuran kinerja metode menggunakan *Chi-square*, Fitur seleksi, dan Framwork RapidMiner . Dari pengukuran kinerja yang dilakukan, diketahui bahwa akurasi metode SVM adalah sebesar 79,33%, sedangkan untuk SVM fitur seleksi memiliki akurasi sebesar 83,12%. Selanjutnya penelitian yang dilakukan oleh Bekir Karlik (Karlik, 2011) yang meneliti tentang Prediksi penyakit hepatitis dengan algoritma *Backpropagation dan Naïve Bayes*. Pengukuran kinerja metode menggunakan *10Fold Cross Valdiation*, *Confusion Matrix*, ROC Area. dan Framwork RapidMiner . Dari pengukuran kinerja yang dilakukan, diketahui bahwa akurasi metode *Backpropagation* adalah sebesar 98%, sedangkan untuk *Naïve Bayes* memiliki akurasi sebesar 86%. Berikutnya penelitian yang dilakukan oleh Senthil Kumar (D. S. Kumar et al., 2011) yang meneliti tentang Prediksi penyakit hepatitis, diabetes dan jantung dengan komparasi algoritma CART, ID3, dan C4.5. Pengukuran kinerja metode menggunakan *10Fold Cross Valdiation*, *Confusion Matrix*, ROC Area. dan Framwork RapidMiner . Dari pengukuran kinerja yang dilakukan, diketahui bahwa akurasi metode CART adalah sebesar 83,2%, ID3 mempunyai akurasi 64,8%, sedangkan untuk C4.5 memiliki akurasi sebesar 71,4%. Penelitian yang dilakukan oleh (Amrin & Pahlevi, 2020) tentang Data Mining Model For Designing Diagnostic Applications Inflammatory Liver Disease. Yang selanjutnya penelitian yang dilakukan oleh Novianto Donna Prayoga (Prayoga, 2018) dan kawan-kawan yang meneliti tentang Diagnosis penyakit hepatitis dengan metode *Naïve Bayes*. Dari pengujian tingkat akurasi yang dilakukan, diketahui bahwa akurasi metode Naïve Bayes adalah sebesar 87,50%.

METODE PENELITIAN

Penelitian ini terdiri dari beberapa tahap seperti terlihat pada kerangka pemikiran Gambar 1. Permasalahan pada penelitian ini adalah belum diketahui algoritma yang akurat untuk diagnosa penyakit peradangan hati. Untuk itu dibuat approach (model) yaitu algoritma C4.5, *Naive Bayes*, C4.5 dengan pso, dan *Naive Bayes* dengan pso untuk memecahkan permasalahan kemudian dilakukan pengujian terhadap kinerja dari beberapa metode tersebut. Pengujian menggunakan metode *Cross Validation*, *Confusion Matrix* dan kurva ROC. Untuk mengembangkan aplikasi (development) berdasarkan model yang dibuat, digunakan Rapid Miner.



Gambar 1. Kerangka Pemikiran Penelitian

HASIL DAN PEMBAHASAN

A. Analisa Data

Pada penelitian ini dataset yang digunakan adalah dari website UCI Machine Learning Repository yaitu Indian Liver Patient Dataset (ILPD). Dataset ini berisi data yang dikumpulkan dari para pasien yang ada di timur laut Andhra Pradesh, India. Setelah dilakukan teknik preprocessing data, maka dataset berisi 414 pasien penderita liver sedangkan 165 pasien bukan penderita liver. Dataset ini memiliki 10 atribut dimana 9 atribut merupakan atribut input sedangkan 1 atribut sebagai output atau class, dan memiliki 579 record. Adapun deskripsi atribut seperti ditunjukkan pada table 1 berikut:.

Tabel 1. Deskripsi Atribut

Atribut	Keterangan
Age	Umur pasien
TB	Total bilirubin pasien
DB	Direct bilirubin pasien
Alkphos	Alkaline phospotase
Sgpt	Almine Aminotransferase
Sgot	Aspartate Aminotransferase
TP	Total protiens
ALB	Albumin
A/G Ratio	Albumin dan Globulin ratio
Class	Class apakah pasien positif liver atau tidak

B. Pengujian Model

Penelitian ini dilakukan dengan eksperimen pengujian pada model yang diusulkan. Kemudian

dilakukan evaluasi dan validasi model untuk menghasilkan nilai accuracy dan AUC. Pengujian menggunakan Rapidminer dengan operator 10-fold cross-validation untuk mendapatkan hasil accuracy dan AUC pada setiap algoritma yang diuji. Evaluasi yang dilakukan adalah dengan Confusion Matrix dan ROC Curve atau Area Under Curve (AUC).

