Journals of Ners Community

Volume 13, Nomor 5, November 2022
Hal. 489 - 495

PERBANDINGAN METODE ALGORITMA DECISION TREE C4.5 DAN NAÏVE BAYES UNTUK MEMPREDIKSI PENYAKIT TIROID

Leli Safitri, Krista Cahayani
Murtiwiyati, Siti Chodidjah, Deasy Indayanti
Universitas Gunadarma, Jawa Barat, Indonesia
Email: leli.s@staff.gunadarma.ac.id, kristacahayani20@gmail.com
murtiwiyati@staff.gunadarma.ac.id, chodi@staff.gunadarma.ac.id,
deasy@staff.gunadarma.ac.id

ABSTRAK

Penyakit tiroid adalah kelenjar endokrin murni terbesar di tubuh manusia, terletak di leher bagian depan. Gangguan fungsi tiroid seringkali sulit dikenali karena gejalanya tidak spesifik, dan sering diabaikan karena gejala penyakit tiroid sangat mirip dengan banyak penyakit gaya hidup modern. pasien seringkali tidak menyadari ada masalah pada dirinya dan tidak memeriksakan diri ke dokter. Oleh karena itu, penelitian dibidang kesehatan dilakukan untuk pengobatan lebih dini, guna mencegah kematian akibat terlambatnya penanganan. Penelitian ini menggunakan metode klasifikasi data mining Algoritma Decision Tree C4.5 dan Naïve Bayes dengan tujuan agar algoritma terpilih merupakan algoritma yang menghasilkan nilai akurasi dan nilai Area Under Curve (AUC) yang lebih baik. Data penelitian menggunakan Thyroid Disease Dataset UCI (University of California, Irvine) Machine Learning Repository. Hasil pengujian menunjukkan bahwa akurasi lebih baik diperoleh dari Algoritma Decision Tree C4.5 sebesar 97,12% sedangkan nilai akurasi Algoritma Naïve Bayes sebesar 76,02%. Nilai Area Under Curve (AUC) pada kurva Receiver Operating Characteristic (ROC) menunjukkan Algoritma Decision Tree C4.5 memiliki nilai lebih tinggi dari Algoritma Naïve Bayes dengan hasil klasifikasi Good Classification.

Kata kunci : Algoritma Decision Tree C4.5, Algoritma Naïve Bayes, Data Mining, Klasifikasi, Perbandingan , Penyakit Tiroid.

ABSTRACT

Thyroid disease is the largest pure endocrine gland in the human body, located in the front of the neck. Disorders of thyroid function are often difficult to identify because the symptoms are non-specific, and are often overlooked because the symptoms of thyroid disease are very similar to many of the diseases of modern lifestyles. Patients often do not realize there is a problem with themselves and do not see a doctor. Therefore, research in the health sector is carried out for early treatment, in order to prevent death due to delay in treatment. This study uses data mining classification methods, Decision Tree C4.5 Algorithm and Naïve Bayes with the aim that the selected algorithm is an algorithm that produces better accuracy and Area Under Curve (AUC) values. Research data using the Thyroid Disease Dataset UCI (University of California, Irvine) Machine Learning Repository. The test results show that better accuracy is obtained from the Decision Tree C4.5 Algorithm of 97.12% while the accuracy value of the Naïve Bayes Algorithm is 76.02%. The Area Under Curve (AUC) value on the Receiver Operating Characteristic (ROC) curve shows the Decision Tree C4.5 Algorithm has a higher value than the Naïve Bayes Algorithm with Good Classification results.

