ISSN: 2548-9704

ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI PENDERITA PENYAKIT JANTUNG

Riski Annisa

Program Studi Sistem Informasi Akuntansi Kampus Kota Pontianak, Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika
Jl. Abdurrahman Saleh No. 18, Pontianak, Kalimantan Barat
E-mail :riski.rnc@bsi.ac.id

ABSTRAK

Penyakit jantung adalah istilah umum untuk semua jenis gangguan yang mempengaruhi jantung. Penyakit jantung berarti sama dengan penyakit jantung tetapi tidak penyakit kardiovaskular. Penelitian ini akan melakukan perbandingan beberapa algoritma klasifikasi yaitu Decision Tree, Naïve Bayes, k-Nearest Neighbour, Random Forest, dan Decison Stump dengan menggunakan uji parametrik dengan *t-test* agar dapat menghasilkan perbandingan metode yang lebih baik untuk data set laki-laki penderita Penyakit jantung. Hasil penelitian mendapatkan nilai akurasi sebesar tertinggi sebesar 80.38%. Hasil penelitian menunjukkan bahwa algoritma random forest dan decision stump melakukan performa terbaik dalam pengklasifikasi di dataset, C4.5 dan Naïve bayes juga tampil baik, kemudian k-NN merupakan algoritma yang kurang baik diimplementasikan dalam dataset.

Kata kunci: Penyakit Jantung, Algoritma, Klasifikasi

ABSTRACT

Heart disease is a general term for all types of disorders that affect the heart. Heart disease means the same as heart disease but not cardiovascular disease. This study will compare several classification algorithms, namely Decision Tree, Naïve Bayes, k-Nearest Neighbor, Random Forest, and Decison Stump by using parametric tests with a t-test in order to produce better method comparisons for data sets of men with Disease heart. The results of the study get the highest accuracy value of 80.38%. The results showed that the random forest algorithm and decision stump performed the best performance in classifiers in the dataset, C4.5 and Naïve Bayes also performed well, then k-NN was an algorithm that was not well implemented in the dataset.

Keyword : Heart Disease, Algorithm, Classification

I. PENDAHULUAN

Penyakit jantung adalah istilah umum untuk semua jenis gangguan yang mempengaruhi jantung. Penyakit jantung berarti sama dengan penyakit jantung tetapi tidak penyakit kardiovaskular. Penyakit kardiovaskular mengacu pada gangguan pembuluh darah dan jantung, sedangkan penyakit jantung mengacu hanya hati. Menurut WHO (Organisasi Kesehatan Dunia) dan CDC, penyakit jantung adalah penyebab utama kematian di Inggris, Amerika Serikat, Kanada dan Australia. Jumlah orang dewasa AS yang

ISSN: 2548-9704

didiagnosis dengan penyakit jantung berdiri di 26,6 juta (11,3% dari populasi orang dewasa) [1].

Klasifikasi adalah jenis analisis data yang dapat membantu orang memprediksi label kelas sampel harus diklasifikasikan. Berbagai macam teknik klasifikasi telah diusulkan dalam bidang-bidang seperti pembelajaran mesin, sistem pakar dan statistik [2].

Sejak 1990-an peneliti telah mengembangkan repositori software untuk mendapatkan pemahaman yang lebih dalam mengenai data [3]. Penggunaan metode Decision tree (C4.5), K-Nearest Neighbour (k-NN), Naïve Bayes, Random Forest, Decision Stump yang membandingkan algoritma C4.5, Random Forest, dan SimpleCART dan menghasilkan analisis bahwa algoritma C4.5 bekerja lebih baik untuk sistem [3]. Selain itu dalam kasus penurunan klasifikasi untuk atribut ketegori dan numerik. Dari dua puluh metode klasifikasi, Bayes Net, Naïve Bayes, Klasifikasi melalui Regresi, Regresi Logistik dan Random Forest merupakan metode klasifikasi terbaik. Untuk dataset atribut campuran Naïve Bayes, Bayes Net dan Random Forest merupakan metode klasifikasi terbaik. Untuk dataset atribut numerik Klasifikasi Regresi, NBTree dan multiclass Classifier adalah metode yang terbaik. Untuk atribut kategorikal dataset NB-Tree, Klasifikasi melalui Regresi dan metode Bayes Net adalah yang terbaik. Dari ini di atas lima aturan metode klasifikasi berdasarkan PART dan metode Decision Tree adalah yang terbaik [4]. Tetapi, beberapa metode yang telah menunjukkan hasil terbaik dimana teknik yang tepat dipilih untuk data yang tepat. Tidak ada pengklasifikasi tertentu yang melakukan yang terbaik untuk semua dataset [5].