1. Confusion Matrix

a. Algoritma C4.5

Tabel 2 adalah *confusion matrix* untuk algoritma C4.5. Diketahui 407 data diklasifikasi "Yes" diprediksi sesuai dengan data sebenarnya, lalu 7 data diprediksi "No" tetapi ternyata "Yes". Kemudian 4 data diklasifikasi *"No"* diprediksi sesuai, dan 161 data diprediksi "Yes" ternyata "No".

Tabel 2. Model *Confusion Matrix* untuk Algoritma C4.5

accuracy: 70.99% +/- 1.98% (mikro: 70.98%)

	true Yes	true No	class precision
pred. Yes	407	161	71.65%
pred. No	7	4	36.36%
class recall	98.31%	2.42%	

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

b. *Naive Bayes*

Tabel 3 adalah *confusion matrix* untuk algoritma *Naive Bayes*. Diketahui 270 data diklasifikasi "Yes" tepat sesuai dengan data sebenarnya, lalu 144 data diprediksi "No" tetapi ternyata "Yes". Kemudian 113 data diklasifikasi"No" diprediksi sesuai, dan 52 data diprediksi "Yes" ternyata "No".

Tabel 3. Model *Confusion Matrix* untuk Metode *Naive Bayes*

accuracy: 66.14% +/- 4.55% (mikro: 66.15%)

	true Yes	true No	class precision
pred. Yes	270	52	83.85%
pred. No	144	113	43.97%
class recall	65.22%	68.48%	

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

c. Algoritma C4.5 dengan pso

Tabel 4 adalah *confusion matrix* untuk algoritma C4.5 dengan pso. Diketahui 320 data diklasifikasi "Yes" diprediksi sesuai dengan data sebenarnya, lalu 1 data diprediksi "No" tetapi ternyata "Yes". Kemudian 2 data diklasifikasi *"No"* diprediksi sesuai, dan 82 data diprediksi "Yes" ternyata "No".

Tabel 4. Model *Confusion Matrix* untuk Metode C4.5 dengan pso

accuracy: 79.51% +/- 1.10% (mikro: 79.51%)

	true Yes	true No	class precision
pred. Yes	320	82	79.60%
pred. No	1	2	66.67%
class recall	99.69%	2.38%	

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

d. *Naive Bayes* dengan pso

Tabel 5 adalah *confusion matrix* untuk algoritma *Naive Bayes* dengan pso. Diketahui 308 data diklasifikasi "Yes" diprediksi sesuai dengan data sebenarnya, lalu 13 data diprediksi "No" tetapi ternyata "Yes". Kemudian 13 data diklasifikasi *"No"* diprediksi sesuai, dan 71 data diprediksi "Yes" ternyata "No".

Tabel 5. Model *Confusion Matrix* untuk Metode *Naive Bayes* dengan pso

accuracy: 79.28% +/- 3.89% (mikro: 79.26%)

	true Yes	true No	class precision
pred. Yes	308	71	81.27%
pred. No	13	13	50.00%
class recall	95.95%	15.48%	

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

2. Kurva ROC

a. Algoritma C4.5

kurva ROC untuk algoritma C4.5 seperti ditunjukkan oleh gambar 2 di bawah ini.

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

Gambar 2 Kurva ROC algoritma C4.5

Kurva ROC pada gambar 2 mengekspresikan *confusion matrix*. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.

b. *Naive Bayes*

.Kurva ROC untuk algoritma *Naive Bayes* seperti ditunjukkaon oleh gambar 3 di bawah ini.

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2020)

Gambar 3 Kurva ROC *Naïve Bayes*

c. Algoritma C4.5 dengan pso

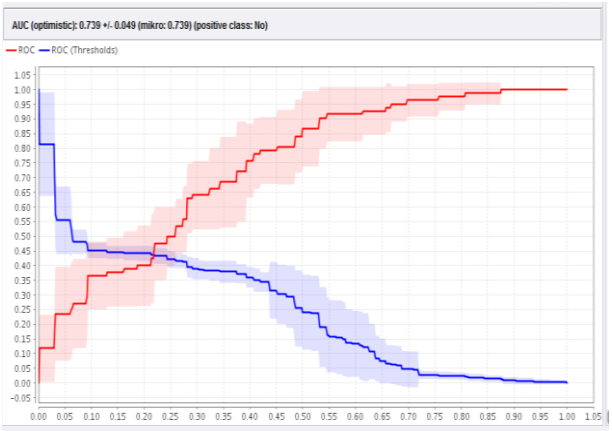
Kurva ROC untuk algoritma C4.5 dengan pso seperti ditunjukkaon oleh gambar 4 di bawah ini.

