Prosiding Annual Research Seminar 2019

Computer Science and ICT

ISBN : 978-979-587-846-9
Vol.5 No.1

Prediksi Penyakit Jantung dengan Algoritma Klasifikasi

1st Pandito Dewa Putra
Fakultas Ilmu Komputer
Universitas Sriwijaya Palembang, Indonesia
panditodewaputra@gmail.com

2nd Dian Palupi Rini
Fakultas Ilmu Komputer
Universitas Sriwijaya Palembang, Indonesia
dprini@unsri.ac.id

Abstract— One of the non-communicable diseases (PTM) that is susceptible occurs especially when an individual is at a productive age, namely heart disease (Heart Disease). Heart disease is also very vulnerable to attack men with age range 60 years and under. The object of this study used a statistic heart disease dataset with 270 data records. The research methodology has been used in this study compared the naïve Bayes classification algorithm, support vector machine, C.45, logistic regression, and backpropagation. Next, this study did cross-validation to see the performance of the accuracy, precision and recall of each of these algorithms.

Keywords — Heart Disease, Classification, Data Mining

***Abstract*— Salah satu Penyakit Tidak Menular (PTM) yang rentan terjadi terutama saat seorang individu berada pada usia produktif yaitu penyakit jantung (*Heart Disease*). Penyakit jantung juga sangat rentan menyerang laki–laki dengan rentang usia 60 tahun kebawah. Objek penelitian ini akan menggunakan *statlog heart disease dataset* dengan 270 *record* data. Metodologi penelitian yang akan digunakan dalam penelitian ini membandingkan Algoritma klasifikasi *naïve bayes, support vector machine, C.45, logistic regression*, dan *back propagation*. Selanjutnya melakukan *cross validation* untuk melihat performa akurasi, presisi dan *recall* dari masing– masing algoritma tersebut.**

Kata Kunci—Penyakit Jantung, Klasifikasi, Data mining

I. PENDAHULUAN

Kesehatan merupakan faktor terpenting yang harus dijaga oleh setiap individu, agar dalam menjalani aktivitas sehari– hari menjadi lebih produktif, tidak mudah lelah dan lebih fokus dalam menyelesaikan suatu *project*. Namun, saat ini PTM menjadi penyebab utama kematian. Setiap tahunnya terdapat lebih dari 36 juta orang yang meninggal disebabkan PTM atau setara 63% dari total seluruh kematian. Secara terperinci PTM terjadi dibawah usia 60 tahun, dan mayoritas sampai 90% dari kematian tersebut terjadi di negara dengan tingkat perekonomian rendah [1].
Salah satu PTM yang rentan terjadi terutama saat seorang individu berada pada usia produktif yaitu penyakit jantung (*Heart Disease*). Tingginya faktor kematian dari penyakit jantung karena kurangnya pengetahuan masyarakat terhadap gejala atau tanda–tanda saat seseorang tersebut mengidap penyakit ini [2]. Penyakit jantung merupakan salah satu penyakit yang cukup berbahaya ketika menyerang seseorang, dimana penyebab utama penyakit jantung yaitu berasal dari pola hidup individu yang kurang sehat, mengkonsumsi makanan berkolesterol tinggi, penggunaan alkohol, tembakau, diet yang ekstrem serta penyebab lainnya. Penyakit jantung lebih rentan diderita oleh laki–laki, dimana perbandingannya sekitar satu dari tiga kemungkinan mengidap penyakit jantung sebelum usia 60 tahun. Pada wanita perbandingannya sekitar satu dari sepuluh kemungkinan yang mengidap penyakit jantung. Rasio yang cukup tinggi terkait penyakit jantung ini, menjadikannya sebagai salah satu penyakit yang akan menghasilkan sejumlah besar data pasien pengidap penyakit jantung. Bahkan industri perawatan kesehatan saat ini mampu memberikan data kompleks tentang pasien, sumber daya rumah sakit, diagnosa penyakit, catatan pasien elektronik, peralatan medis, dll. Penambangan data menjadi kian disoroti terutama pada berbagai layanan kesehatan. Besarnya jumlah data merupakan sumber daya utama yang akan diproses dan dianalisis guna diekstraksi pengetahuan yang memungkinkan dukungan untuk penghematan biaya dan pengambilan keputusan. Data mining merupakan proses untuk menemukan pola dan tren yang sebelumnya tidak diketahui dalam *database* serta menggunakan informasi tersebut dalam membangun model prediksi [3]. Penambangan data menyediakan seperangkat alat dan teknik yang dapat diterapkan pada data yang diproses ini untuk menemukan pola

tersembunyi dan juga memberikan sumber pengetahuan tambahan kepada profesional kesehatan untuk membuat keputusan yang lebih akurat.