Penelitian ini akan melakukan perbandingan beberapa algoritma klasifikasi yaitu Decision Tree, Naïve Bayes, k-Nearest Neighbour, Random Forest, dan Decison Stump dengan menggunakan uji parametrik dengan *t-test* agar dapat menghasilkan perbandingan metode yang lebih baik untuk data set laki-laki penderita Penyakit jantung.

Paper ini disusun dengan urutan sebagai berikut: Pada bagian 2, menjelaskan penelitian terkait dan teori yang digunakan. Pada bagian 3 hasil dan pembahasan akan dipaparkan yaitu disajikan perbandingan hasil eksperimen metode. Kemudian pada bagian akhir akan disampaikan kesimpulan dari penelitian yang dilakukan.

2. METODOLOGI

Penelitian tentang komparasi algoritma data mining telah banyak dilakukan dan dipublikasikan. Untuk melakukan penelitian ini perlu ada kajian terhadap penelitian yang terkait sebelumnya agar dapat mengetahui metode apa saja dan hasil seperti yang dihasilkan.

Penelitian yang dilakukan oleh Yu, Chen, Koronios, Zhu, dan Guo [2] melakukan penelitian untuk membantu bank meningkatkan kualitas kredit dan menurunkan resiko dengan mengimplementasikan teknik klasifikasi utama, yang meliputi model statistik biasa (LDA, QDA, dan regresi logistik), k-nearest neighbour, Bayesian (Naïve Bayes dan TAN), decision tree (C4.5), associative classification (CBA), neural network dan support vector machines (SVM), dan algoritma tersebut untuk mengendalikan resiko kredit. Percobaan dilakukan pada 244 perusahaan terutama dari Industrial and Commercial Bank of China. Dengan menggunakan metode Delong-Pearson untuk memverifikasi dan membandingkan algoritma tersebut. Hasil menunjukkan bahwa model statistika biasa menghasilkan hasil yang buruk, C4.5 atau SVM tidak menunjukkan hasil yang memuaskan, dan CBA menjadi pilihan.

Penelitian yang dilakukan oleh

ISSN: 2548-9704

Wahono, Suryana, dan Ahmad [5], melakukan penelitian terhadap Software Defect Prediction dengan beberapa jenis algoritma klasifikasi untuk memprediksi kerusakan perangkat lunak. Dalam penelitian ini, kerangka perbandingan diusulkan sebagai patokan kinerja berbagai model klasifikasi dalam bidang prediksi kerusakan perangkat lunak. Dengan menggunakan 10 pengklasifikasi dipilih dan diterapkan untuk membangun model klasifikasi dan menguji kinerja pada 9 dataset NASA MDP. AUC digunakan sebagai indikator akurasi untuk mengevaluasi kinerja metode klasifikasi. Friedman dan Nemenyi post hoc tes digunakan untuk menguji signifikansi perbedaan AUC antara metode klasifikasi. Hasil penelitian menunjukkan bahwa regresi logistik tampil terbaik dalam kebanyakan dataset NASA MDP. Naïve bayes, neural network, support vector machine dan k* classifiers juga tampil baik. Klasifikasi decision tree, linear discriminant analysis, dan k-nearest neighbour cenderung underperform.

