Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Penyakit Jantung

[1]Dimas Anugrah Firdlous
[1]Program Studi Pendidikan Sistem dan Teknologi Informasi
[1,]Jl. Veteran No.8, Nagri Kaler, Kec. Purwakarta, Kabupaten Purwakarta, Jawa Barat 41115
email : [1]firdlous@upi.edu

ABSTRACT
Heart disease is the disease with the largest number of deaths with 17.9 million deaths every year. Early disease mortality can be prevented by controlling disease risk factors and identifying high-risk people. Data mining is a science that can extract new information from data and become a new model and is widely used to predict an event. classification data mining technique used to predict heart disease naive bayes classification algorithm, random forest, decision tree, and support vector machine. The purpose of this study was to find the best Yahoo with the highest accuracy value to be used in predicting heart disease. the data in this study were sourced from kaggle. This research method is carried out by means of preprocessing data, classification process, evaluation of accuracy results, and comparison of the highest accuracy measures. the classification process is carried out using rapidminer. from the results of the classification carried out by the random forest algorithm, the algorithm that has the highest accuracy value is 85.7% and the lowest accuracy value uses the support vector machine algorithm with an accuracy of 68.7%. so that the random forest algorithm is the best to be used in predicting heart disease.

Keywords **-** Heart disease, data mining, classification, accuracy

ABSTRAK
Penyakit jantung merupakan penyakit dengan jumlah kematian terbesar dengan jumlah 17,9 juta kematian setiap tahunnya. Kematian dini penyakit jantung dapat dicegah dengan mengendalikan faktor risiko penyakit jantung dan mengidentifikasi orang-orang beresiko tinggi. data mining merupakan sebuah ilmu yang dapat menggali informasi baru dari sebuah data dan menjadi model baru dan banyak digunakan untuk memprediksi suatu kejadian. teknik data mining klasifikasi digunakan untuk memprediksi penyakit jantung algoritma klasifikasi *naive bayes, random forest, decision tree,*dan *support vector machine*. tujuan dari penelitian ini adalah untuk mencari algoritma terbaik dengan nilai akurasi tertinggi untuk digunakan dalam memprediksi penyakit jantung. data dalam penelitian ini bersumber dari kaggle. metode penelitian ini dilakukan dengan cara *preprocessing* data*,* proses klasifikasi, evaluasi hasil akurasi, dan komparasi ukuran akurasi tertinggi. proses klasifikasi dilakukan menggunakan rapidminer. dari hasil klasifikasi yang dilakukan algoritma *random forest* merupakan algoritma yang memiliki nilai akurasi tertinggi yaitu 85,7% dan nilai akurasi terendah menggunakan .algoritma *support vector machine* dengan akurasi 68,7%. sehingga algoritma *random forest* menjadi yang terbaik untuk digunakan dalam memprediksi penyakit jantung.

Kata Kunci **-** Penyakit jantung, data mining, klasifikasi, akurasi

1. Introduction

Penyakit jantung merupakan salah satu penyakit yang mendapatkan perhatian besar dalam dunia medis dikarenakan dampaknya kepada kesehatan manusia. penyakit kardiovaskuler merupakan penyakit yang menyebabkan kematian nomor satu di negara-negara industri, penyakit ini bukan hanya berdampak kepada kesehatan individu melainkan berdampak juga kepada kualitas hidup, biaya kesehatan, dan ekonomi negara[1]. 17,9 juta kematian disebabkan oleh penyakit jantung di dunia setiap tahunnya, 50

