JURNAL TEKNIK MESIN, ELEKTRO DAN ILMU KOMPUTER

Halaman Jurnal: https://journal.amikveteran.ac.id/index.php/teknik
Halaman UTAMA Jurnal : https://journal.amikveteran.ac.id/index.php

COMPARISON OF DATA MINING CLASSIFICATION TECHNIQUES FOR HEART DISEASE PREDICTION SYSTEM

Rezty Amalia Aras [a], Noor Akhmad Setiawan [b]
[a] Department of Electrical Engineering and Information Technology, Faculty of Engineering, rezty.mti15@mail.ugm.ac.id, Universitas Gadjah Mada, Indonesia
[b] Department of Electrical Engineering and Information Technology, Faculty of Engineering, noorwewe@ugm.ac.id, Universitas Gadjah Mada, Indonesia

ABSTRAK

DM is the process of analyzing data from different perspectives and gathering knowledge that can be used for different applications. Classification as one of the data mining techniques used to predict group membership. For example, the healthcare industry. DM provides a set of techniques for discovering hidden patterns from data. In this paper, we examine the heart disease dataset in order to obtain information or patterns that can be useful for making a decision. The test in this paper is a prediction of heart disease using three classification methods, namely OneR, decision tree and naive bayes. The results of this experiment show predictions from each experiment with different levels of prediction accuracy in each method used with 91.48% accuracy for the decision tree, 85.18% for naive Bayes and 76.3% for OneR.

Keywords: Data mining; classification techniques; heart disease; OneR; Decision Trees; naive bayes.

Abstrak

DM merupakan proses menganalisis data dari perspektif yang berbeda dan mengumpulkan pengetahuan yang dapat digunakan untuk aplikasi yang berbeda. Klasifikasi sebagai salah satu teknik data mining yang digunakan untuk prediksi keanggotaan kelompok. misalnya industri kesehatan. DM memberikan seperangkat teknik untuk menemukan pola-pola tersembunyi dari data. Pada makalah ini dilakukan pegujian terhadap dataset penyakit jantung guna mendapatkan informasi atau pola yang dapat berguna untuk penentuan pengambila sebuah keputusan. Pengujian pada makalah ini merupakan prediksi terhadap penyakit jantung dengan menggunakan tiga metode klasifikasi, yakni *OneR, decision tree* dan *naive bayes.* Hasil dari percobaan ini menunjukkan prediksi dari setiap percobaan dengan tingkat akurasi prediksi yang berbeda-beda disetiap metode yang digunakan dengan masing-masing akurasi 91,48% untuk *decision tree* 85,18% untuk *naive bayes* dan 76,3% untuk *OneR.*

Kata Kunci: Data mining; teknik klasifikasi; penyakit jantung; OneR; Decision Tree; naive bayes.

1. PENDAHULUAN

Menerapkan teknik data mining dalam pegaturan pengajaran disebut sebagai *Eduational Data Mining* (EDM); dan merupakan bidang yang mengeksploitasi algoritma statistik, *machine learning*, dan data mining (DM) atas varian mengatur data. Tujuan utamanya adalah untuk menganalisis jenis data untuk menyelesaikan masalah penelitian skolastik [1]. DM merupakan ekstraksi yang tidak mudah dari implisit, yang sebelumnya tidak diketahui dan informasi yang berguna mengenai data [2]. Singkatnya, DM merupakan proses menganalisis data dari perspektif yang berbeda dan mengumpulkan pengetahuan yang dapat digunakan untuk aplikasi yang berbeda. Klasifikasi sebagai salah satu teknik data mining yang digunakan untuk prediksi keanggotaan kelompok. misalnya industri kesehatan. Saat ini industri kesehatan menghasilkan sejumlah besar data tentang pasien, diagnosis penyakit dll. DM memberikan seperangkat teknik untuk menemukan pola-pola tersembunyi dari data. Tantangan utama yang dihadapi industri kesehatan adalah kualitas layanan. Kualitas layanan menyiratkan mendiagnosis penyakit dengan benar dan memberikan pengobatan yang efektif untuk pasien [3].