Penelitian yang dilakukan Saha dan Nandi [4], menerapkan 20 metode klasifikasi data mining pada dataset medis dengan memvariasikan jumlah atribut kategorikal dan numerik, jenis atribut dan jumlah kasus di dataset. Classification Accuracy (CA), Root Mean Square Error (RMSE) dan Area Under Curve (AUC) dari Receiver's Operational Characteristics (ROC) digunakan sebagai metric dari: 1) Performa klasifikasi metode tergantung pada jenis variabel dataset atau atribut seperti kategori, numerik, dan keduanya, 2) Performa metode klasifikasi dengan atribut kategorikal lebih baik dari pada atribut numerik dari dataset, 3) Tingkat akurasi, RMSE, dan AUC dari metode klasifikasi tergantung pada jumlah kasus di dataset, 4) Dalam kasus penurunan performa klasifikasi dari atribut kategori dan numerik, 5) Tiga metode klasifikasi tertinggi ditetapkan setelah membandingkan performa dari 20 metode klasifikasi untuk atribut kategori, numerik, dan keduanya, 6) Dari dua puluh metode klasifikasi yang berbeda, Bayes Net, Naïve Bayes, classification via regression, logistic regression, dan Random Forest merupakan metode terbaik untuk dataset medis.

2.1 Decision Tree (C4.5)

Decision Tree adalah algoritma klasifikasi yang dinyatakan sebagai partisi rekursif dari ruang contoh. Decision Tree terdiri dari node yang membentuk pohon berakar, yang berarti pohon diarahkan dengan simpul yang disebut akar. Sebuah node dengan tepi keluar disebut internal atau tes node. Semua node yang lain disebut daun. Dalam pohon keputusan, setiap simpul internal membagi ruang misalnya menjadi dua atau lebih sub ruang sesuai dengan fungsi diskrit tertentu dari atribut nilai [6].

Sebuah pohon keputusan terdiri dari internal node yang menentukan tes pada variabel masukan individu atau atribut yang membagi data menjadi himpunan bagian yang lebih kecil, dan serangkaian node daun menetapkan kelas untuk masing-masing pengamatan di segmen yang dihasilkan. Pada penelitian ini, C4.5 membangun pohon keputusan dengan menggunakan konsep entropi informasi. Entropi sampel S dari pengamatan yang diklasifikasikan diberikan oleh

$$\text{Entropy (S)} = -p_1 \log_2 (p_1) - p_0 \log_2 (p_0),$$

Dimana p_1 adalah proporsi dari kelas yang nilainya 1 dan p_0 proporsi dari kelas yang nilainya 0 dalam sample S. C4.5 memeriksa informasi *Gain* normalisasi (perbedaan entropi) yang dihasilkan dari memilih atribut untuk membagi data. Atribut dengan informasi *Gain* normalisasi tertinggi adalah yang digunakan untuk membuat keputusan. Algoritma kemudian terbagi lagi menjadi subset lebih kecil [7].

2.2 K-Nearest Neighbour (k-NN)

K-nearest neighbour merupakan metode klasifikasi pertama yang menempatkan

ISSN: 2548-9704

point *k* data yang paling mirip dengan titik data *k* terdekat untuk menentukan kelas target titik terdekat. Untuk menentukan k-nearest neighbour dari titik data perlu mengukur persamaan atau perbedaan antara titik data, titik data tersebut dapat diukur dengan beberapa teknik misalnya, Euclidean distance, Minkowski distance, Hamming distance, koefisian korelasi Pearson, dan persamaan cosine [8]. Euclidean distance didefinisikan sebagai berikut:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{l=1}^{p} \left(x_{i,l} - x_{j,l}\right)^2}, \quad i \neq j.$$

Euclidean distance adalah teknik membandingan perbedaan dari dua data (x_i dan x_j). Semakin besar jarak Euclidean, semakin beda dua point data, dan dua titik data yang jauh terpisah dipisahkan dalam ruang data dimensi *p*.