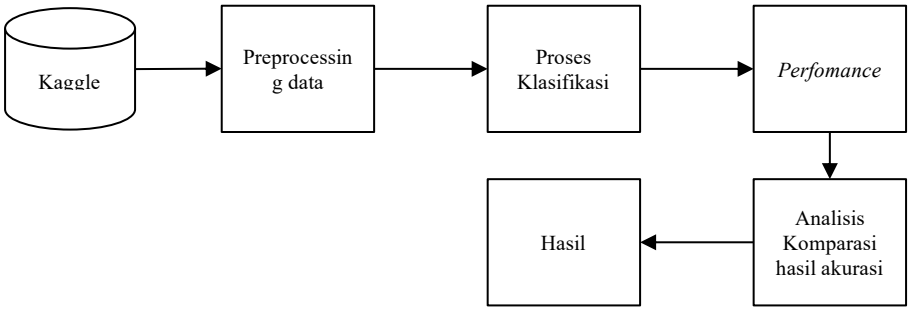
persen di antaranya dapat dicegah dengan mengendalikan faktor risiko. Penyakit jantung diharapkan menjadi alasan utama untuk 35 hingga 60 persen dari total kematian yang diperkirakan di seluruh dunia pada tahun 2025 [2]. Mengidentifikasi seseorang yang berisiko tinggi terkena penyakit jantung dan memastikan mereka menerima perawatan yang tepat dapat mencegah kematian dini. Data mining merupakan sebuah proses penggalian informasi dari sebuah data yang berjumlah besar untuk menghasilkan sebuah informasi tersembunyi di dalamnya. Teknik data mining banyak digunakan di berbagai industri untuk menemukan sebuah pola atau informasi dari industri tersebut. Penerapan data mining di industri kesehatan menjadi model baru dan banyak digunakan dengan penerapan data mining dapat dilakukan ekstraksi dan ditemukan sebuah pola tersembunyi dari data tersebut yang dapat dijadikan sebagai pendukung keputusan. Teknik data mining klasifikasi merupakan salah satu metode data mining yang tujuannya untuk memberikan prediksi dari variabel-variabel yang terkait. *Naive bayes, decision tree, random forest dan support vector machine* merupakan bagian dari algoritma klasifikasi, penggunaan algoritma klasifikasi dapat dijadikan sebagai prediksi awal untuk seorang pasien terkena pasien demi mencegah kematian dini.

Efisiensi data mining sangat bervariasi pada teknik yang digunakan dan fitur-fiturnya terpilih. Redundansi data dan inkonsistensi dalam kumpulan data mentah mempengaruhi prediksi hasil dari algoritma[3]. Pada penelitian yang menggunakan algoritma genetika, SVM dan SSVM dalam klasifikasi pasien jantung fitur telah dipilih oleh algoritma genetik untuk membantu SSVM dalam mode pemilihan input terbaik, presisi yang diperoleh adalah 72,55%, sedangkan presisi yang diperoleh GA-SSVM telah meningkatkan hasil dan presisi sama dengan 90,57%[4]. Selanjutnya penelitian mengenai prediksi penyakit jantung koroner (PJK) berdasarkan faktor risiko menggunakan jaringan syaraf tiruan backpropagation dengan 9 fitur atau faktor resiko sebagai masukan menghasilkan pola sebesar 80%[5]. kemudian penelitian untuk memprediksi penyakit jantung menggunakan algoritma K-NN dengan banyaknya tetangga 9 menghasilkan akurasi sebesar 70%[6].

Penelitian ini dilakukan dengan menggunakan data yang berasal dari kaggle, dimana penelitian ini bertujuan untuk melakukan perbandingan dari algoritma klasifikasi data mining yaitu *naive bayes, decision tree, random forest,* dan *support vector machine* untuk mencari algoritma mana yang dapat menghasilkan akurasi tertinggi dalam melakukan klasifikasi, menganalisis dan mendapatkan nilai *confusion matrix* beserta akurasi menggunakan aplikasi RapidMiner nilai akurasi tertinggi akan digunakan sebagai pendukung keputusan untuk dapat memprediksi seseorang terkena penyakit jantung dalam menghindari kematian dini.

2. Research Method

Prosedur dalam penelitian ini disajikan dalam gambar 1. yang menunjukan langkah-langkah dari penelitian ini untuk mendapatkan hasil prediksi[7].



Gambar 1. Tahapan Penelitian

Dari model penelitian diatas maka langkah-langkah penelitian yang akan dilakukan adalah sebagai berikut.

1. *Kaggle Dataset* merupakan sumber data yang digunakan pada penelitian ini yaitu *heart dataset* dengan jumlah data sebanyak 919 data dengan 11 atribut dan 1 kelas untuk dijadikan sebagai prediksi.
2. *Preprocessing* teknik *preprocessing* data yang digunakan pada penelitian ini yaitu dengan melakukan pengecekan terhadap data yang tidak valid atau hilang berdasarkan pengecekan tersebut dalam data ini tidak terdapat satupun data pada atribut yang tidak bernilai. kemudian setelah melakukan pengecekan tersebut dilakukan nilai pada kelas *heart* untuk menentukan seseorang terkena penyakit jantung atau tidak ditransformasikan nilainya yang tadinya 0 dan 1 menjadi 0 untuk no dan 1 untuk yes.
3. Proses klasifikasi pada penelitian ini dilakukan menggunakan *software rapidminer* dengan menggunakan fitur automodel untuk melakukan klasifikasi dengan algoritma *random forest, decision tree, naive bayes,* dan *support vector machine.*
4. Evaluasi *Performance* pada evaluasi *performance* klasifikasi diukur menggunakan tiga kinerja ukuran: akurasi, f-measure dan presisi. Akurasi adalah persentase dari prediksi yang benar contoh di antara semua contoh. F-measure adalah rata-rata tertimbang dari presisi dan recall. Presisi adalah persentase prediksi yang benar untuk kelas positif. kemudian Bagian ini menyampaikan beberapa hal mengenai metode yang dipergunakan untuk memecahkan solusi permasalahan. Adapun penjelasan tersebut dapat menggunakan tabel atau gambar sehingga dapat mengikuti ketentuan penulisan seperti dibawah ini.
5. Komparasi hasil pada penelitian ini analisis komparasi hasil dilakukan dengan melakukan identifikasi untuk mencari algoritma mana yang memiliki nilai akurasi tertinggi.
6. Hasil pada penelitian ini hasil yang diperoleh untuk memprediksi penyakit jantung diambil dari hasil komparasi nilai akurasi dimana algoritma dengan nilai akurasi tertinggi dijadikan prediksi terbaik untuk penyakit jantung.