Received Juni 2, 2022; Revised Juni 16, 2022; Accepted Juli 13, 2022

Banyak algoritma baru serta teknik data mining yang telah digunakan untuk memprediksi penyakit jantung. Diagnosis menggunakan teknik *machine learning* merupakan salah satu teknik yang ada yang memiliki pengetahuan diagnostik transparan yang memberikan hasil yang lebih akurat [4].

Menurut *World Health Organization* (WHO) pada pencegahan dan pengendalian penyakit *cardiovascular* menyatakan bahwa *cardiovascular disease* (CVDs) adalah penyebab utama kematian dan kecacatan di dunia. Meskipun sebagian besar CVDs dicegah, mereka terus meningkat terutama karena langkah-langkah pencegahan yang tidak memadai. Lebih dari 17.3 juta diperkirakan 17.3 juta orang meninggal karena CVDs pada tahun 2008, lebih dari 80% kematian CVDs berlangsung di negara-negara berpenghasilan rendah dan menengah, 23.6 juta orang akan meninggal karena CVDs [5]. Jika kondisi didiagnosis lebih awal dan jika mereka mengikuti rencana pengobatan mereka secara teratur maka orang-orang yang mengalami gagal jantung dapat hidup lebih lama. Jantung merupakan orga penting dari bagain tubuh manusia. Hal ini tidak lebih seperti pompa, yang memompa darah keseluruh tubuh. Jika sirkulasi darah dalam tubuh tidak efisien, organ seperti otak menderita dan jika jantung berhenti bekerja sama sekali, kematian terjadi dalam beberapa menit. Beberapa faktor resiko penyakit jantung adalah obesitas, kadar kolesterol darah yang tidak normal, darah tinggi dan kegiatan fisik lainnya, merokok dll. Oleh karena itu, sistem diagnosa medis otomatis dirancang yang memenfaatkan dataset yang dikumpulkan dan sistem pendukung keputusan. Sistem ini dapat membantu dalam mendiagnosis penyakit dengan tes medis yang kurang dan pengobatan efektif.

Tujuan dari makalah ini adalah untuk membandingkan tiga metode teknik klasifikasi yang berbeda dari *statlog* (*heart*) dataset yakni *OneR, decision tre* dan *naive bayes*. Makalah ini ditulis dalam beberapa bagian. Dalam bagian II, tinjauan literatur mengenai beberapa teknik klasifikasi data mining. Selanjutnya bagian III menjelaskan dataset dan metode yang digunakan. Dalam bagian IV, hasil dan analisis dan pada bagian terakhir menyajikan kesimpulan.

2. TINJAUAN PUSTAKA

Beberapa penelitian telah dilakukan yang fokus pada diagnosis penyakit jantung. Mereka telah menerapkan teknik data mining yang berbeda untuk diagnosis dan dicapai probabilitas yang berbeda untuk metode yang berbeda dengan berbagai teknik data mining, misalnya, klasifikasi *Naive bayes*, *Support Vector Machine* (SVM), *Decision Tree* (DT), *Artificial Neural Network* (ANN), dan *Logistic Regression* dan lainnya. Aditya *et al.* [6] menganalisis teknik klasifikasi data mining menggunakan *Naive Bayes* pada dataset penyakit jantung. Sementara Kalaiselvi [4] mendiagnosis penyakit jantung menggunakan algoritma *Average K-Nearest Neighbor* dari data mining.

Sistem prediksi penyakit jantung yang cerdas yang diusulkan oleh Sellapan *et al.* [7]. Dia mengusulkan teknik data mining seperti *decision tree*, *naive bayes* dan *neural network*. Setiap metode data mining memiliki kelebihan dan kelemahan untuk memprediksi dan memberi hasil yang terbatas dan tepat. Sistem ini dibangun menggunakan pola tersembunyi dan hubungan yang terhubung dengan pola yang lainnya. Sementara Chaitrali *et al.* [3] meningkatkan studi prediksi sistem penyakit jantung menggunakan teknik klasifikasi data mining seperti *Neural Network*, *Decision Tree*, dan *Naive Bayes*.