Minkowski didefinisikan sebagai berikut:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{l=1}^{p} \left| x_{i,l} - x_{j,l} \right|^r \right)^{1/r}, \quad i \neq j.$$

Berikut ini merupakan koefisien korelasi Pearson *p*:

$$\rho_{\boldsymbol{x}_i \boldsymbol{x}_j} = \frac{s_{\boldsymbol{x}_i \boldsymbol{x}_j}}{s_{\boldsymbol{x}_i} s_{\boldsymbol{x}_j}},$$

Persamaan cosine menganggap dua titik data x_i dan x_j sebagai dua vektor dalam ruang dimensi *p* dan menggunakan cosinus dari sudut θ antara dua vektor untuk mengukur kesamaan dari dua titik data:

$$\cos(\theta) = \frac{\boldsymbol{x}_i' \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|},$$

2.3 Naïve Bayes

Naïve Bayes didasarkan pada teorema Bayes. Oleh karena itu, meninjau teorema Bayes dan kemudian menggambarkan klasifikasi. List dari paket *software* data mining yang mendukung pembelajaran klasifikasi Naïve Bayes tersedia. Beberapa aplikasi klasifikasi Naïve Bayes tersedia dengan referensi [8]. Teorema Bayes berasal dari persamaan:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Klasifikasi Naïve Bayes memperkirakan persamaan probabilitas berikut:

$$P(y) = \frac{n_y}{n}$$

$$P(x_i|y) = \frac{n_{y\&x_i}}{n_y},$$

Keterangan:

n , total nomor dari point data pada data set training
n_y , nomor dari point data target class *y*
$n_{y\&xi}$, nomor dari point data dengan target class *y*
i , variabel atribut yang mengambil nilai dari x_i

2.4 Random Forest

Random forest didefinisikan sebagai kelompok klasifikasi atau *regresi tree* tidak ditebang, dilatih pada sample *bootstrap* dari data pelatihan menggunakan pilihan fitur acak dalam proses *generate tree*. Setelah sejumlah besar tree telah di-generate, setiap tree divoting untuk mendapatkan kelas yang paling populer. Prosedur *voting tree* secara kolektif ini didevinisikan sebagai random forest. Untuk teknik klasifikasi random forest ini membutuhkan dua parameter yaitu jumlah tree dan jumlah atribut yang digunakan untuk memperbanyak tree.

2.5 Decision Stump

Decision Stump pada dasarnya merupakan pohon keputusan dengan lapisan tunggal. Perbandingan dari pohon yang memiliki beberapa lapisan, Decision Stump pada dasarnya berhenti setelah split pertama. Decision Stump biasanya digunakan dalam segmentasi populasi untuk data yang besar. Kadang-kadang, juga digunakan untuk membantu membuat simpleyes atau tidak ada model keputusan untuk data yang lebih

ISSN: 2548-9704

kecil dengan sedikit data. Decision Stump umumnya lebih mudah untuk dibangun dibandingkan dengan Decision Tree. Pada saat yang sama, coding SAS untuk Decision Stump lebih mudah dikelola dibandingkan dengan CART dan CHAID. Alasannya adalah bahwa Decision Stump adalah hanya satu single run dari algoritma tree dan dengan demikian tidak perlu mempersiapkan data untuk split berikutnya. Pada saat yang sama, tidak butuh menentukan data untuk split berikutnya yang membuat penggantian nama dari pengelolaan output sederhana [3].

3. HASIL DAN PEMBAHASAN

Metode yang diusulkan pada penelitian ini yaitu dengan menggunakan algoritma untuk mengkomparasi algoritma klasifikasi Decision Tree, Naïve Bayes, k-Nearest Neighbour, Random Forest, dan Decison Stump diuji pada dataset penderita penyakit jantung laki-laki. Untuk mengukur kinerja algoritma klasifikasi ini dengan menggunakan dataset yang tersedia secara publik di UCI Repository yaitu dengan dataset penderita penyakit jantung laki-laki. Dataset yang digunakan mempunyai 8 atribut terdiri dari data nominal dan numerik. Selanjunya untuk validasi menggunakan *10-fold cross validation*. Hasil pengukuran algoritma menggunakan uji t (*t-test*) untuk mengetahui perbedaan kinerja model.

