PERBANDINGAN MODEL ATURAN CN2 *RULE IDUCER* DAN POHON KEPUTUSAN C4.5 TERHADAP DATA PENYAKIT JANTUNG KORONER

[1]Anis Fitri Nur Masruriyah
[2] Hilda Yulia Novita

Teknik Informatika, Fakultas Ilmu Komputer, Universitas Buana Perjuangan Karawang
anis.masruriyah@ubpkarawang.ac.id[1,] hilda.yulia@ubpkarawang.ac.id [2]

ABSTRAK

Penyakit jantung koroner (PJK) menjadi peringkat tertinggi penyebab kematian di Indonesia khususnya pada usia-usia produktif. Terlebih di masa pandemi COVID-19 pasien yang memiliki komorbid dengan PJK memiliki risiko keselamatan yang mengakibatkan perburukan bahkan kematian. Maka pasien pada usia produktif yang memiliki komorbid PJK akan terancam nyawanya. Penelitian ini hendak melakukan perbandingan model algoritma menerapkan teknik ekstraksi fitur untuk mengetahui variabel yang paling mempengaruhi penyakit jantung berdasarkan komputasi.

Kata kunci: data mining, penyakit jantung koroner, prediksi

PENDAHULUAN

Berdasarkan data yang dihimpun oleh Kementrian Kesehatan Republik Indonesia (2021), PJK salah satu penyakit yang meningkat setiap tahunnya, sekaligus menempati peringkat tertinggi penyebab kematian di Indonesia terutama terjadi pada usia-usia produktif. Dijelaskan bahwa perubahan gaya hidup yang tidak sehat dan pola makan yang tidak seimbang menyebabkan prevalensi PJK semakin tinggi. Penelitian Mezzatesta et al. (2019) menunjukkan bahwa memprediksi tingkat kematian pada pasien PJK yang mengidap dialisis yang menerapkan algoritma algoritma Support Vector Machine dengan kernel Radial Basis Function yang dioptimasi menggunakan GridSearch. Hasil model prediksi yang diuji memperoleh akurasi sebesar 95.25%. Selanjutnya, Ghosh et al. (2021) menerapkan algoritma ekstraksi fitur RELIEF dan LASSO pada kasus kardiovaskular. Penelitian tersebut memanfaatkan machine learning untuk proses prediksi penyakit kardiovaskular dan mendapatkan akurasi mencapai 99.05%. Kemudian, penelitian yang dilakukan oleh El-Hasnony et al. (2022) membuat model pencegahan penyakit stroke dan jantung. Active learning diterapkan pada penelitian tersebut untuk mengetahui faktor yang paling berpengaruh pada penyakit stroke dan jantung.

Sehingga, berdasarkan faktor yang berpengaruh tersebut, para tenaga medis dapat melakukan perawatan yang tepat dalam pencegahan penyakit stroke dan jantung.

Penelitian lain juga menjelaskan dan membuktikan bahwa model prediksi mampu membantu tenaga medis untuk menyelesaikan masalah-masalah kesehatan (Maniruzzaman et al. 2017; Nilashi et al. 2017; Masruriyah et al. 2019). Di sisi lain, pada bidang selain kesehatan model prediksi juga terbukti kompeten (Morota et al. 2018; Deo et al. 2020; Wang et al. 2020).

Maka dari itu, penelitian ini hendak mengetahui dan membangun model terbaik untuk kasus penyakit jantung dengan menerapkan teknik ektraksi fitur serta tanpa ekstraksi fitur untuk mengevaluasi kinerja model prediksi penyakit jantung.