Purushottam *et al.* [8] mengusulkan sistem prediksi penyakit jantung yang efisien menggunakan DT. Radhimeenakshi [9] melakukan penelitian dengan mengklasifikasikan dan prediksi resiko penyakit jantung menggunakan teknik data mining SVM dan ANN. Sedangkan dalam [10] melakukan penelitian mengenai madical data mining dan analisis dataset penyakit jantung menggunakan teknik klasifikasi seperti algoritma ID3 dan *Naive Bayes*.

3. METODOLOGI PENELITIAN

Pada penelitian, menggunakan dataset *statlog* (*Heart*) yang diambil dari dataset online *Machine Learning Repository* (UCI) [11]. Dataset ini berisi 13 atribut (yang telah diekstrak dari satu set yang lebih besar dari 75). Dataset ini merupakan dataset penyakit jantung yang sama dengan dataset yang sudah ada dalam repositori (dataset penyakit jantung) tetapi dalam bentuk yang sedikit berbeda. Selanjutnya, tiga teknik klasifikasi data mining yang berbeda seperti *OneR, Decision Tree*, dan *Naive Bayes* digunakan untuk analisis dataset dengan menggunakan *training set* pada tool weka.

3.1 Input atribut

a. attribute age numeric
b. attribute sex numeric
c. attribute chest_pain_tipe numeric
d. attribute resting_blood_pressure numeric
e. attribute serum_cholestoral_in_mg/dl numeric
f. attribute fasting_blood_sugar_>_120_mg/dl numeric
g. attribute resting_electrocardiographic_results numeric
h. attribute maximum_heart_rate_achieved numeric
i. attribute exercise_induced_angina numeric
j. attribute oldpeak numeric
k. attribute the_slope_of_the_peak_exercise_ST_segment numeric
l. attribute number_of_major_vessels numeric
m. attribute thal {3,6,7}

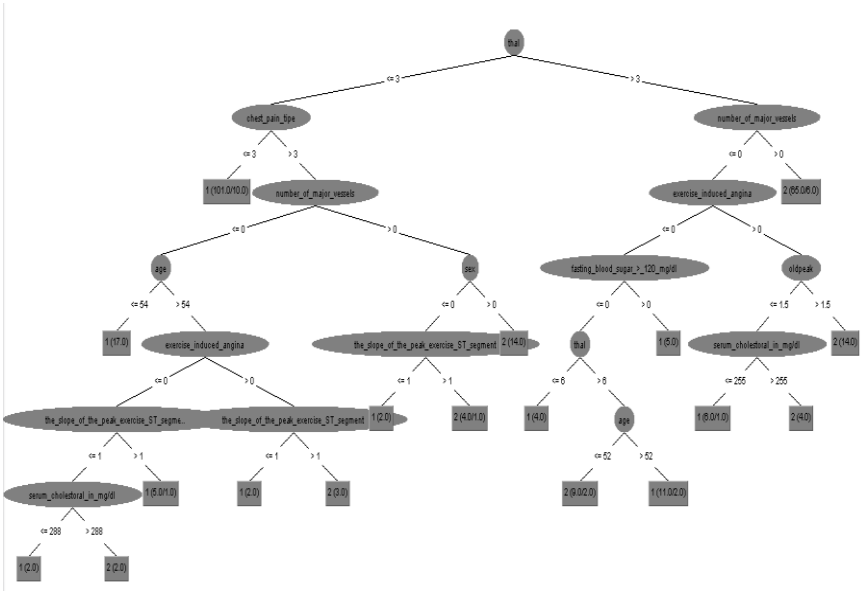
3.2 Teknik klasifikasi

3.2.1 OneR

OneR merupakan algoritma klasifikasi sederhana yang membengun pohon keputusan satu level (tingkat). Algoritma *OneR* membuat tanking atribut rerata eror pada data training seperti pengukuran entropi pada C4.5. *OneR* memperlakukan nilai numerik atribut sebagai nilai continous dan menggunakan metode sederhana untuk membagi cakupan nilai ke dalam beberapa interval terpisah. *OneR* membuat satu aturan untuk masing-masing atribut daam data training kemudian memilih *rule* dengan *error* terkecil yang disebut satu aturan (*one rule*).

