e- ISSN: 2828–299X

Perbandingan Tingkat Akurasi Algoritma *Support Vector Machines* (SVM) dan C45 dalam Prediksi Penyakit Jantung

Herliyani Hasanah[1], Nurmalitasari[2]
[1]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Duta Bangsa Surakarta
[2]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Duta Bangsa Surakarta
E-mail: *[1]*herliyani_hasanah@udb.ac.id*, [2]*nurmalitasari@udb.ac.id*

Abstrak – *Penyakit jantung merupakan penyakit yang menyumbang angka kematian relatif tinggi. Tingkat kematian manusia yang disebabkan oleh penyakit jantung merupakan masalah yang tersebar luas di dunia. Tujuan utama dari penelitian ini adalah untuk membandingkan dua algoritma prediksi seseorang dengan penyakit jantung menggunakan kumpulan data yang tersedia untuk umum di Repositori Kaggle dengan kumpulan data penyakit jantung. Algoritma yang digunakan dalam penelitian ini adalah algoritma Support Vector Machines (SVM) dan algoritma C.45. Hasil perbandingan menunjukkan bahwa algoritma SVM merupakan yang tepat dan akurat digunakan untuk memprediksi orang dengan hati penyakit dengan nilai accuracy sebesar 87%.*

Kata Kunci *— jantung, C45, Support Vector Machines (SVM)*

1. PENDAHULUAN

Jantung merupakan organ penting dari tubuh manusia. Jantung memompa darah ke setiap bagian dari anatomi tubuh manusia. Jika gagal berfungsi dengan benar, maka otak dan berbagai organ lainnya akan berhenti bekerja, dan dalam waktu singkat menit, orang tersebut akan mati. Perubahan gaya hidup, terkait pekerjaan stres dan kebiasaan makan yang buruk berkontribusi pada peningkatan angka beberapa penyakit yang berhubungan dengan jantung.

Menurut data WHO, penyakit kardiovaskular adalah penyebab utama kematian secara global, mengambil sekitar 17,9 juta jiwa setiap tahun [1]. Kardiovaskular adalah sekelompok gangguan jantung dan pembuluh darah dan termasuk penyakit jantung koroner, penyakit serebrovaskular, penyakit jantung rematik dan kondisi lainnya. Dengan demikian, prediksi penyakit terkait jantung yang layak dan akurat adalah sangat penting. Menurut VV. Ramalingam, algoritma dan teknik pembelajaran mesin telah diterapkan ke berbagai kumpulan data medis untuk mengotomatisasi analisis data yang besar dan kompleks [2]. Teknik pembelajaran mesin banyak digunakan dalam penelitian untuk membantu industri kesehatan dalam diagnosis penyakit salah satunya penyakit jantung.

Penelitian yang sudah dilakukan oleh Alham, yaitu penggunaan algoritma C4.5 untuk diagnosis penyakit jantung koroner dapat diimplementasikan dengan baik dengan pengujian perbandingan data latih dan data uji sebesar 70:30 dan menghasilkan akurasi sebesar 94,4% [3]. Penelitian dilakukan oleh Nugroho, untuk mengetahui Intelegensi (IQ) pasien setelah koma menggunakan metode *Naïve Bayes* dan Metode C45 telah diimplementasikan dengan hasil bahwa metode C45 mempunyai tingkat akurasi lebih tinggi yaitu sebesar 63% [4].

Permana menggunakan algoritma *Support Vector Machine* untuk prediksi penyakit jantung dengan nilai metrik terbaik didapatkan jika menggunakan kernel linear [5]. Nilai metrik akurasi sama dengan 90.11%, presisi 90.38% dan *recall* 92.15% dengan kernel linear. Berdasarkan latar belakang di atas, serta penelitian – penelitian sebelumnya maka dapat dirumuskan masalah, yaitu bagaimana perbandingan performansi model klasifikasi pasien penyakit jantung dengan algoritma *decision tree* C45 dan *Support Vector Machine (SVM)*.

