Aisyah Journal of Informatics and Electrical Engineering
Universitas Aisyah Pringsewu

Journal Homepage
http://jti.aisyahuniversity.ac.id/index.php/AJIEE

MACHINE LEARNING UNTUK KLASIFIKASI PENYAKIT JANTUNG

Ratnasari[1*], Ahmad Jurnaidi Wahidin[2], Agustinus Eko Setiawan[3], Panji Bintoro[4]

[1,4]Rekayasa Perangkat Lunak, Universitas Aisyah Pringsewu
[3]Teknik Informatika, Universitas Aisyah Pringsewu
[2]Teknologi Informasi, Universitas Bina Sarana Informatika

Email: ratnasari@aisyahuniversity.ac.id*, Ahmad.ajn@bsi.ac.id, agustinus@aisyahuniversity.ac.id, panjibintoro09@aisyahuniversity.ac.id

ABSTRACT

Heart disease is caused by abnormal conditions of the heart and blood vessels, widely considered a direct threat to human life and health. Correct diagnosis in the early phase is a very challenging task due to the complex dependencies that must be taken into account on various factors. Therefore, it is necessary to develop a medical diagnosis system in such a way that it can assist in making decisions in the diagnostic process. This research aims to find a machine learning algorithm that has the highest accuracy for predicting whether someone has heart disease or not based on a medical database. Our research compares six machine learning classification methods namely Naïve Bayes, kNN, Random Forest, Logistic Regression, SVM, Decision Tree and AdaBoost with the Cleveland Clinic Foundation dataset available in the "UCI Machine Learning Repository". The results of this research show that the Naive Bayes algorithm has the highest accuracy, namely 84.67%, then Logistic Regression is in second place with an accuracy of 84.30%, then Random Forest 81.70%, SVM 81%, Tree 74%, kNN 73%, AdaBoost 71.30%.

Keywords: *Machine Learning, Classification Models, Heart Disease, Cleveland dataset*

ABSTRAK

Penyakit jantung disebabkan oleh kondisi abnormal jantung dan pembuluh darah, secara luas dianggap sebagai ancaman langsung terhadap kehidupan dan kesehatan manusia. Diagnosa yang tepat pada fase awal merupakan tugas yang sangat menantang karena adanya ketergantungan yang kompleks yang harus dipertimbangkan pada berbagai faktor. Oleh karena itu dibutuhkan pengembangan sistem diagnosis medis sedemikian rupa sehingga dapat membantu dalam mengambil keputusan pada proses diagnostik. Penelitian ini bertujuan untuk mencari algoritma mechine learning yang memiliki akurasi yang paling tinggi untuk menprediksi apakah seseorang mengidap penyakit jantung atau tidak berdasarkan database medis. Penelitian kami membandingkan enam metode klasifikasi mechine learning yaitu Naïve Bayes, kNN, Random Forest, Logistic Regression, SVM, Decision Tree dan AdaBoost dengan dataset Cleveland Clinic Foundation yang tersedia di “UCI Machine Learning Repository”. Hasil dari penelitian ini menunjukan bahwa algoritma Naive Bayes memiliki akurasi paling tinggi yaitu sebesar 84.67%, lalu Logistic Regression diurutan kedua dengan akurasi 84.30%, Kemudian Random Forest 81.70%, SVM 81%, Tree 74%, kNN 73%, AdaBoost 71.30%.

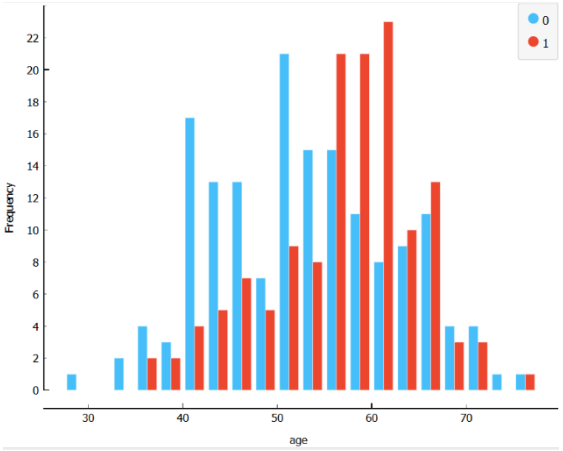
Kata Kunci: *Mechine Learning*, Model Klasifikasi, Penyakit Jantung, Cleveland dataset