Keywords: *Decision Tree C4.5 Algorithm, Naïve Bayes Algorithm, Data Mining, Classification, Comparison, Thyroid Disease.*

PENDAHULUAN

Pentingnya mengetahui gejala suatu penyakit merupakan Langkah awal untuk

mengantisipasi timbulnya penyakit yang dapat membahayakan kesehatan, bahkan berujung pada kematian. Tiroid merupakan merupakan kelenjar endokrin murni terbesar dalam tubuh manusia yang terletak di leher bagian depan. Faktor risiko penyakit atau gangguan tiroid dapat dipengaruhi oleh berbagai faktor, seperti usia di atas 60 tahun maka semakin berisiko terjadinya hipotiroid atau hipertiroid. Jenis kelamin perempuan biasanya lebih berisiko terjadi gangguan tiroid. (Kementrian Kesehatan, 2015).

Berdasarkan data Globocan tahun 2020, kanker tiroid merupakan kanker peringkat 5 tertinggi yang diderita oleh perempuan di seluruh dunia. Kanker tiroid menempati urutan ke-9 dari 10 kanker terbanyak di Indonesia, serta angkanya mengalami peningkatan tiap tahunnya. Gangguan fungsi tiroid seringkali sulit diidentifikasi karena gejalanya tidak spesifik, gejala gangguan tiroid sangat mirip dengan berbagai keluhan akibat gaya hidup modern sehingga sangat sering diabaikan. Akibatnya pasien seringkali tidak menyadari ada masalah pada dirinya dan tidak memeriksakan diri ke dokter (Dr. Dr. EM Yunir SpPD-KEMD, 2017). Oleh karena itu, perlu adanya penelitian yang menerapkan metode untuk memprediksi penyakit tiroid yang nantinya akan mempermudah pasien dalam memprediksi penyakit tersebut.

Salah satu metode yang dapat digunakan untuk memprediksi penyakit adalah dengan menggunakan data mining. Dalam penelitian ini data mining yang digunakan adalah teknik klasifikasi. Sedangkan metode atau algoritma yang digunakan dalam penelitian ini adalah algoritma Decision Tree C4.5 dan Naïve Bayes.

Banyak penelitian mengenai penyakit tiroid dengan menggunakan metode data mining, diantaranya penelitian yang dilakukan oleh Bambang Wijonarko. Menurut jurnal Bambang Wijonarko yang berjudul Perbandingan Algoritma Data Mining Naïve Bayes dan Bayes Network untuk Mengidentifikasi Penyakit Tiroid. Dalam penelitian ini melakukan komparasi algoritma diantaranya Naïve Bayes dan Bayes Network yang menyatakan bahwa Bayes Network memiliki Akurasi yang lebih tinggi dibandingan Naïve Bayes (Wijonarko, 2018).

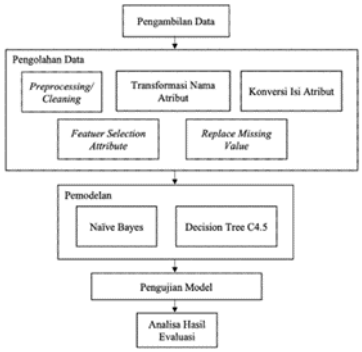
Penelitian yang lain dilakukan oleh Sarifah Agustina, Ali Mustopa, Andi Saryoko, Windu Gata, dan Siti Khotimatul Wildah (2020) mengenai Penerapan Algoritma J48 untuk Deteksi Penyakit Tiroid. Dalam penelitian tersebut diperoleh nilai akurasi sebesar 99.645%. Penelitian yang dilakukan oleh Umar Sidiq, Dr. Syed Mutahar Aaqib, dan Dr. Rafi Ahmad Khan (2019) mengenai Diagnosis Berbagai Penyakit Tiroid Menggunakan Data Mining Teknik Klasifikasi. Dalam penelitian tersebut dinyatakan bahwa Decision Tree memperoleh akurasi yang lebih tinggi daripada algoritma lainnya (Sidiq, Aaqib, dan Khan, 2019).