2. LANDASAN TEORI

2.1 Algoritma *Decision Tree* C45

Algoritma C4.5 adalah algoritma klasifikasi data dengan teknik pohon keputusan yang memiliki kelebihan-kelebihan. Kelebihan ini misalnya dapat mengolah data numerik dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan - aturan yang mudah diintrepetasikan dan tercepat di antara algoritma algoritma yang lain [6] untuk membangun pohon keputusan pertama yansg akan dilakukan yaitu memilih atribut sebagai akar. Kemudian membuat cabang untuk

e- ISSN: 2828–299X

setiap nilai didalam akar tersebut. Langkah selanjutnya dengan membagi kasus dalam cabang. Kemudian ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama[6][7]. Perhitungan nilai entropy yang dapat dilihat pada persamaan di bawah ini :

$$Entropy(S) = \sum_{i=1}^{n} -p_i * log_2\, p \quad (1)$$

Keterangan:

S = Himpunan Kasus

N = Jumlah Partisi S

P_i = Porsi P_i terhadap S

Gain adalah salah satu atribute selection measure yang digunakan untuk memilih test atribut tiap node pada tree. Atribut dengan information gain tertinggi dipilih sebagai test atribut dari suatu node [6][7].

$$Gain(S, A) = Entropy\ (S) - \sum_{t=i}^{n} \frac{|S_i|}{|S|} * Entropy\ (s) \quad (2)$$

Keterangan :

S = Himpunan Kasus

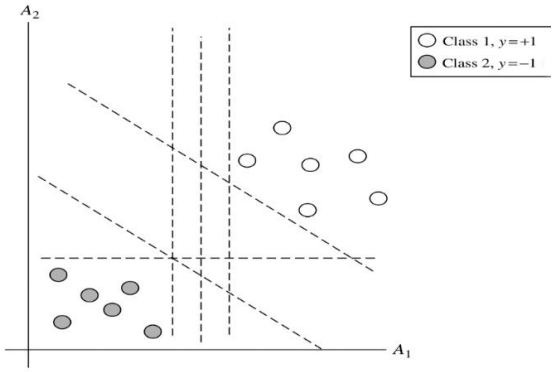
A = Atribut

n = Jumlah partisi atribut A

$|S_i|$ = Jumlah kasus pada partisi ke - i

2.2 Algoritma *Support Vector Machine (SVM)*

Menurut Jiawei, *Support Vector Machine* adalah metode klasifikasi untuk data linear dan nonlinear dan termasuk ke kategori pembelajaran mesin dengan pengawasan [8]. Metode ini menggunakan pemetaan nonlinier untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi. Dalam dimensi baru ini akan dicari optimasi linier yang memisahkan dua kelas target dengan hyperplane. Hyperplane adalah batas keputusan yang memisahkan tipe dari satu kelas dengan kelas yang lain. SVM menemukan hyperplane menggunakan vektor pendukung dan margin.



Gambar 1. Hyperplane di SVM yang Terbagi Atas Dua Kelas Target

3. METODE PENELITIAN

3.1 Sumber Data

Penelitian ini menggunakan dataset yang diperoleh dari website *Kaggle* yaitu *Heart Failure Prediction Dataset*.

3.2 Variabel Data

Dataset ini berisi data yang dikumpulkan dari lima data set jantung yang digabungkan, dengan 11 atribut, dimana 10 atribut merupakan atribut biasa sedangkan 1 atribut sebagai *class* dan memiliki 918 *instance*. Lima dataset tersebut adalah *Cleveland* dengan jumlah 303 observasi, Hungarian dengan jumlah 294 observasi,

e- ISSN: 2828–299X

Switzerland dengan jumlah 123 observasi, *Long Beach* dengan jumlah 200 observasi, dan *Stalog (Heart)* dengan jumlah 270 observasi.

3.3 Tahapan Penelitian

Gambar 2. Tahapan Penelitian

Berikut ini langkah-langkah yang dilakukan dalam penelitian:

1) Identifikasi Masalah

Identifikasi masalah yang ingin dicari solusinya, dalam penelitian ini adalah bagaimana kita memprediksi apakah seorang pasien terkena penyakit jantung atau tidak berdasarkan kondisi medis mereka.

2) Pengumpulan dan PraProses Data.

Tahap ini menentukan sumber data yang berupa data sekunder diperoleh dari website *Kaggle* yaitu *Heart Failure Prediction Dataset*. Jumlah data dari data set tersebut adalah 1190 *instance*. *Instance* pada dataset ini memiliki nilai *duplicate* sebanyak 272. Sehingga final dataset yaitu 918 *instance* yang siap untuk diproses.