I. PENDAHULUAN

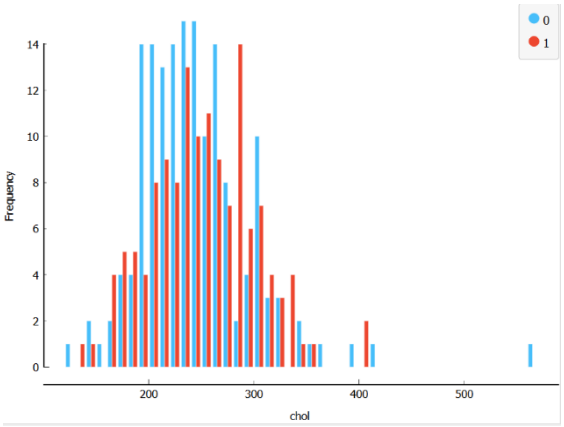
Penyakit jantung tetap menjadi isu kesehatan yang meresahkan dan kritis secara global. Penyakit ini disebabkan oleh kondisi abnormal jantung dan pembuluh darah, secara luas dianggap sebagai ancaman langsung terhadap kehidupan dan kesehatan manusia. Dari semua penyakit yang disebabkan adanya gangguan pada jantung, penyakit arteri koroner adalah penyumbang penyebab kematian terbanyak [1]. Heron menyatakan bahwa penyakit jantung merupakan salah satu penyakit yang memberikan efek tidak dapat dipulihkan pada banyak orang dewasa dan lansia, di mana komplikasi fatal sangat mungkin terjadi [2].

Analisis statistik telah mengidentifikasi beberapa faktor risiko terkait penyakit jantung, seperti usia, tekanan darah, kebiasaan merokok, kolesterol, diabetes, hipertensi, riwayat keluarga penyakit jantung, obesitas, dan kurangnya aktivitas fisik [3]. Diagnosa yang tepat pada fase awal merupakan tugas yang sangat menantang karena adanya ketergantungan yang kompleks yang harus dipertimbangkan pada berbagai faktor. Oleh karena itu dibutuhkan pengembangan sistem diagnosis medis sedemikian rupa sehingga dapat membantu dalam mengambil keputusan pada proses diagnostik.

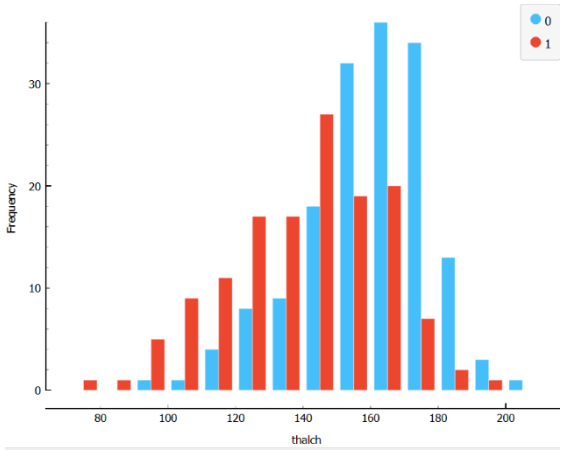
Untuk menjawab permasalahan diatas, Teknik machine learning dapat dimanfaantkan untuk membantu praktisi medis dalam mendeteksi penyakit ini. Di antara teknik yang ada, metode supervised learning adalah yang paling populer dalam diagnosis penyakit jantung [3]. Didalam penelitian ini kami membandingkan enam metode klasifikasi mechine learning yaitu Naïve Bayes, kNN, Random Forest, Logistic Regression, SVM, Decision Tree dan AdaBoost. Penelitian ini bertujuan untuk mencari algoritma mechine learning yang memiliki akurasi yang paling tinggi untuk menprediksi apakah seseorang mengidap penyakit jantung atau tidak berdasarkan database medis.



Gambar 1. Histogram Variabel Usia



Gambar 2. Histogram Variabel Kolestrol



Gambar 3. Histogram Variabel denyut jantung maksimum

Histogtram diatas adalah gambaran data medis mewakili beberapa atribut yang digunakan diantaranya adalah umur, kolestrol, dan denyut jantung maksimum terhadap diagnosis penyakit jantung, dimana simbol 1 menggambarkan pasien didiagnosis mengidap penyakit jantung dan 0 untuk pasien yang didiagnosis tidak mengidap penyakit jantung.