3.2.2 Decision Tree (DT)

Pendekatan DT lebih kuat untuk masalah klasifikasi. Ada dua langkah dalam teknik ini membangun pohon dan menerapkan pohon untuk dataset. Ada banyak algoritma pohon keputusan CART, ID3, C5.5, CHAID, dan J48. Dari algoritma J48 ini digunakan untuk sistem ini. Algoritma J48 menggunakan metode pemangkasan untuk membangun sebuah pohon. Pemangkasan adalah teknik yang mengurangi ukuran pohon dengan menghapus *overfitting* data, yang mengarah ke akurasi yang kecil pada predikat. Algoritma J48 secara berulang mengklasifikasikan data sampai telah dikategorikan sesempurna mungkin. Teknik ini memberikan akurasi maksimum pada data training. Konsep keseluruhan adalah membangun sebuah pohon yang memberikan keseimbangan fleksibilitas dan akurasi.



Gambar 1. Hasil klasifikasi tree

3.2.1 Naive Bayes

Klasifikasi *naive bayes* didasarkan pada teorema bayes. Algoritma klasifikasi ini menggunakan kemerdekaan bersyarat, artinya menganggap bahwa nilai atribut pada kelas tertentu independen dari nilai-nilai atribut lainnya.

Teorema bayes adalah sebagai berikut:

$X= \{X_1, X_2, ..., X_n\}$ menjadi satu set dari atribut n. Dalam bayes, X dianggap sebagai bukti dan H menjadi beberapa hipotesis, data dari X milik kelas C. Kita harus menentukan P (H │X), probabilitas bahwa hipotesis H memegang bukti yaitu data sampel X. Menurut teorema Bayes P (H │X) dinyatakan sebagai:

$$P(H \mid X)=P(X \mid H)P(H)/P(X)$$

4. HASIL DAN PEMBAHASAN

Tujuan dari makalah ini yaitu untuk membandingkan tiga tektik klasifikasi data mining yang berbeda yakni OneR, Naive Bayes, dan Decision Tree dengan menggunakan bantuan tool machine learning yang disebut weka (Wakaito Environment for Knowledge Analysis) versi 3.8.0. Jika metode yang diusulkan akurat, metode kami dapat memiliki potensi untuk diterapkan untuk membantu dalam prediksi penyakit jantung. Data penyakit jantung diambil dari data set UCI.

Evaluasi diukur dengan menghitung nilai sensitivity, specivicity, dan accuracy menggunakan true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Seperti yang dirumuskan (2) to (4).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \qquad (4)$$

Sebuah *confusion matrix* diperoleh untuk menghitung akurasi klasifikasi. *Confusion matrix* memnunjukkan berapa banyak contoh yang ditugaskan untuk masing-masing kelas. Dalam percobaan kami, kami memiliki 2x2 *confusion matrix.*

Kelas 1 = absence (no heart disease)

Kelas 2 = presence (has heart disease)

Tabel 1 Confusion Matrix

	1 (no heart disease)	**2 (no heart disease)**
1 (no heart disease)	TP	FN
2 (no heart disease)	FP	TN

TP menunjukkan jumlah record diklasifikasikan benar ketika mereka benar-benar benar. FN menunjukkan jumlah record diklasifikasikan salah ketika mereka benar-benar benar. FP menunjukkan jumlah record diklasifikasikan benar ketika benar-benar palsu dan TN menunjukkan jumlah record sebagai salah ketika mereka benar-benar salah. Berikut di bawah ini hasil *confusion matrix* dari masing-masing metode yang diusulkan menggunakan dataset:

Tabel 2 Confusion Matrix For Oner

	1 (no heart disease)	**2 (no heart disease)**
1 (no heart disease)	119	31
2 (no heart disease)	33	87

Tabel 3 Confusion Matrix For Decision Tree

	1 (no heart disease)	2 (no heart disease)
1 (no heart disease)	141	9
2 (no heart disease)	14	106