3) Model Mining

Penelitian ini menggunakan model terbaik menggunakan algoritma *Support Vector Machines* (SVM) dan C45. Proses perancangan, pelatihan dan pengujian data diimplementasikan dengan bahasa python dan pustaka scikit.

4) Learner

Pada tahap ini akan dilakukan tahap pelatihan dari dataset.

5) Evaluasi

Tahap evaluasi prediksi dengan AUC, CA, F1, *Precission, Recall, Confussion Matrix* dan ROC analysis.

4. HASIL DAN PEMBAHASAN

4.1 Olah Data

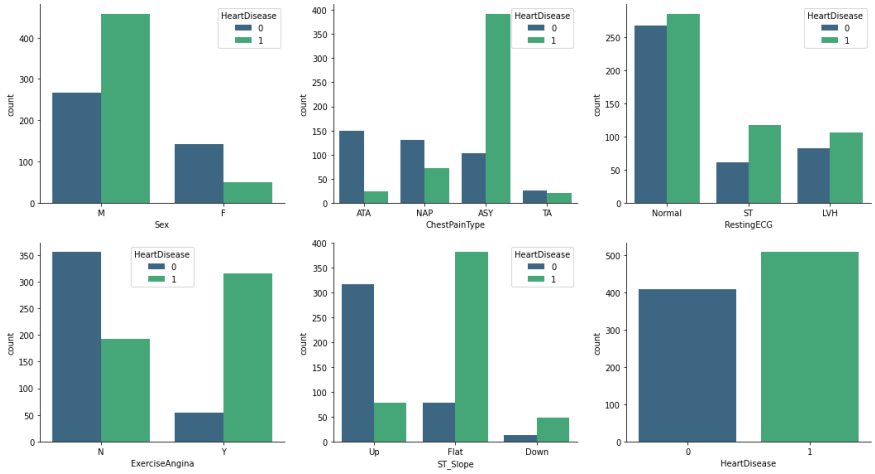
Database yang digunakan adalah basisdata pasien jantung dan kondisi klinisnya dari website *Kaggle* yaitu *Heart Failure Prediction Dataset*. Gejala penyakit kardiovaskular terdiri dari berbagai kondisi yang mempengaruhi kerja jantung dan vena darah dan cara darah dipompa dan diedarkan melalui tubuh [9]. Basisdata ini mempunyai 918 *instance* hasil praproses data dengan 12 atribut, distribusi data disajikan dalam bentuk grafik batang pada gambar 3. Sebanyak 11 atribut merupakan kondisi klinis pasien yang digunakan sebagai atribut prediksi. Sedangkan kelas target adalah atribut ke 12, atribut ini memiliki dua nilai. Nilai 1 untuk pasien terkena penyakit jantung dan 0 untuk pasien bukan penyakit jantung. Berikut ini rincian atribut dalam bentuk tabulasi:

Tabel 1. Atribut Prediksi dan Kelas Target

No	Atribut	Tipe Data	Keterangan
X_1	*Age*	numerik	Umur pasien
X_2	*Sex*	text	Jenis kelamin
X_3	*ChestPainType*	text	Jenis sakit dada. TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic

e- ISSN: 2828–299X

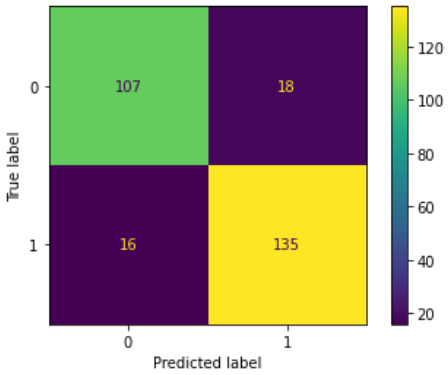
X_4	*RestingBP*	numerik	Tekanan darah
X_5	*Cholesterol*	numerik	Jumlah kolestrol
X_6	*FastingBS*	numerik	Tekanan gula darah
X_7	*RestingECG*	text	Hasil tes tekanan electrodiogram
X_8	*MaxHR*	numerik	Detak jantung maksimum
X_9	*ExerciseAngina*	text	Nyeri dada ketika olahraga
X_{10}	*Oldpeak*	numerik	Oldpeak
X_{11}	*ST_Slope*	text	Kemiringan detak jantung setelah olahraga
Y	*HeartDisease*	numerik	1: *positif heart disease* 0: *negatif heart disease*