II. TINJAUAN PUSTAKA

Dalam penelitian ini kami menggunakan tujuh metode mechine learning diantaranya:

1. Naïve Bayes
   algoritme naïve bayes merupakan metode yang digunakan dalam machine learning untuk menangani masalah klasifikasi berdasarkan pada probabilitas. Menurut teorema Bayesian :

   P(A|B)=P(A)*P(B/A)/P(B)
   Dimana P(B|A) = P(A∩B)/P(A)

   Pengklasifikasi Bayesian menghitung probabilitas bersyarat dari suatu instance yang termasuk dalam setiap kelas, berdasarkan rumus di atas, dan berdasarkan pada data probabilitas bersyarat tersebut, instance tersebut diklasifikasikan sebagai kelas dengan probabilitas bersyarat tertinggi.

2. K-Nearest Neighbors (KNN)
   K-Nearest Neighbors (KNN) merupakan algoritma klasifikasi yang menguji kemungkinan dari suatu titik data milik suatu kelompok menurut jarak terdekatnya titik[8]. Penelitian ini memilih 1 sampai 20 sebagai jumlah tetangga.

3. Random Forest
   Random Forest adalah algoritma yang terdiri dari pohon Keputusan[8]. Algoritma ini merupakan kombinasi masing-masing tree dari decision tree yang kemudian digabungkan menjadi satu model.

4. Logistic Regression
   Regresi logistik adalah cara paling sederhana untuk menangani masalah klasifikasi[9]. Regresi Logistik adalah model untuk memprediksi hasil biner dengan menggunakan observasi kumpulan data. Penelitian ini memilih model ini karena variabel outputnya adalah hasil biner yang mengambil risiko tinggi atau tidak berisiko terkena penyakit jantung.

5. Support Vector Machine (SVM)
   Algoritma Support Vector Machine (SVM) paling efektif untuk masalah klasifikasi. SVM dapat digunakan untuk klasifikasi linier dan nonlinier tergantungpada fungsi kernel yang berbeda. SVM merupakan alternatif dari metode pembelajaran Bayesian [9]. SVM tergantung pada vektor dukungan yang dipilih berdasarkan jaraknya. Poin-poinnya yaitu dekat dengan batas keputusan dikenal sebagai vektor pendukung. Semakin besar jaraknya poin dari margin, semakin tinggi kepercayaannya. Model ini sepenuhnya tergantung pada vektor dukungan yang dipilih dan metrik jarak. Dalam penelitian ini, Algoritma SVM diterapkan menggunakan parameter default dan diperoleh akurasi 81%.

6. Decision Tree
   Decision Tree merupakan pengklasifikasi berupa pohon yang mempunyai dua jenis node, yaitu: node keputusan (*decision nodes*) dan node daun (*leaf nodes*)[10]. Decision Tree sama seperti pohon biner lainnya dan dapat dengan mudah diikuti untuk mencapai simpul daun.

7. AdaBoost
   Algoritma AdaBoost (Adaptive Boosting), adalah sebuah teknik Boosting yang digunakan sebagai metode ensemble dalam machine learning. Algoritma ini disebut Adaptive Boosting karena bobot diberikan ulang pada setiap instance, dengan bobot yang lebih tinggi diberikan pada instance yang salah diklasifikasikan. Pada penelitian ini algoritma AdaBoost memiliki tinggkat akurasi 71.30%

Model Evaluation

1. Confusion Matrik
   Confusion Matrix adalah pengukuran performa untuk masalah klasifikasi machine learning dimana keluaran dapat berupa dua kelas atau lebih. Confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu True Positif, True Negatif, False Positif, dan False Negatif.

Tabel 1. Confusion Matrik

TP	FN

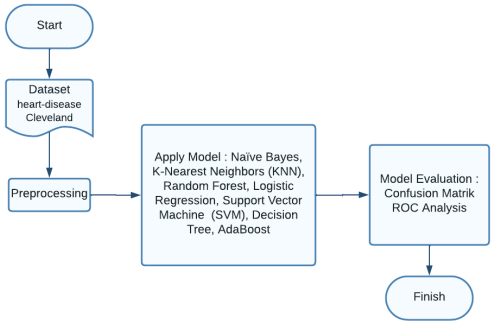
FP	TN

2. ROC Analysis

ROC Analysis adalah metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, terutama dalam konteks biner (dua kelas). Metrik ini fokus pada kemampuan model untuk membedakan antara kelas positif dan negatif dengan memperhatikan trade-off antara tingkat True Positive Rate (TPR) dan tingkat False Positive Rate (FPR).