Proses pengujian metode dimulai dari pembagian dataset dengan metode *10-fold cross validation* yang membagi dataset menjadi dua yaitu data training dan data testing. Selanjutnya diterapkan tahapan evaluasi menggunakan *Area Under Curve* (AUC) untuk mengukur hasil akurasi dari performa model klasifikasi. Hasil akurasi dilihat menggunakan curva *Receiver Operating Characteristic* (ROC) dan hasil *confusion matrix*. ROC menghasilkan dua garis dengan bentuk true positive sebagai garis vertikal dan false positive sebagai garis horizontal. Pengukuran akurasi dengan *confusion matrix* dapat dilihat pada tabel berikut ini:

Tabel 1. *Confusion Matrix*

		Aktual	
		True	***False***
Predicted	***True***	*True Positive* (TP)	*False Negative* (FN)
	False	*False Negative* (FN)	*True Negative* (TN)

Formulasi perhitungan adalah sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Sensitivity = TP_{rate} = \frac{TP}{TP + FN}$$

$$Spesificity = TN_{rate} = \frac{TN}{TN + FP}$$

$$FP_{rate} = \frac{FP}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F\ Measure = \frac{2RP}{R + P}$$

$$G - Mean = \sqrt{sensitivity * specificity}$$

Dalam pengklasifikasian data menggunakan AUC penjelasannya sebagai berikut:

Tabel 2. Nilai AUC dan Keterangan

Nilai AUC	Klasifikasi
0.90 - 1.00	*excellent classification*
0.80 - 0.90	*good classification*
0.70 - 0.80	*fair classification*
0.60 - 0.70	*poor classification*
0.50 - 0.60	*failure*

Evaluasi dalam penelitian ini adalah menggunakan uji t (*t-test*). Uji t adalah membandingkan hubungan antara dua variabel yaitu variabel respon dan variabel *predictor*. Uji t sample berpasangan

ISSN: 2548-9704

(*paired-sample t-test*) digunakan untuk menguji perbandingan selisih dua rata-rata dari sua sample yang berpasangan dengan asumsi bahwa data terdistribusi normal.

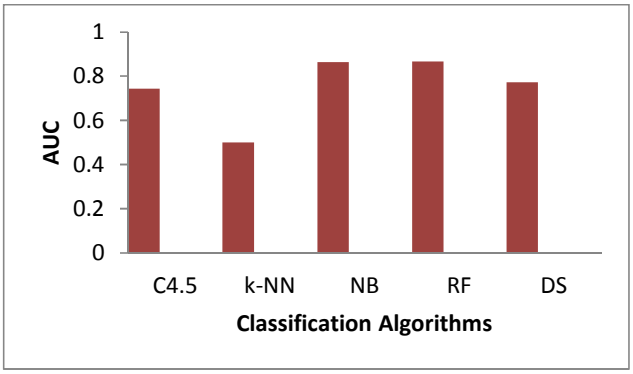
Eksperimen yang dilakukan dalam penelitian ini menggunakan sebuah platform komputer berbasis Intel(R) Celeron(R) CPU 1007U @1.50GHz, RAM 2GB, dan sistem operasi Microsoft Windows 10 Pro 64-bit. Sedangkan lingkungan pengembangan aplikasi menggunakan tools rapid miner 7.0 library.

Dalam eksperimen ini menggunakan dataset yang terdiri dari 8 atribut, datanya berupa numerik dan nominal. Metode yang diuji adalah algoritma klasifikasi decision tree, Naïve bayes, k-nearest neighbour, random forest, decision stump. Hasil eksperimen yang disajikan adalah *accuracy*, *precision*, *recall*, dan AUC.

Tabel 3. Hasil Eksperimen Algoritma Klasifikasi

	C4.5	**k-NN**	**NB**	**RF**	**DS**
Accuracy	76.08 %	60.77 %	78.95%	80.38 %	78.95 %
Precision	81.90 %	64.71 %	79.67%	78.36 %	76.64 %
Recall	73.50 %	65.81 %	83.76%	89.74 %	89.74 %
AUC	0.743	0.500	0.863	0.867	0.772

Dari tabel diatas algoritma klasifikasi yang menunjukkan accuracy paling tinggi adalah algoritma random forest yang tingkat accuracy sebesar 80.38% dan berdasarkan klasifikasi nilai AUC algoritma yang tingat accuracy failure adalah k-Nearest Neighbour, C4.5 dan Decison Stump tergolong *fair classification*, Naïve Bayes dan Random Forest tergolong *good classification*.