Gambar 3. Grafik Distribusi Data

4.2 Evaluasi

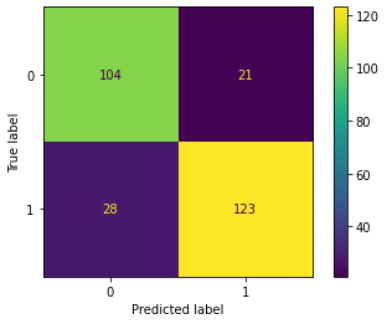
Tahap evaluasi prediksi dengan AUC, CA, F1, *Precission*, *Recall*, *Confussion Matrix* dan ROC *analysis*. CA dapat berfungsi untuk akurasi dari dataset yang dipilih. *Precission* adalah akurasi data yang memungkinkan dua kejadian yaitu 1 dan 0. *Recall* berfungsi untuk mengukur rasio. F1 yaitu perbandingan anatara *recall* dan presisi. AUC digunakan untuk mewakili probabilitas.



Gambar 4. *Confusion Matrix* SVM

Algoritma *Support Vector Machine* (SVM) pada gambar 4 terdapat 107 + 135 prediksi yang benar dan 18 + 16 prediksi yang salah dari total 276 data testing. Sehingga tingkat akurasi algoritma dapat dihitung 242/ 276 * 100% = 88%.

e- ISSN: 2828–299X



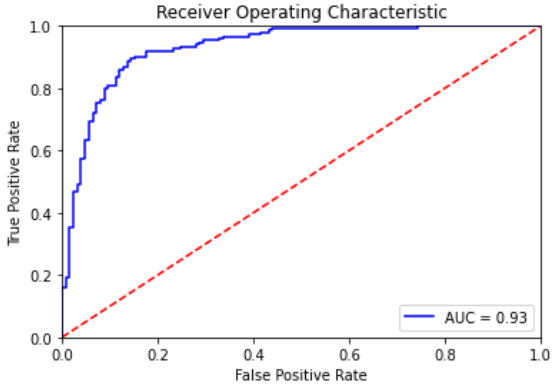
Gambar 5. *Confusion Matrix* C45

Algoritma C45 pada gambar 5 terdapat 104 + 123 prediksi yang benar dan 21 + 28 prediksi yang salah dari total 276 data testing. Sehingga tingkat akurasi algoritma dapat dihitung 227/ 276 * 100% = 82%.

Tabel 2. Kinerja Algoritma

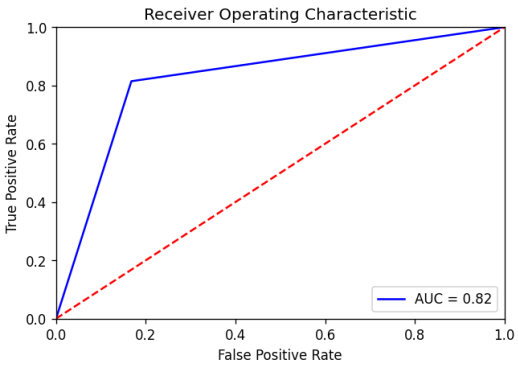
Model	AUC	CA	F1	*Precision*	*Recall*
SVM	0.93	0.88	0.86	0.86	0.87
C45	0.82	0.82	0.81	0.83	0.79

Pada tabel 2 dapat diketahui bahwa algoritma *Support Vector Machine* (SVM) memiliki tingkat akurasi tertinggi untuk data klinis jantung. Besarnya nilai AUC 0.93, CA 0.88, F1 0.086, *Precision* 0.86 dan *Recall* 0.87.



Gambar 6. ROC SVM

Pada gambar 6 ROC SVM terlihat bahwa nilai *sensitivity* (*TP Rate*) dan *specificity (FP Rate)* tidak lebih dari 1. Besarnya nilai AUC adalah 0.93 dengan kualitas *classifier* adalah *excellent*, yang berarti bahwa algoritma SVM memiliki klasifikasi yang bagus dalam menyelesaikan masalah klasifikasi penyakit Kardiovaskular.