Penelitian sebelumnya [4] menggunakan The Statlog (*Heart Disease*) *dataset*, dimana akurasi tertinggi (92,59%) didapat dengan menggunakan klasifikasi *ensemble* C4.5 *decision tree* dibanding dengan algoritma lain. Kemudian pada penelitian sebelumnya [5] juga menggunakan The Statlog (*Heart Disease) dataset*, dimana hasil akurasi tertinggi didapat oleh algoritma *Logistic Regression* dengan akurasi 85%. Selanjutnya, pada penelitian lain [6] juga Menggunakan The Statlog (*Heart Disease) dataset*, hasil akurasi tertinggi didapat oleh metode Naive Bayes sebesar 84%. Kemudian pada penelitian sebelumnya [7] menggunakan Cleveland *heart dataset*, dimana *Bagging* dan *boosting Accuracy* tertinggi dimiliki algoritma naïve bayes dengan akurasi 83.17% dan 84.16%. Terakhir pada penelitian sebelumnya [8] menggunakan *Cleveland heart dataset* dengan akurasi tertinggi menggunakan algoritma *Support Vector Machine* (SVM) didapatkan sebesar 86.87%. Berdasarkan penelitian– penelitian di atas, maka penelitian ini akan melakukan sebuah komparasi terhadap algoritma-algoritma yang memiliki tingkat akurasi tertinggi tersebut dengan menggunakan objek *statlog heart disease dataset*. Sehingga berbagai perbandingan algoritma tersebut akan menghasilkan algoritma dengan akurasi paling baik.

II. DASAR TEORI

A. Naïve Bayes

Langkah kerja algoritma naïve bayes :

1. Biarkan D menjadi seperangkat pelatihan tupel dan label kelas yang terkait. Seperti biasa, setiap tupel diwakili oleh vektor atribut n-dimensi, X = (x1, x2, ..., xn), yang menggambarkan pengukuran n yang dilakukan pada tupel dari n atribut, masing-masing, A1, A2, .., An.
2. jika kelas m, C1, C2, ..., Cm. Diberikan tuple, X, classifier prediksi X dengan probabilitas posterior tertinggi, dikondisikan pada X. Artinya, classifier Bayesian naif memperkirakan bahwa tuple x milik kelas Ci jika dan hanya jika
   P (Ci|X) > P (Cj|X) for $1 \leq j \leq m, j \neq i$
   Kemudian maximum P(Ci|X). Kelas Ci yang P(Ci|X) dilakukan hipotesis posteriori maksimum.
   $P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$
3. Karena P(X) konstan untuk semua kelas, hanya P(X|Ci) P(Ci) yang perlu dimaksimalkan. Jika probabilitas kelas sebelumnya tidak diketahui, maka umumnya diasumsikan bahwa kelas-kelas tersebut kemungkinan besar sama, yaitu, P(C1) = P(C2) = ... = P(Cm), dan oleh karena itu kami akan memaksimalkan P (X|Ci). Jika tidak, kami memaksimalkan P(X|Ci) P(Ci). Perhatikan bahwa probabilitas kelas sebelumnya dapat diperkirakan dengan P(Ci) = |Ci,D|/|D|, di mana |Ci,D| adalah jumlah tupel pelatihan kelas Ci di D.
4. Mengingat set data dengan banyak atribut, akan sangat mahal secara komputasi untuk menghitung P(X|Ci). Untuk mengurangi perhitungan dalam mengevaluasi P(X|Ci), asumsi naif tentang independensi kondisional kelas dibuat. Ini mengasumsikan bahwa nilai-nilai atribut secara kondisional independen satu sama lain, diberi label kelas tuple (yaitu, bahwa tidak ada hubungan ketergantungan antara atribut). Jadi
   $P(X|Ci) = \prod_{k=1}^{n} P(xk|Ci)$
   = P(x1|Ci)x P(x2|Ci)x… P(xm|Ci).
   Kita dapat dengan mudah memperkirakan probabilitas P(x1|Ci), P(x2|Ci), ..., P(xm|Ci) dari tuple pelatihan. Ingatlah bahwa di sini xk merujuk pada nilai atribut Ak untuk tuple X. Untuk setiap atribut, kita melihat apakah atribut tersebut kategorikal atau bernilai kontinyu. Misalnya, untuk menghitung P(X|Ci), kami mempertimbangkan hal berikut:
   (a) Jika Ak adalah kategorikal, maka P(Xk |Ci) adalah jumlah tupel kelas Ci dalam D yang memiliki nilai xk untuk Ak, dibagi dengan |Ci,D|, jumlah tupel kelas Ci di D.
   (b) Jika Ak adalah nilai berkelanjutan, maka kita perlu melakukan sedikit lebih banyak pekerjaan, tetapi perhitungannya cukup mudah. Atribut bernilai kontinu biasanya diasumsikan memiliki distribusi Gaussian dengan rata-rata μ dan standar deviasi σ, yang didefinisikan oleh