III. METODOLOGI

Berikut langkah-langkah penelitian :



Gambar 4. Langkah Penelitian

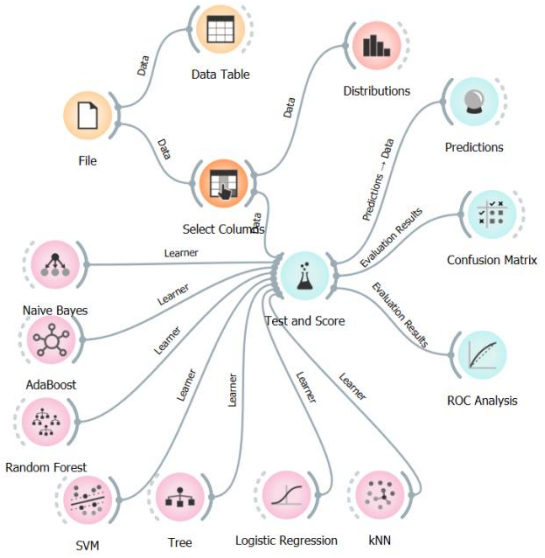
Dataset yang digunakan yaitu “heart-disease directory” tersedia di “UCI Machine Learning Repository”[4]. Dataset heart-disease directory memiliki empat sumber data yaitu Cleveland, Hungary, Switzerland, dan VA Long Beach. Dalam penelitian ini data yang digunakan yang bersumber dari Cleveland Clinic Foundation[5][6][7] dengan alasan kelengkapan data yang tersedia, data yang tersedia memiliki 303 record data dengan 14 atribut, yang terdiri dari 13 variabel bebas dan 1 atribut variable diagnosis yang akan dijadikan sebagai variable target, namun 6 record data diantanya memiliki atribut yang tidak lengkap sehingga hanya 297 record yang akan digunakan. Atribut yang tersedia dikonversikan kedalam kode biner kemudian diolah menggunakan aplikasi Orange Data Mining.

Tabel 1. Detail atribut yang digunakan

Atribut	Keterangan
Age	Umur Pasien
Jenis Kelamin	1 = Laki-laki 0 = Perempuan
cp	Jenis Nyeri dada 1 = typical angina 2 = atypical angina 3 = non -anginal pain 4 = asymptomatic
Trestbps	Tekanan Darah (mm Hg)
Chol	Kolestrol (mg/dl)
Fbs	Gula darah puasa
Restecg	Resting Relectrocardiographic 0 = normal 1 = ST-T abnormality 2 = showing probable or define left ventricular hypertrophy by Estes’criteria
thalch	Denyut jantung maksimum
Exang	Olahraga 1 = ya 0 = tidak
oldpeak	Depresi
Slope	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Fluoroskopi
Thal	3 = normal 6 = cacat tetap 7 = cacat reversibel
Diagnosis (num)	Diagnosa 0 = sehat 1 = Mempunyai penyakit jantung

IV. HASIL DAN PEMBAHASAN

Setelah data melewati tahapan preprocessing, maka kami berekperimen menggunakan aplikasi orange untuk menguji algoritma mechine learning yang memiliki akurasi paling tinggi seperti gambar berikut :



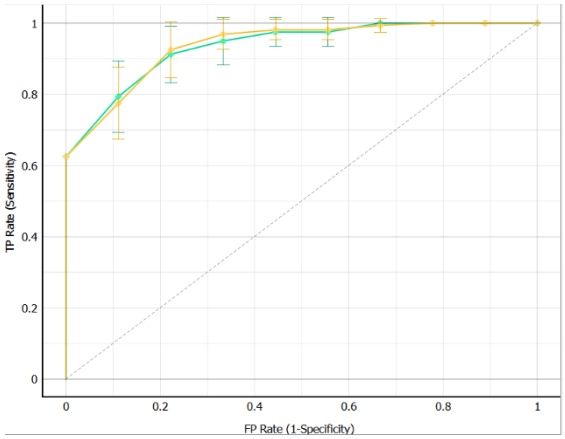
Gambar 5. Penerapan dan Evaluasi Model

Dari percobaan diatas penelitian kami mendapatkan hasil berikut :