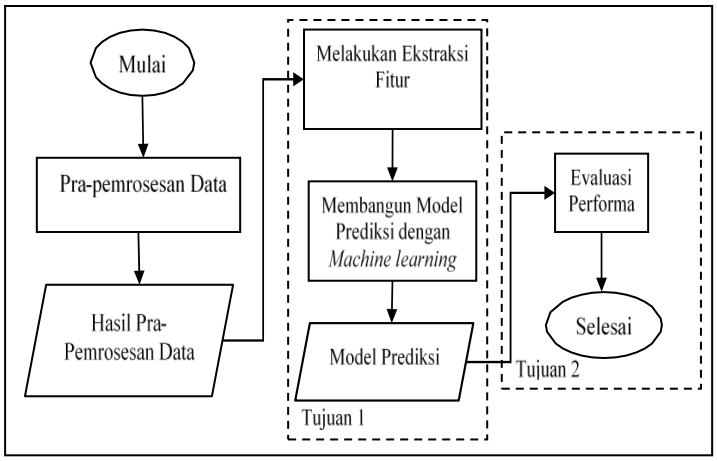
METODE PENELITIAN

Waktu dan Tempat Penelitian

Penilitian dilakuan di Laboratorium Riset, Fakultas Ilmu Komputer, UBP Karawang pada Januari 2022 hingga Oktober 2022.

Prosedur Penelitian

Proses penelitian ditunjukkan pada gambar 1 yang menggunakan empat tahap analitika (*data quality analytics*, *descpritive analytics*, *diagnostic analytics* dan *predictive analytics*).



Gambar 1 Alur Prosedur Penelitian

Secara umum, proses analisis data dimulai dengan prapemrosesan, ekstraksi fitur hingga dihasilkan pengetahuan baru. Pada tahap pertama pra-pemrosesan data sudah termasuk data *quality analytics* dan *descpritive analytics*. Selanjutnya hasil pra-pemrosesan data diolah untuk mendapatkan hasil *diagnostic analytics* dan *predictive analytics*.

Data, Instrumen, dan Teknik Pengumpulan Data

Data dikumpulkan dari catatan medis yang telah dietujui oleh Organisasi Kesehatan Dunia dan dapat diakses pada (Service 2020). Atribut pada data dan penjelasannya ditunjukkan pada Tabel 1 berikut.

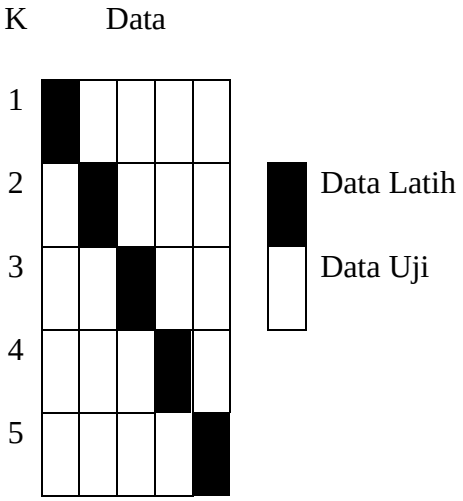
Tabel 1 Atribut Data Penelitian

Atribut	**Keterangan**
Age	: Usia Pasien dengan satuan tahun
Sex	: Jenis kelamin pasien dengan F sebagai perempuan dan M sebagai pria
ChestPainType	: Nyeri dada yang diderita pasien dengan tipe [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	: Tekanan darah yang dihitung saat sedang istirahat dalam satuan mm/Hg
Cholesterol	: Jumlah kolesterol pasien dalam satuan mm/dl
FastingBS	: Jumlah gula darah pasien saat puasa dengan kondisi jika nilainya 1 FastingBS > 120 mg/dl, dan 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 tercapai maksimum [Nilai numerik antara 60 dan 202]
ExerciseAngina	: angina yang diinduksi oleh 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	: Hasil berdasarkan variabel independent dengan kriteria 1 sebagai pengidap penyakit jantung dan 0 sehat atau normal