   $G(x, \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
   Hingga
   $P(xk \mid Ci) = g(xk, \mu ci, \sigma ci)$

   Selanjutnya mengeksekusi μci dan σci, yang merupakan mean dan standar deviasi, dari nilai-nilai atribut Ak untuk tupel pelatihan kelas Ci. Kami kemudian memasukkan dua jumlah ini ke dalam persamaan di atas.
5. kemudian melakukan prediksi label kelas X, P(X|Ci)P(Ci) dievaluasi untuk setiap kelas Ci. Pengklasifikasi memprediksi bahwa label kelas tuple X adalah kelas Ci jika dan hanya jika
   P(X|Ci)P(Ci)>P(X|Cj)P(Cj) for $1 \leq j \leq m, j \neq i$
   Dengan kata lain, label kelas yang diprediksi adalah kelas Ci yang P(X|Ci) P(Ci) maksimum [9].

B. Support Vector Machine

SVM dapat digunakan untuk klasifikasi pola dan regresi nonlinier. Lebih tepatnya, SVM adalah perkiraan penerapan metode minimalisasi risiko struktural. SVM menerapkan tingkat *error* pada *machine learning* pada data uji, dimana tingkat *error train* dan istilah yang bergantung pada dimensi Vapnik-Chervonenkis (VC). *Support vector machine* dapat memberikan kinerja generalisasi yang baik pada masalah klasifikasi pola [10].

Hyperplane optimal untuk pola: Pertimbangkan sampel

pelatihan {(xi,yi)}Ni=1 dimana xi adalah pola input untuk instance engan dan yi adalah output target yang sesuai. Dengan pola yang diwakili oleh subset yi = +1 dan pola yang diwakili oleh subset yi = -1 dapat dipisahkan secara linear. Persamaan dalam bentuk hyperplane yang melakukan pemisahan adalah $w^T x + b=0$
di mana x adalah vektor input, w adalah vektor bobot yang dapat disesuaikan, dan b adalah bias. Jadi,

$w^T x + b \geq 0$ for yi = +1 $w^T x + b \leq 0$ for yi = -1

C. C.45

Algoritma C4.5 dibentuk berdasarkan kriteria pengambilan keputusan. Pohon keputusan adalah metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mewakili aturan. Secara umum, algoritma C4.5 membangun pohon keputusan dengan memilih atribut sebagai *root*, membuat cabang untuk setiap nilai untuk kasus di cabang, dan proses akan diulang untuk setiap cabang sampai kasus di cabang memiliki kelas yang sama [11]. Pemilihan atribut didasarkan pada nilai *gain* tertinggi, menggunakan persamaan.

$$\text{Gain(S,A)} = \text{Entropy(S)} - \sum_{i=0}^{n} \frac{|Si|}{|S|} Entropy(Si)$$

di mana S adalah set kasus, A adalah atribut, N adalah sejumlah partisi atribut A, |Si| adalah jumlah case pada partisi ke-i, dan |S| adalah jumlah kasus di S. Nilai *entropy* dihitung sebelumnya mendapatkan nilai *gain*. *Entropy* digunakan untuk menentukan seberapa informatif atribut untuk menghasilkan atribut. Rumus dasar *entropy* adalah seperti dalam persamaan.

$$\text{Entropy(S)} = \sum_{i=1}^{n} pi * log2pi$$

D. Logistic Regression (LR)

LR sangat berguna untuk memprediksi ada atau tidaknya karakteristik atau hasil berdasarkan nilai dari set variabel prediktor. Hal ini mirip dengan model *regresi linier* tetapi cocok untuk model dimana variabel dependennya adalah dikotomis [12]. Model LR untuk p variabel independen dapat ditulis sebagai

$$P(Y=1) = \frac{1}{1+e^{-(\beta 0+\beta 1x\beta 1+\beta 2x\beta 2+\cdots+\beta p+q)}}$$

di mana P (Y = 1) adalah probabilitas kehadiran CAD, dan $\beta_0, \beta_1, \ldots \beta_p$ dalah koefisien regresi. Ada model linier yang tersembunyi dalam model regresi logistik. Logaritma natural dari rasio P (Y = 1) ke (1 P (Y = 1)) memberikan model linier dalam Xi:

$$g(x) = 1n(\frac{p(y=1}{1-p(y=1)}$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

G(x), memiliki banyak sifat yang diinginkan dari model regresi linier. Variabel independen dapat berupa kombinasi variabel kontinyu dan kategorikal.