3. Result and Analysis

Pada penelitian ini digunakan perangkat keras dengan spesifikasi RAM 8GB Hardisk 1TB dengan prosesor intel i5-8250U. Proses klasifikasi dilakukan melalui auto model pada *software RapidMiner* Tujuan utama dari penelitian ini adalah untuk mencari nilai akurasi tertinggi pada algoritma klasifikasi yang digunakan untuk memprediksi penyakit jantung. tabel 1. menunjukan gambaran singkat mengenai deskripsi atribut dataset yang digunakan dalam penelitian ini.

3.1 Data Set

Dataset yang digunakan dalam penelitian ini memiliki jumlah data sebanyak 919 data dengan 12 atribut. tabel 1 akan mendeskripsikan mengenai atribut yang digunakan dalam penelitian.

Tabel 1. Deskripsi atribut

Atribut	Nilai
Age	28-77 Tahun
Sex	M untuk *male*, F untuk *female*
ChestPainType	TA: *Typical Angina,* ATA*: Atypical Angina,* NAP*: Non-Anginal Pain,* ASY*: Asymptomatic*
RestingBP	Tekanan darah mmhg
cholesterol	Serum Kolesterol mm/dl

RestingBS	Gula darah (1: jika Puasa BS > 120 mg/dl, 0: sebaliknya)
RestingECG	hasil elektrokardiogram istirahat (Normal: Normal, ST: memiliki kelainan gelombang ST-T (inversi gelombang T dan/atau elevasi atau depresi ST > 0,05 mV), LVH: menunjukkan kemungkinan atau pasti hipertrofi ventrikel kiri menurut kriteria Estes)
MaxHR	detak jantung maksimum tercapai (Nilai numerik antara 60 dan 202)
ExerciseAngina	angina akibat olahraga (Y: Ya, N: Tidak)
Oldpeak	ST Nilai numerik diukur dalam depresi
ST_Slope	kemiringan puncak latihan segmen ST (Up: upsloping, Flat: flat, Down: downsloping)
HeartDisease	yes : penyakit jantung, no: tidak penyakit jantung

3.2 *Confusion Matrix*

Perhitungan evaluasi performa klasifikasi dilakukan dengan *confusion matrix.* Nilai *confusion matrix* dari masing-masing algoritma ditunjukan oleh tabel 2.

Tabel 2. *Confusion Matrix*

Algoritma	**TP**	**FP**	**TN**	**FN**
Naive Bayes	128	19	93	22
Decision Tree	133	8	56	65
Random Forest	130	17	96	20
SVM	101	46	79	36

TP merupakan data yang positif yaitu yes (memiliki penyakit jantung) yang terklasifikasi sebagai yes, FP merupakan data yes namun terklasifikasi no oleh sistem, kemudian TN merupakan no orang yang tidak memiliki penyakit jantung dan terklasifikasi no, dan FN adalah dat no yang terklasifikasi yes oleh sistem.

3.3 Pengukuran Akurasi

Algoritma *naive bayes, random forest,* dan *support vector machine* digunakan dalam penelitian ini, nilai *precision, recall,* akurasi*,* dan *F Measure* digunakan sebagai hasil klasifikasi penelitian ini. dimana nilai akurasi akan dipakai sebagai hasil untuk memprediksi penyakit jantung. Tabel 3. akan menggambarkan hasil ukuran akurasi.