Gambar 7. ROC C45

e- ISSN: 2828–299X

Pada gambar 7 ROC C45 terlihat bahwa nilai *sensitivity (TP Rate)* dan *specificity (FP Rate)* tidak lebih dari 1. Besarnya nilai AUC adalah 0.82 dengan kualitas *classifier adalah good*.

Dari analisis kurva ROC Gambar 6 dan Gambar 7 dapat dilihat kinerja algoritma klasifikasi. Semakin dekat kurva mengikuti batas kiri dan kemudian batas atas ruang ROC, semakin akurat classifier tersebut. Dari ROC Analisis tersebut dapat diketahui bahwa model SVM memiliki keakuratan dari *classifier* lebih baik dibandingkan dari model C45. Hal ini terlihat pada setiap class dapat dilihat bahwa kurva dari SVM rata - rata mendekati sumbu axis atau sumbu Y yang menandai bahwa SVM memiliki keakuratan *classifier* yang optimal.

5. SIMPULAN

1. Algoritma *Support Vector Machine* (SVM) memiliki nilai akurasi sebesar 88% dalam menyelesaikan masalah klasifikasi penyakit Kardiovaskular
2. Algoritma C45 memiliki nilai akurasi sebesar 82% dalam menyelesaikan masalah klasifikasi penyakit Kardiovaskular
3. Hasil ROC Analisis dapat diketahui bahwa model SVM memiliki keakuratan dari *classifier* lebih baik dibandingkan dari model C45 dengan nilai AUC adalah 0.93 dengan kualitas *classifier* adalah *excellent.*

6. SARAN

Penelitian selanjutnya diharapkan dapat ditambahkan algoritma optimasi lain untuk meningkatkan tingkat akurasi. Penerapan algoritma baik berupa penggabungan ataupun optimasi algoritma diharapkan dapat meningkatkan nilai performa algoritma tersebut.

DAFTAR PUSTAKA

[1] S. Mendis, P. Puska, and B. Norrving, “Global atlas on cardiovascular disease prevention and control,” *World Heal. Organ.*, pp. 2–14, 2011.

[2] S. Guruprasad, V. L. Mathias, and W. Dcunha, “Heart Disease Prediction Using Machine Learning Techniques,” *2021 5th Int. Conf. Electr. Electron. Commun. Comput. Technol. Optim. Tech. ICEECCOT 2021 - Proc.*, vol. 7, pp. 762–766, 2021, doi: 10.1109/ICEECCOT52851.2021.9707966.

[3] S. R. J. I. Alham, “Sistem Diagnosis Penyakit Jantung Koroner Dengan Menggunakan Algoritma C4.5 Berbasis Website (Studi Kasus: RSUD Dr. Soedarso Pontianak),” *Petir*, vol. 14, no. 2, pp. 214–222, 2021, doi: 10.33322/petir.v14i2.1338.

[4] F. A. Nugroho, “Implementasi Data Mining Pada Pasien Setelah Koma dengan Menggunakan Metode Algoritma C 4 . 5 dan Naivy Bayes untuk Mengetahui Intelegensi (IQ) Pasien,” *Semin. Nas. Din. Inform. 2020*, pp. 26–28, 2020.

[5] D. S. Permana and A. Silvanie, “Prediksi Penyakit Jantung Menggunakan Support Vector Machinedan Pythonpada Basis Data Pasiendi Cleveland,” *J. Nas. Inform.*, vol. 2, no. 1, pp. 30–34, 2021.

[6] A. Sonita and R. Kundari, “Aplikasi Seleksi Calon Pendonor,” *J. Pseudocode*, vol. VI, no. September, pp. 96–103, 2019.

[7] L. Hermawanti, “Penerapan Algoritma Klasifikasi C4.5 untuk Diagnosis Penyakit Kanker Payudara,” *J. Tek. Unisfat*, vol. 7, no. 1, pp. 57–64, 2012.

[8] J. Han, M. Kamber, and M. Kaufmann, “Data Mining: Concepts and Techniques (2nd edition) Classification and Prediction,” 2006, [Online]. Available: www.rulequest.com.

[9] P. L. Latthe, R. Champaneria, and K. S. Khan, “Women ’ s health Women ’ s health Dysmenorrhoea,” *Clin. Evid. (Online).*, vol. 9, no. January 2010, pp. 1–59, 2011, doi: 10.1186/1472-6874-4-S1-S15.