E. Back Propagation

Algoritma *back propagation* adalah teknik yang digunakan dalam mengembangkan jaringan saraf multilayer dengan cara yang diawasi. Algoritma ini memiliki dua lintasan melalui lapisan jaringan yang berbeda: lintasan maju dan lintasan mundur. Dalam umpan maju, pola aktivitas diterapkan pada node input jaringan, dan efeknya merambat melalui lapisan jaringan demi lapisan. Akhirnya, serangkaian output dihasilkan sebagai respons aktual dari jaringan [13].
Dua fungsi aktivasi *neuron* yang biasa digunakan untuk *neuron* adalah fungsi *sigmoidal* dan tansig. Kedua fungsi secara terus-menerus dapat dibedakan dimana-mana dan biasanya memiliki bentuk matematika berikut.

Sigmoidal $f(x) = \frac{1}{1+\exp(ax)}, a > 0$
Tansig f(x) = a tanh(bx), a & b > 0

III. METODOLOGI

Metodologi penelitian yang akan digunakan dalam penelitian ini yaitu dengan menerapkan beberapa Algoritma klasifikasi naïve bayes, *support vector machine*, C.45, *logistic regression*, dan *back propagation*. Algoritma tersebut akan melakukan klasifikasi terhadap dataset penyakit jantung.

A. Dataset

Dataset yang akan digunakan yaitu *Statlog Heart Disease Dataset* yang berasal dari UCI *machine learning repository* (http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/) yang mengandung 270 *record* data. Parameter yang digunakan dengan 13 parameter seperti *age, sex, chest pain, serum cholesterol,* dan lain–lain yang berhubungan dengan penyakit jantung.

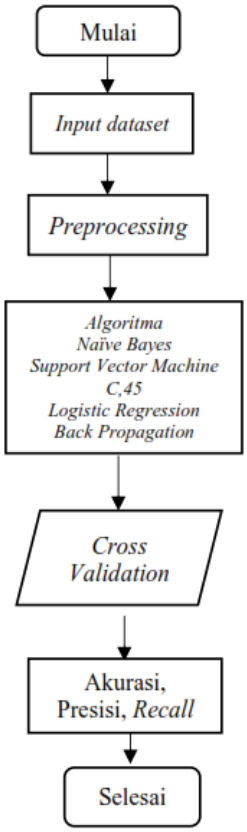
B. Validation

Tahapan ini dilakukan validasi dengan menggunakan *Cross Validation* untuk melihat performa masing–masing algoritma berdasarkan dari segi akurasi, presisi dan *recall*.

ISBN : 978-979-587-846-9
Vol.5 No.1

C. Tahapan Sistem

Secara umum sistem ini terdiri dari beberapa tahap diantaranya *preprocessing* untuk menghilangkan data–data yang *noisy* ataupun redudansi, kemudian melakukan klasifikasi dengan algoritma naïve bayes, *support vector machine*, C.45, *logistic regression,* dan *back propagation*. Setelah itu dilakukan validasi untuk melihat akurasi, presisi dan *recall* dari masing–masing algoritma guna mendapatkan algoritma dengan akurasi terbaik.



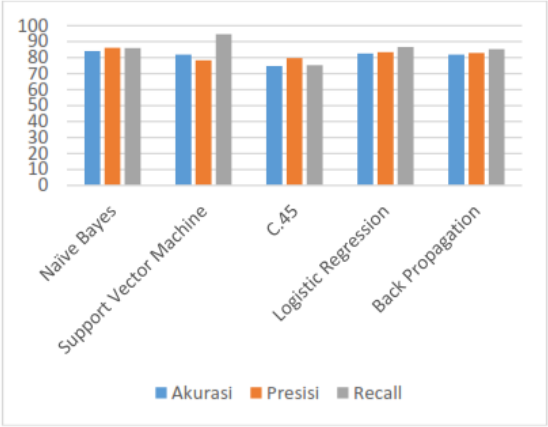
Gambar 1 *tahapan sistem*

I. HASIL

Penelitian ini dilakukan guna pengeksekusian algoritma dan pengujian validasi data. Sehingga hasil yang didapatkan terlihat pada tabel 1.