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

Gambar 4 Kurva ROC Metode C4.5 dengan pso

d. *Naive Bayes* dengan pso

Kurva ROC untuk algoritma *Naive Bayes* dengan pso seperti ditunjukkaon oleh gambar 5 di bawah ini.



Sumber: Hasil Pengolahan Menggunakan RapidMiner (2021)

Gambar 5 Kurva ROC Metode *Naive Bayes* dengan pso.

Pebandingan hasil perhitungan nilai AUC untuk metode C4.5, Naïve Bayes, C4.5 dengan pso, dan Naïve Bayes dengan pso dapat dilihat pada Tabel 7.

3. Analisis Hasil Komparasi

perbandingan nilai *accuracy dan ROC Curve atau AUC* untuk algoritma C4.5, *Naive Bayes*, C4.5 dengan pso, dan *Naive Bayes* dengan pso ditunjukkan oleh tabel 6 di bawah ini.

Tabel 6 Komparasi Nilai A*ccuracy* dan AUC

	C4.5	*NB*	*C4.5+ pso*	*NB+pso*
Accuracy	70.99%	66.14%	**79.51%**	79.28%
AUC	0.950	0.742	**0.950**	0.739

Sumber: Hasil Pengolahan Menggunakan RapidMiner (2020)

Tabel 6 membandingkan *accuracy* dan AUC dari tiap algoritma. Terlihat bahwa nilai *accuracy algoritma C4.5* dengan pso paling tinggi dibandingkan dengan algoritma lainnya begitu pula dengan nilai AUC-nya. Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = klasifikasi sangat baik
b. 0.80-0.90 = klasifikasi baik
c. 0.70-0.80 = klasifikasi cukup
d. 0.60-0.70 = klasifikasi buruk
e. 0.50-0.60 = klasifikasi salah

Berdasarkan pengelompokkan di atas maka dapat disimpukan bahwa model algoritma C4.5 dan C4.5 dengan pso termasuk katagori klasifikasi sangat baik , sedangkan *naïve bayes* dan *naïve bayes* dengan pso termasuk katagori klasifikasi cukup.

KESIMPULAN

Dari hasil penelitian dan pengujian, performa model C45 untuk diagnosa penyakit peradangan hati memberikan tingkat akurasi kebenaran sebesar 70,99% dengan nilai area under the curva (AUC) sebesar 0,950. Performa model Naive Bayes memberikan tingkat akurasi kebenaran sebesar 66,14% dengan nilai area under the curve (AUC) sebesar 0,742. Performa model C4.5 dengan pso memberikan tingkat akurasi kebenaran sebesar 79,51% dengan nilai area under the curve (AUC) sebesar 0,950. Sedangkan Performa model naïve bayes dengan pso memberikan tingkat akurasi kebenaran sebesar 79,28% dengan nilai area under the curve (AUC) sebesar 0,739. Hal ini membuktikan bahwa optimasi particle swarm optimization dapat meningkatkan kinerja metode klasifikasi yang digunakan. Berdasarkan tingkat akurasi dan nilai area under the curve (AUC), maka performa metode C4.5 dengan pso adalah yang paling baik untuk mendiagnosa penyakit peradangan hati.

REFERENSI

Amrin, A., & Pahlevi, O. (2020). Data Mining Model For Designing Diagnostic Applications Inflammatory Liver Disease. *SinkrOn*, *5*(1), 51. https://doi.org/10.33395/sinkron.v5i1.10589

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques*. Springer.

Hannan, A., Manza, R., & Remteke, R. (2010). Generalized Regression Neural Network and Radial Basis Function for Heart Disease diagnosis. *International Journal of Computer Application (0975-8887)*, *7*(13), 7–13.

Karlik, B. (2011). Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers. *Turkey : Journal of Science and Technology*, *1*(1).

Kemenkes. (2017). *Situasi Penyakit Hepatisis B Di Indonesia*. InfoDatin Kemenkes RI.

Kumar, D. S., Sathyadevi, G., & Sivanesh, S. (2011). Decision Support System for Medical Diagnosis Using Data Mining. *India : International Journal of Computer Science Issues*, *8*(1).

Kumar, V., Sharaty, V., & Devi, G. (2012). Hepatitis Prediction Model based on Data Mining Algorithm and Optimal Feature Selection to Improve Predictive Accuracy. *Vellore : International Journal of Computer Applications (0975-8887)*, *51*(19).

Neshat, M., & Yaghoobi, M. (2009). Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System. *Proceeding of the World Congress on Engineering and Computer Science 2009,Vol II, WCECS 2009, ISBN:978-988-18210-2-7*, 1–6.

Prayoga, N. D. (2018). Sistem Diagnosis Penyakit Hati Menggunakan Metode Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(8), 2666–2671.