Teknik Analisis Data

Data dianalisis dengan teknik komputasi yang menerapkan kaidah statistik pada data mining. Pada tahap pertama pra-pemrosesan data sudah termasuk data quality analytics dan descpritive analytics. Selanjutnya hasil pra-pemrosesan data diolah untuk mendapatkan hasil diagnostic analytics dan predictive analytics. Beberapa teknik klasifikasi digunakan untuk membuat model prediksi. Selanjutnya, Evaluasi kinerja dilakukan terhadap model dengan tujuan untuk mengetahui seberapa baik kinerja model dengan menggunakan data uji. Evaluasi didasarkan pada akurasi, presisi, sensitivitas, dan spesifisitas (Balali dan Golparvar-Fard 2015). Akurasi adalah tingkat kedekatan pengukuran kuantitas dengan nilai sebenarnya dari kuantitas, presisi,

sensitivitas adalah bagian dari sampel relevan yang diambil, dan spesifisitas menghitung proporsi positif aktual yang diidentifikasi dengan benar. Akurasi, sensitivitas dan spesifisitas dihitung berdasarkan matriks konfusi. Evaluasi kinerja pada penelitian ini menggunakan teknik K-Fold Cross Validation. Cara kerja dari Teknik K-Fold Cross Validation adalah dengan membagi data menjadi data uji dan data latih sebanyak K. Ilustrasi untuk teknik ini ditunjukkan pada Gambar 2.



Gambar 2 Ilustrasi K-Fold

HASIL PENELITIAN DAN PEMBAHASAN

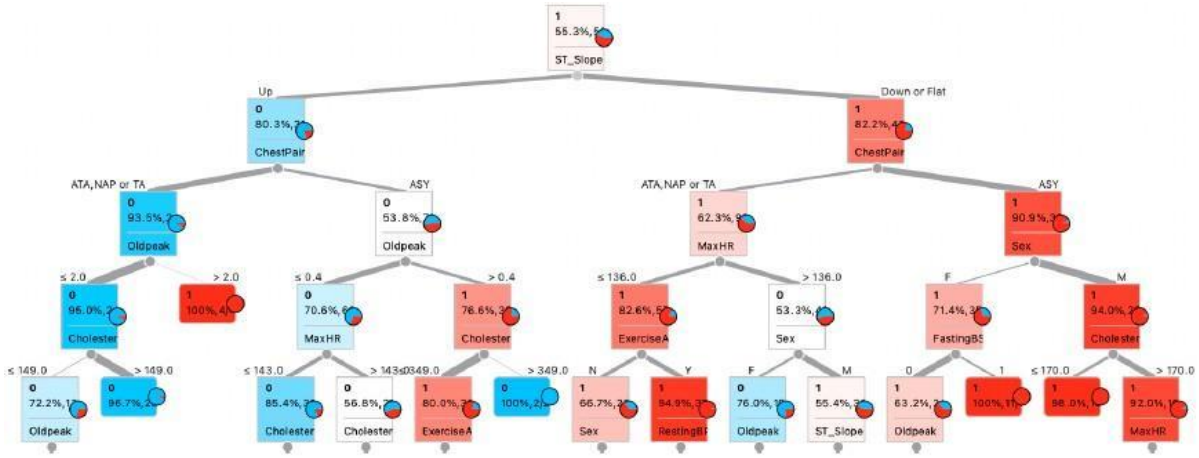
Hasil Penelitian

Tahapan prapemrosesan data pada penelitian ini dilakukan agar data lebih ideal untuk digunakan pada tahap selanjutnya, Jenis prapemrosesan data yang dilakukan antara lain menghapus data dengan komponen yang tidak lengkap untuk menghindari manipulasi pengisian data lebih jauh karena data yang digunakan lebih dari 1000. Selanjutnya setelah data dengan komponen tidak lengkap telah dihapus maka dilakukan normalisasi pada data-data yang memiliki lebih dari 4 kategori. Hal ini dilakukan untuk mengurangi perulangan dan memetakan kejadian yang serupa.