Selain itu, penelitian lain dilakukan oleh Khalid Salman, Emrullah Sonuc di Universitas Karabuk, Turki (2021) mengenai "Thyroid Disease Classification Using Machine Learning Algorithms". Dalam penelitian ini melakukan komparasi algoritma diantaranya Support Vector Machines, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, Multi-Layer Perceptron's, dan Linear Discriminant Analysis. Dalam penelitian tersebut menunjukkan bahwa terdapat dua algoritma terbaik yang digunakan yaitu Random Forest dengan akurasi 98,93% dan Multi-Layer Player dengan akurasi 96,4% diikuti algoritma terbaik ketiga yaitu Naïve Bayes dengan akurasi 90,67%.

Berdasarkan latar belakang yang telah diuraikan di atas, maka dilakukan penelitian dengan menggunakan teknik data mining untuk melakukan pengolahan dataset penyakit tiroid dengan menggunakan metode Decision Tree C4.5 kemudian dibandingkan dengan metode data mining lainnya yang memiliki kemampuan sama baik dengan Decision Tree C4.5 yaitu Naïve Bayes.

METODE PENELITIAN

Metode Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1 sebagai berikut:



Gambar 1. Tahapan Penelitian

Pengambilan data, penelitian ini menggunakan data yang diambil dari website UCI (Universitas California Invene) Machine Learning Repository. Dataset tersebut berisi pasien yang terkena penyakit tiroid sebanyak 3711 pasien.

Pengolahan data, pada tahap ini dilakukan pengolahan data tiroid, mulai dari transformasi dengan mengubah sebagian nama atribut- atribut, membuang duplikasi yang ada pada data, menangani nilai yang hilang dan memperbaiki kesalahan pada data.

Pemodelan, pada tahap ini dilakukan pengujian model dengan menggunakan dataset penyakit tiroid. Algoritma yang dipakai pada tahap ini adalah Decision Tree C4.5 dan Naïve Bayes. Pada tahap pemodelan ini dilakukan pemrosesan data training sehingga akan membahas metode algoritma yang diujikan, kemudian dianalisa dan dikomparasi. Pemodelan dilakukan dengan menggunakan software RapidMiner yang merupakan salah satu software pemodelan terbaik untuk data mining.

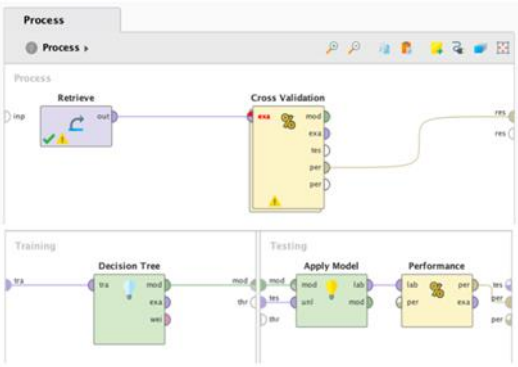
Pengujian model, setelah melakukan perancangan dan pemodelan dari kedua model Algoritma Decision Tree C4.5 dan Naïve Bayes, maka dapat diuji tingkat akurasinya untuk melihat kinerja dari kedua model tersebut. Pengujian tingkat Accuracy, Precision, Recall menggunakan confusion matrix dan kurva ROC/AUC (Area Under Cover). Analisa hasil evaluasi, dilakukan analisa terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

HASIL DAN PEMBAHASAN

A. Penerapan Model Algoritma

Pada tahap ini dilakukan pengujian model dengan menggunakan dataset penyakit tiroid. Algoritma yang dipakai pada tahap ini adalah Decision Tree C4.5 dan Naïve Bayes. Pada tahap pemodelan ini dilakukan pemrosesan data training sehingga akan membahas metode algoritma yang diujikan, kemudian dianalisa dan dikomparasi. Pemodelan dilakukan dengan menggunakan software RapidMiner yang merupakan salah satu software pemodelan terbaik untuk data mining. Proses dalam tahapan ini antara lain dengan validasi menggunakan ten-folds cross validation, proses ini berguna untuk membagi dataset menjadi 10 bagian secara acak lalu 9 bagian digunakan untuk data training sedangkan 1 bagian digunakan untuk data testing. Proses ini berulang sampai dengan 10 kali sampai dengan seluruh data mendapatkan porsinya atau mendapat giliran menjadi data testing.