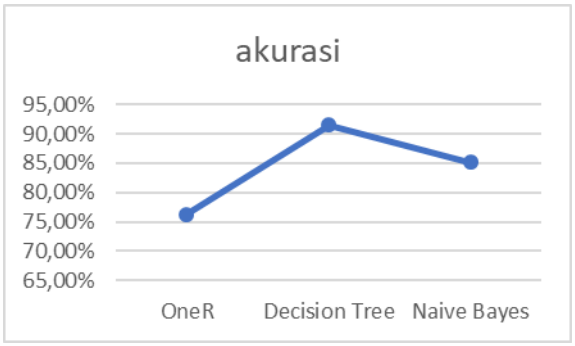
Tabel 4. Confusion Matrix For Naïve Bayes

	1 (no heart disease)	2 (no heart disease)
1 (no heart disease)	133	17
2 (no heart disease)	23	97

Tabel 5. Perbedaan Performa Teknik Klasifikasi Data Mining

Metode	Sensitivity %	Specificity %	Accuracy %
OneR	79,3	72,5	76,3
Decision Tree	94	88,3	91,48
Naive Bayes	88,7	80,8	85,18

Dari hasil confusion matrix diperoleh masing-masing akurasi 76.3% untuk *OneR*, 91,48% untuk *Decision Tree* dan 85,18% untuk *Naive Bayes*. Akurasi dari masing-masing metode di plot pada gambar 2 grafik di bawah ini:



Gambar 2. Representasi grafis dari akurasi setiap metode

5. KESIMPULAN DAN SARAN

Hasil klasifikasi penyakit jantung yang dilakukan dengan menggunakan beberapa metode klasifikasi data mining dapat membuktikan bahwa pola kejadian yang terjadi riil dapat ditransformasikan dan diprediksi dengan menggunakan machine learning. Akurasi dan performansi dari machine learning sangat tergantung dari data yang ada serta pemahaman akan metode yang diterapkan, dalam penelitian ini diperoleh akurasi prediksi tertinggi dari metode Decision tree yaitu sebesar 91,48%, selanjutnya metode Naive bayes sebesar 85,18% dan metode OneR memperoleh akurasi yang terkecil yaitu 76,3% dari metode-metode yang diusulkan. Sistem ini dapat dikembangkam untuk penelitian yang lebih lanjut disarankan perlu dilakukan pengujian lagi terhadap data yang lebih besar untuk dapat membandingkan performa akurasi teknik machine learning yang digunakan.

DAFTAR PUSTAKA

[1] S. Member, “Educational Data Mining : A Review of the State of the Art,” vol. 40, no. 6, pp. 601–618, 2010.

[2] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus, “in Databases : An Overview,” vol. 13, no. 3, pp. 57–70, 1992.

[3] C. S. Dangare and M. E. Cse, “Improved Study of Heart Disease Prediction System using Data

Mining Classification Techniques," vol. 47, no. 10, pp. 44–48, 2012.
[4] O. Ri, D. W. D. Plqlqj, P. Wr, D. Lpsuryhg, and D. Dqg, "Diagnosing of Heart Disease using Average K-Nearest Neigbor Algorithm of Data Mining," pp. 3099–3103, 2016.
[5] M. Takore and P. R. R. Shelke, "Data Mining Techniques to Find Out Heart Diseases : An Overview," vol. 4, no. Iii, pp. 80–84, 2016.
[6] N. A. Sundar, P. P. Latha, and M. R. Chandra, "Performance Analysis of Classification Data Mining Techniques over Heart Disease Data Base," no. 3, pp. 470–478, 2012.
[7] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," pp. 108–115, 2008.
[8] R. Sharma, "Efficient Heart Disease Prediction System using Decision Tree," 2015.
[9] S. Radhimeenakshi, "Classification and Prediction f Heart Disease Risk Using Data Mining Techniques of Support Vector Machine and Artificial Neural Network," pp. 3107–3111, 2016.
[10] S. Ranganatha, P. R. H. R, C. Anusha, and S. K. Vinay, "Medical Data Mining and Analysis for Heart Disease Dataset using Classification Techniques."
[11] "Machine Learning Repository (UCI)." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Statlog+(Heart).