Tabel 2. Hasil Comparasi Alogoritma mechine learning

Model	Akurasi (%)	Confusion Matrik
Naive Bayes	84.67%	0: 141, 19, 160 / 1: 27, 113, 140 / Σ: 168, 132, 300
Logistic Regression	84.30%	0: 140, 20, 160 / 1: 27, 113, 140 / Σ: 167, 133, 300
Random Forest	81.70%	0: 134, 26, 160 / 1: 29, 111, 140 / Σ: 163, 137, 300
SVM	81%	0: 136, 24, 160 / 1: 33, 107, 140 / Σ: 169, 131, 300
Tree	74%	0: 125, 35, 160 / 1: 43, 97, 140 / Σ: 168, 132, 300
kNN	73%	0: 119, 41, 160 / 1: 40, 100, 140 / Σ: 159, 141, 300
AdaBoost	71.30%	0: 109, 51, 160 / 1: 35, 105, 140 / Σ: 144, 156, 300

Dari tabel diatas terlihat bahwa hasil akuasi paling tinggi adalah algoritma Naive Bayes sebesar 84.67% lalu Logistic Regression diurutan kedua dengan akurasi 84.30%.
Characteristic (ROC) adalah plot grafis yang menggambarkan kinerja model klasifikasi pada nilai ambang batas yang bervariasi. Kurva ROC merupakan plot dari tingkat positif sebenarnya terhadap tingkat positif palsu pada setiap pengaturan ambang batas [65]. ROC pada model Naive Bayes yaitu 0.502 (warna biru), dan Logistic Regression sebesar 0.48 (warna orange), seperti yang ditunjukkan pada Gambar 3 berikut.



Gambar 6. Kurva ROC algoritma Naive Bayes dan Logistic Regression

V. PENUTUP

Penelitian ini membandingkan tujuh metode Mechine Learning yaitu Naïve Bayes, kNN, Random Forest, Logistic Regression, SVM, Decision Tree dan AdaBoost untuk memprediksi apakah seseorang mengidap penyakit jantung atau tidak berdasarkan database medis. Hasil dari penelitian ini menunjukan bahwa algoritma Naive Bayes memiliki akurasi paling tinggi yaitu sebesar 84.67%, lalu Logistic Regression diurutan kedua dengan akurasi 84.30%, Kemudian Random Forest 81.70%, SVM 81%, Tree 74%, kNN 73%, AdaBoost 71.30%.

DAFTAR PUSTAKA

[1] Shi, Z. Tao, P. Wei, and J. Zhao, "Epidemiological aspects of heart diseases (Review)," Experimental and Therapeutic Medicine, vol. 12, no. 3. pp. 1645–1650, 2016.

[2] Heron, M. (2012) Deaths: Leading Causes for 2008. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 60, 1-94.

[3] D. Chaki, A. Das and M. I. Zaber (2015) A comparison of three discrete methods for classification of heart disease data, Bangladesh J. Sci. Ind. Res,2015, 293-296.

[4] K. Bache and M. Lichman, "UCI Machine Learning Repository," University of California Irvine School of Information, vol. 2008, no. 14/8. p. 0, 2013

[5] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, 1989.

[6] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognit., vol. 30, no. 7, pp. 1145–1159, 1997.

[7] K. U. Rani, "Analysis of heart diseases dataset using neural network approach," Int. J. Data Min. Knowl. Manag. Process, vol. 1, no. 5, pp. 1–8, 2011.

[8] Huating Sun, Jianan Pan, "Heart Disease Prediction Using Machine Learning Algorithms with Self-Measurable Physical Condition Indicators", Journal of Data Analysis and Information Processing, 2023, 11, 1-10.

[9] Akansh Gupta (eds.), "Heart Disease Prediction Using Classification (Naive Bayes)" Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019), Lecture Notes in Networks and Systems 121, https://doi.org/10.1007/978-981-15-3369-3_42

[10] Chavda, P., Bhavsar, H., Pithadia, Y., Kotecha, R.: Early Detection of Cardiac Disease Using Machine Learning. Available at SSRN 3370813 (2019).

[11] Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.P.P. (2013) Computational Intelligence for Heart Disease Diagnosis: A Medical Knowledge Driven Approach. Expert Systems with Applications, 40, 96-104. https://doi.org/10.1016/j.eswa.2012.07.032

[12] Xing, Y.W., Wang, J., Zhao, Z.H. and Gao, Y.H. (2007) Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease. Convergence Information Technology, Gwangju, 21-23 November 2007, 868-872. https://doi.org/10.1109/ICCIT.2007.204.

[13] Nahar, J., Imam, T., Tickle, K.S. and Chen, Y.-P.P. (2013) Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females. Expert Systems with Applications, 40, 1086-1093. https://doi.org/10.1016/j.eswa.2012.08.028.