Didalam proses validasi dilakukan perhitungan tingkat akurasi dengan menggunakan confussion matrix. Tabel ini digunakan untuk mengukur tingkat akurasi dari algoritma. Dalam 10 kali percobaan keseluruhan tingkat akurasi dihitung rerata untuk mendapatkan tingkat akurasi dari algoritma tertentu. Gambar 2 merupakan hasil dari perhitungan algoritma Decision Tree C4.5 dengan menggunakan Rapid Miner.



Gambar 2. Pemodelan Algoritma Decision Tree C4.5

Untuk perhitungan algoritma Naive Bayes dilakukan dengan tahapan yang sama dengan perhitungan Decision Tree C4.5. Hanya saja algoritma yang digunakan adalah algoritma Naive Bayes.

B. Analisa Hasil Algoritma

Setelah melakukan perancangan dan pemodelan dari kedua model Algoritma Decision Tree C4.5 dan Naïve Bayes, maka dapat diuji tingkat akurasinya untuk melihat kinerja dari kedua model tersebut. Pengujian tingkat Accuracy, Precision, Recall menggunakan confusion matrix dan kurva ROC/AUC (Area Under Cover). Dari hasil komparasi didapatkan nilai akurasi dari keseluruhan algoritma yang dikomparasi. Gambar 3 merupakan hasil akurasi dari algoritma Decision Tree C4.5. Sedangkan Gambar 4 merupakan hasil akurasi dari algoritma Naive Bayes.

accuracy: 97.12% +/- 0.72% (micro average: 97.12%)

	true positif	true negatif	class precision
pred. positif	151	27	84.83%
pred. negatif	80	3453	97.74%
class recall	65.37%	99.22%	

Gambar 3. Hasil Akurasi Decision Tree C4.5

accuracy: 76.02% +/- 2.71% (micro average: 76.02%)

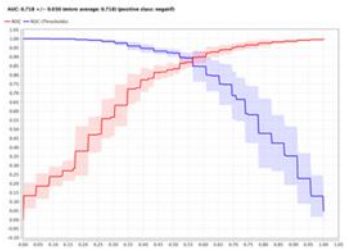
	true positif	true negatif	class precision
pred. positif	112	771	12.68%
pred. negatif	119	2709	95.79%
class recall	48.48%	77.84%	

Gambar 4. Hasil Akurasi Naive Bayes

Selain nilai akurasi dari keseluruhan algoritma yang dikomparasi, di dapatkan pula hasil AUC Gambar 5 merupakan hasil AUC dari algoritma Decision Tree C4.5. Sedangkan Gambar 6 merupakan hasil kurva ROC dari algoritma Naive Bayes.

Gambar 5. Hasil AUC Decision Tree C4.5

Setelah melakukan penelitian dan mendapatkan hasil tingkat akurasi dan nilai AUC dari kedua algoritma. Tahap berikutnya adalah membandingkan dari kedua hasil tersebut. Tabel 2 merupakan perbandingan hasil tingkat akurasi dari algoritma Decission Tree C4.5 serta Naive Bayes.

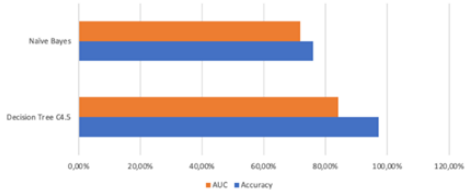


Gambar 6. Hasil AUC Naive Bayes

Setelah melakukan penelitian dan mendapatkan hasil tingkat akurasi dan nilai AUC dari kedua algoritma. Tahap berikutnya adalah membandingkan dari kedua hasil tersebut. Tabel 2 merupakan perbandingan hasil tingkat akurasi dari algoritma Decission Tree C4.5 serta Naive Bayes.