Tabel 3. Ukuran Akurasi

Algoritma	**Akurasi**	***Recall***	***Precision***	***F Measure***
Naive Bayes	84,4%	87,1%	85,3%	86,2%
Decision Tree	72,1%	94,5%	67,5%	78,5%
Random Forest	85,9%	88,3%	86,7%	87,5%
SVM	68,7%	68,5%	73,9%	71%

Dari data tabel 3. diatas dapat dilakukan evaluasi dan identifikasi mengenai algoritma mana yang memiliki nilai akurasi paling tinggi dalam penelitian ini ditemukan bahwa klasifikasi dengan menggunakan algoritma *random forest* memiliki ukuran akurasi paling tinggi yaitu sebesar 85,9%.

Efisiensi data mining tergantung kepada algoritma mana yang digunakan dan fitur-fitur atau atribut apa saja yang dijadikan sebagai *input* dalam melakukan proses data mining[3] dalam penelitian ini ditemukan bahwa dengan data yang memiliki jumlah atribut 12 seperti pada tabel1. menghasilkan algoritma *random forest* sebagai algoritma dengan ukuran akurasi tertinggi sehingga dalam langkah pertama untuk memprediksi penyakit jantung berdasarkan pada temuan penelitian adalah dengan menggunakan hasil klasifikasi dari algoritma *random forest* yang memiliki akurasi sebesar 85,9% kemudian di urutan kedua menggunakan *naive bayes* dengan akurasi 84,4%, kemudian menggunakan *decision tree* dengan akurasi sebesar 72,1%, dan terakhir adalah menggunakan *support vector machine* dengan akurasi 68,7%.

4. Conclusion

Penyakit jantung merupakan penyakit dengan jumlah kematian terbesar di dunia, salah satu cara untuk mencegah terjadinya kematian dini oleh penyakit jantung adalah dengan melakukan identifikasi terhadap orang yang memiliki resiko tinggi. Kontribusi atau tujuan utama dari penelitian ini adalah untuk mencari algoritma klasifikasi terbaik dengan nilai akurasi tertinggi untuk memprediksi penyakit jantung untuk mencegah kematian dini. Percobaan pada penelitian ini menghasilkan bahwa algoritma *random forest* menjadi algoritma yang paling direkomendasikan untuk melakukan prediksi penyakit jantung dengan akurasi sebesar 85,7%. Selanjutnya dapat dilakukan penelitian dengan menggunakan algoritma lain untuk memprediksi penyakit jantung dan penyakit lainnya. Kemudian dapat juga dilakukan penelitian dengan melakukan pemisahan data menjadi data training dan data testing secara manual menggunakan persentase 80% data training dan 20% data latih sebagai perbandingan klasifikasi yang dihasilkan nantinya.

References

[1] Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I., & Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. International Journal of Electrical & Computer Engineering.

[2] Bahrami, B., & Shirvani, M. H. (2015). Prediction and diagnosis of heart disease by data mining techniques. Journal of Multidisciplinary Engineering Science and Technology (JMEST), 2(2), 164-168.

[3] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics.

[4] Sumit B, Praveen P, G.N. Pillai. (2008). SVM Based Decision Support System for Heart Disease Classification with Integer- Coded Genetic Algorithm to Select Critical Features. WCECS. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA.

[5] Effendy, N., Subagja, S., & Faisal, A. Prediksi penyakit jantung koroner (PJK) berdasarkan faktor risiko menggunakan jaringan syaraf tiruan backpropagation. In Seminar Nasional Aplikasi Teknologi Informasi (SNATI).

[6] Lestari, M. E. I. (2015). Penerapan algoritma Klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung. Faktor Exacta, 7(4), 366-371.

[7] Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi*

[8] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*.

[9] Permana, D. S., & Silvanie, A. (2021). PREDIKSI PENYAKIT JANTUNG MENGGUNAKAN SUPPORT VECTOR MACHINE DAN PYTHON PADA BASIS DATA PASIEN DI CLEVELAND. *JUNIF: Jurnal Nasional Informatika*.

[10] Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications*. IEEE.

[11] Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In 2012 Japan-Egypt Conference on Electronics, Communications and Computers.IEEE.
[12] Rifai, B. (2013). Algoritma Neural Network Untuk Prediksi Penyakit Jantung. *Jurnal Techno Nusa Mandiri*.
[13] Septiani, W. D. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis. *Jurnal Pilar Nusa Mandiri*.
[14] Andayani, S., & Astuti, Y. (2017). Prediksi Kejadian Penyakit Tuberkulosis Paru Berdasarkan Usia di Kabupaten Ponorogo Tahun 2016-2020. *Indonesian Journal for Health Sciences*.