Selanjutnya, ektraksi fitur pada algoritma C4.5 dan CN2 Rule Inducer pemilihan fitur berpengaruh dilakukan ketika menentukan atribut yang menjadi akar dan cabang pada pohon keputusan. Berdasarkan hasil perhitungan menggunakan persamaan 1 hingga Persamaan 3 pada algoritma C4.5, 5 atribut yang mempengaruhi penyakit jantung antara lain ST Slope, Chest Pain, Oldpeak, Jenis Kelamin dan Detak Jantung Maksimum. Selanjutnya, pada CN2 Rule Inducer 5 atribut yang berpengaruh terhadap penyakit jantung yaitu ST Slope, Chest Pain, Jenis kelamin, Detak Jantung Maksimum dan ECG Resting. Berdasarkan hasil ekstraksi fitur, ST Slope dan Chest Pain menjadi fitur paling berpengaruh pada dua algoritma yang digunakan.

Pembahasan

Model prediksi dimulai dengan membagi seluruh data menjadi data latih dan data uji dengan K-Fold cross validation. Data latih menjadi basis pembangunan model prediksi dalam penelitian ini. Model yang dihasilkan dengan menjalankan Persamaan 1 hingga Persamaan 4 pada algoritma C4.5 mencapai akurasi prediksi sebesar 81.9%. Di sisi lain, algoritma CN2 berhasil membangun model prediksi dengan nilai akurasi lebih rendah dibanding dengan C4.5 yaitu sebesar 77.3%. Hasil model pohon keputusan ditunjukkan pada Gambar 4.1. Di mana, ST Slope dan Chest Pain sebagai akar dan dahan utama dari pohon keputusan yang dibangun.



Gambar 3 Model Pohon Keputusan C4.5

Selanjutnya model aturan yang dibangun menggunakan Algoritma C2 ditunjukkan pada Tabel 2 Model aturan yang ditampilkan adalah 10 model tertinggi dengan tingkat kemunculan yang dominan.Proses evaluasi menggunakan data uji yang telah dipisahkan pada proses sebelumnya dan hasil evaluasi menggunakan matriks konfusi yang ditunjukkan pada Tabel 2.

Tabel 2 Matriks Konfusi C4.5 dan CN2 Ruler

		C4.5			CN2 Ruler	
		Prediksi			Prediksi	
		1	0		1	0
Aktual	1	419	89		407	101
	0	77	333	0	107	303

Menggunakan matriks konfusi dari Tabel 2, diperoleh nilai akurasi untuk C4.5 sebesar 81.9% dan 77.3% untuk CN2. Artinya, model algoritma mampu memetakan data dengan kelas yang

sesuai cukup baik. Kemudian, masih dengan matriks yang sama evaluasi presisi C4.5 terhadap model yang dibangun sebesar 82% dan CN2 77.3%. Sehingga, model algoritma terbukti memiliki kemampuan yang cukup untuk memprediksi data pada kelas yang tepat. Terakhir adalah *recall* yang diperoleh dari matriks konfusi yang sama, pada C4.5 diperoleh nilai *recall* sebesar 81.9% dan CN2 77.3%.

KESIMPULAN DAN IMPLIKASI

Berdasarkan hasil dapat disimpulkan bahwa model algoritma telah berhasil dibangun dengan informasi dari C4.5 atribut yang berpengaruh pada penyakit jantung adalah ST Slope, *Chest Pain*, *Oldpeak*, Jenis Kelamin dan Detak Jantung Maksimum. Kemudian, pada CN2 *Rule Inducer* 5 atribut yang berpengaruh yaitu ST Slope, *Chest Pain*, Jenis kelamin, Detak Jantung Maksimum dan ECG *Resting*. Evaluasi model juga telah berhasil dilakukan dengan nilai akurasi mencapai 81.9% untuk C4.5 dan 77.3% untuk CN2 *ruler inducer*.

selanjutnya, disarankan melakukan optimasi pada algoritma CN2 ruler, sehingga akurasi dapat mencapai lebih dari 80%. Karena suatu model algoritma dikatakan efektif jika sudah melampaui 80%.

DAFTAR PUSTAKA

Alkhusari, Handayani M, Saputra MAS, Romadhon M. 2020. Analisis Kejadian Penyakit Jantung Koroner Di Poliklinik Jantung. J ‘Aisyiyah Med. 5:99–110.