Gambar 1. AUC 5 algoritma klasifikasi

Pada penelitian ini dilakukan pengujian dengan menggunakan uji *t-test*. Dalam signifikansi uji nilai menggunakan tingkat signifikasi statistik menjadi 0.05. Berarti secara statistik kurang dari 0.05 menunjukkan perbedaan signifikan antara nilai rata-rata, dengan demikian harus menolak hipotesis nol dan berarti harus menolak hipotesis nol (H_0). Selanjutnya dilakukan uji *t-test* untuk mendeteksi pengklasifikasi berbeda secara signifikan.

Tabel 4. Hasil uji *t-test*

	C4.5	**k-NN**	**NB**	**RF**	**DS**
C4.5	-	0.003	0.511	0.189	0.474
k-NN	0.003	-	0.001	0.000	0.001
NB	0.511	0.001	-	0.707	0.991
RF	0.189	0.000	0.707	-	0.654
DS	0.474	0.001	0.991	0.654	-

Berdasarkan tabel diatas, diketahui bahwa terdapat perbedaan signifikan (H_1) antara algoritma C4.5 dengan k-NN, random forest, dan decision stump sedangkan k-nearest neighbour dengan algoritma Naïve bayes, random forest, dan decision stump. H_0 menjelaskan tidak terdapat perbedaan signifikan antara algoritma C4.5 dengan Naïve bayes, Naïve bayes dengan random forest dan decision stump. Dengan demikian algoritma C4.5 tidak ada perbedaan signifikan dengan Naïve Bayes dan Naïve bayes tidak ada perbedaan

ISSN: 2548-9704

signifikan dengan random forest dan decision stump namun C4.5 terdapat perbedaan dengan random forest dan decision stump. Sedangkan k-nearest neighbour berbeda secara signifikan dengan algoritma lain.

4. KESIMPULAN

Penelitian dengan menggunakan dataset penderita penyakit jantung dengan mengkomparasi 5 algoritma yaitu decision tree, k-nearest neighbour, Naïve bayes, random forest, dan decision stump. Dengan menggunakan validasi *10-fold cross validation* dan uji *t-test*. Hasil penelitian mendapatkan nilai akurasi sebesar tertinggi sebesar 80.38%. Hasil penelitian menunjukkan bahwa algoritma random forest dan decision stump melakukan performa terbaik dalam pengklasifikasi di dataset, C4.5 dan Naïve bayes juga tampil baik, kemudian k-NN merupakan algoritma yang kurang baik diimplementasikan dalam dataset.

DAFTAR PUSTAKA

[1] Nordqvist, Christian. (2016). Heart Disease Heart Disease: Definition, Causes, Research. 7 April 2016. http://www.medicalnewstoday.com/articles/237191.php

[2] Yu, L., Chen, G., Koronios, A., Zhu, S., & Guo, X. (n.d.). Application and Comparison of Classification Techniques in Controlling Credit Risk. *World*, 2007–2007.

[3] Undavia, J. N. (2014). Comparison of Classification Algorithms to Predict Comparison of Decision Tree Classification Algorithm to Predict Student ' s Post Graduation Degree in Weka Environment, *1*(2), 17–22.

[4] Saha, S. (n.d.). Data Classification based on Decision Tree , Rule Generation , Bayes and Statistical Methods : An Empirical Comparison, *129*(7), 36–41.

[5] Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). A comparison framework of classification models for software defect prediction. *Advanced Science Letters*, *20*(10–12), 1945–1950. http://doi.org/10.1166/asl.2014.5640

[6] Rokach, L., & Maimon, O. (2010). Classification Trees. *Data Mining and Knowledge Discovery Handbook*, 149–174. http://doi.org/10.1007/978-0-387-09823-4

[7] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. http://doi.org/10.1016/j.eswa.2011.09.033

[8] Fu, Y. F. Y. (1997). *Data mining*. *IEEE Potentials* (Vol. 16). http://doi.org/10.1109/45.624335