Tabel 1 akurasi, presisi, dan recall dari masing – masing algoritma

No	Algortima	Akurasi	Presisi	Recall
1	Naïve Bayes	84.07%	86.16%	86.00%
2	Support Vector Machine	81.85%	78.40%	94.67%
3	C.45	74.81%	79.73%	75.33%
4	Logistic Regression	82.59%	83.51%	86.67%
5	Back Propagation	81.85%	82.99%	85.33%



Gambar 2 akurasi, presisi, dan recall dari masing masing algoritma.

Berdasarkan *cross validation* dengan masing–masing algoritma yang ditetapkan, sehingga menghasilkan akurasi tertinggi didapat oleh algoritma naïve bayes dengan akurasi 84.07%. kemudian untuk presisi tertinggi didapat oleh algoritma naïve bayes dengan presisi 86.16%. Selanjutnya untuk *recall* tertinggi didapat oleh algoritma *support vector machine* dengan *recall* 94.67%.

I. KESIMPULAN

Penyakit jantung merupakan salah satu dari jenis PTM yang rentan menyerang terutama pria dengan usia dibawah 60 tahun. Oleh sebab itu penelitian ini berfokus untuk menyelidiki suatu algoritma, apakah memiliki tingkat akurasi yang tinggi guna pendeteksi penyakit jantung melalui objek menggunakan dataset (*heart disease*). Setelah mengeksekusi dataset dengan algoritma yang dipilih didapatlah algoritma naïve bayes dengan akurasi 84.07% mengungguli dari algoritma–algoritma lainnya.

Untuk pengembangannya, peningkatan akurasi bisa dilakukan dengan menghibridkan algoritma klasifikasi ini dengan beberapa algoritma lain, sehingga tidak menutup kemungkinan bisa menghasilkan kinerja yang berbeda dari algoritma tersebut.

ISBN : 978-979-587-846-9
Vol.5 No.1

REFERENCES

[1] Kemenkes RI. 2014. Info Datin: Situasi Kesehatan Jantung. http://www.depkes.go.id/folder/view/01/structure-publikasi-pusdatin- info-datin.html

[2] Sabransyah, M., dkk. 2017. *Aplikasi Metode Naive Bayes dalam Prediksi Risiko Penyakit Jantung.* Jurnal EKSPONENSIAL, Vol. 8 No.2, November. ISSN: 2085-7829.

[3] Bhatla, Nidhi & Jyoti, Kiran. 2012. *An Analysis of Heart Disease Prediction using Different Data Mining Techniques*. International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8, October. ISSN: 2278-0181.

[4] Liu, Xiao., et al. 2017. *A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method*. Hindawi Computational and Mathematical Methods in Medicine. Vol.2017, Article ID 8272091, Page 1-11.

[5] Dwivedi, Ashok Kumar. 2016. *Performance evaluation of different machine learning techniques for prediction of heart disease*. Neural Comput & Applic. The Natural Computing Applications Forum.

[6] Wijaya, Sugeng Hendra., et al. 2018. *Improving Classifier Performance Using Particle Swarm Optimization on Heart Disease Detection*. International Seminar on Application for Technology of Information and Communication (iSemantic).

[7] Latha, C. Beulah Christalin and Jeeva, S. Carolin. 2019. *Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques*. Informatics in Medicine Unlocked 16, 100203, 2352-9148, July.

[8] Amin, Mohammad Shafenoor., et al. 2018. *Identification of significant features and data mining techniques in predicting heart disease*. Telematics and Informatics 1191.

[9] Subbalakshmi, G., et al. 2011. *Decision Support in Heart Disease Prediction System using Naive Bayes*. Indian Journal of Computer Science and Engineering (IJCSE), Vol.2 No.2 Apr-May. ISSN: 0976-5166.

[10] Burges, Christopher J.C. 1998. *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery 2, pp.121-167.

[11] Anwar, N., et al. 2018. *Grouping the community health center patients based on the disease characteristics using C4.5 decision tree*. 1st International Conference on Engineering and Applied Technology (ICEAT), Materials Science and Engineering.

[12] Kurt, Imam., et al. 2008. *Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease*. Expert Systems with Applications Vol.34, p.366-374.

[13] Al-Milli, Nabeel. 2013. *Backpropogation Neural Network For Prediction of Heart Disease*. Journal of Theoretical and Applied Information Technology, Vol.56 No.1 p.131-135. E-ISSN: 1817-3195.