Barhate A, Gupta S, Kinage S, Parvatikar P. 2018. Study of Data Mining Concepts. Int J New Innov Eng Technol. 9(1):30–34.

Cherfi A, Nouira K, Ferchichi A. 2018. Very Fast C4.5 Decision Tree Algorithm. Appl Artif Intell. 32(2):119–137. doi:10.1080/08839514.2018.1447479.

Deo GS, Mishra A, Jalaluddin ZM, Mahamuni CV. 2020. Predictive Analysis of Resource Usage Data in Academic Libraries using the VADER Sentiment Algorithm. Proc - 2020 12th Int Conf Comput Intell Commun Networks, CICN 2020. July:221–228. doi:10.1109/CICN49253.2020.9242575.

Djatna T, Hardhienata MKD, Masruriyah AFN. 2018. An intuitionistic fuzzy diagnosis analytics for stroke disease. J Big Data. 5(1). doi:10.1186/s40537-018-0142-7.

El-Hasnony IM, Elzeki OM, Alshehri A, Salem H. 2022. Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. Sensors. 22(3). doi:10.3390/s22031184.

Ghosh A, Sufian A, Sultana F, Chakrabarti A, De D. 2019. Fundamental concepts of convolutional neural network. Volume ke-172. Ghosh P, Azam S, Jonkman M, Karim A, Shamrat FMJM, Ignatious E, Shultana S, Beeravolu AR, De Boer F. 2021. Efficient

prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. IEEE Access. 9:19304–19326. doi:10.1109/ACCESS.2021.3053759.

Kementrian Kesehatan Republik Indonesia. 2021. Penyakit Jantung Koroner Didominasi Masyarakat Kota. [diakses 2022 Jan 20]. https://www.kemkes.go.id/article/view/21093000002/penyakit-jantungkoroner-didominasi-masyarakat-kota.html.

Maniruzzaman M, Kumar N, Menhazul Abedin M, Shaykhul Islam M, Suri HS, El-Baz AS, Suri JS. 2017. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed. 152:23–34. doi:10.1016/j.cmpb.2017.09.004.

Masruriyah AFN, Djatna T, Dewi Hardhienata MK, Handayani HH, Wahiddin D. 2019. Predictive Analytics For Stroke Disease. Di dalam: Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019.

Mezzatesta S, Torino C, De Meo P, Fiumara G, Vilasi A. 2019. A machine learningbased approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Comput Methods Programs Biomed. 177:9–15. doi:10.1016/j.cmpb.2019.05.005.

Morota G, Ventura R V., Silva FF, Koyama M, Fernando SC. 2018. Big data analytics and precision animal agriculture symposium: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. J Anim Sci. 96(4):1540–1550. doi:10.1093/jas/sky014.

Nilashi M, Ibrahim O, Dalvi M, Ahmadi H, Shahmoradi L. 2017. Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset. Fuzzy Inf Eng. 9(3):345–357. doi:10.1016/j.fiae.2017.09.006.

Pradono J, Werdhasari A. 2018. Faktor Determinan Penyakit Jantung Koroner pada Kelompok Umur 25-65 tahun di Kota Bogor, Data Kohor 2011-2012. Bul Penelit Kesehat. 46(1):23–34. doi:10.22435/bpk.v46i1.48.

Service D of H and H. 2020. Behavioral Risk Factor Surveillance System. [diakses 2022 Feb 2]. https://www.cdc.gov/brfss/annual_data/annual_2020.html.

Sullivan W. 2017. Machine Learning For Beginners Guide Algorithms. Volume ke-4.

Wang F, Li M, Mei Y, Li W. 2020. Time Series Data Mining: A Case Study with Big Data Analytics Approach. *IEEE Access*. 8:14322–14328. doi:10.1109/ACCESS.2020.2966553.

Wang L. 2017. Data Mining, Machine Learning and Big Data Analytics. Int Trans Electr Comput Eng Syst. 4(2):55–61. doi:10.12691/iteces-4-2-2.