Tabel 1. Hasil Perbandingan Algoritma

Algoritma	Tingkat Akurasi	AUC
Decision Tree C4.5	97,12%	0,841
Naïve Bayes	76.02%	0,718



Gambar 7. Komparasi Algoritma

KESIMPULAN

Penelitian Hasil penelitian telah dilakukan dari 21 atribut independen menggunakan dua algoritma data mining yaitu Decision Tree C4.5 dan Naïve Bayes. Kedua algoritma ini sudah cukup baik diterapkan dalam memprediksi penyakit tiroid. Hal ini menunjukkan oleh hasil pengujian, dimana nilai akurasi Algoritma Decision Tree C4.5 adalah sebesar 97,12% dan nilai AUC 0,841, sedangkan nilai akurasi Algoritma Naïve Bayes sebesar 76,02% dan nilai AUC 0,718.

Dengan demikian maka dapat disimpulkan bahwa Algoritma Decision Tree C4.5 memiliki tingkat akurasi lebih baik sebesar 21,1% dibandingkan dengan Naïve Bayes. Selain itu, hasil perbandingan kedua algoritma tersebut dapat dinyatakan bahwa Algoritma Decision Tree C4.5 lebih unggul dari Naïve Bayes karena memiliki nilai AUC 0,841 dengan kategori Good Classification.

DAFTAR PUSTAKA

Abdalsatar, Khalid. (2021). The Efficiency of Classification Techniques in Predicting Thyroid Disease. Diakses tanggal 4 Juni 2022 http://acikerisim.karabuk.edu.tr:8080/xmlui/bitstream/handle/123456789/1355/10406874.pdf?sequence=1&isAllowed=y.

Abriana, Recha., Widagdo, Galih., Setya, Arief., Qomaruddin, M. (2019). Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naive Bayes, Jurnal Sistem dan Teknologi Informasi, Vol. 7, No. 1, Hal 48-49.

Agustiani, S., Ali, M., Andi, S., Windu, G., Siti,K. (2020). Paradigma – Jurnal Informatika dan Komputer, Vol. 22, No. 2 (September 2020). P: ISSN 1410- 5063, E-ISSN: 2579-3500.

Agustiani, S., Ali, M., Andi, S., Windu, G., Siti,K. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, Citec Journal, Vol. 2, No. 3, Hal 209-210.

Alfisahrin, S. N. (2014). Komparasi Algoritma C4.5, Naive Bayes dan Neural Network Untuk Memprediksi Penyakit Jantung. Jakarta: Pascasarjana Magister Ilmu Komputer STMIK Nusa Mandiri.

Astuti, T., Mujiati, I., Ayu, D., Ristianah, V., Lestari, W. A. (2016). Penerapan Algoritme J48 Untuk Prediksi. Jurnal Telematika, Vol.9 No. 2, Hal 1–10.

B. Santoso. (2007). Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu.

Bustami. (2013). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI : Jurnal Penelitian Teknik Informatika, Vol. 3, No.2, Hal. 127-146.

Buulolo, E. (2020). Data Mining Untuk Perguruan Tinggi. ISBN 978-623-02-0508- 8. Deepublish.

Darmawan, A., Kustian, N., Rahayu, W., & Tabebuya. (2018).Implementasi Data Mining Menggunakan Model Svm, Vol. 2, No. 3, Hal 299–307. Fatmawati. (2016). Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 dan Naive Bayes Untuk Prediksi Penyakit Diabetes. Jurnal Techno Nusa Mandiri, Vol.XIII, No.1, STMIK Nusa Mandiri, Jakarta.

GmbH, R. (2020). RapidMiner 9 Operator Reference Manual. Diakses dari www.rapidminer.com.

Gorunescu, Florin. (2011). Data Mining: Concepts, Models, and Techniques. Verlag Berlin Heidelberg : Springer.

Hairul, U., Victor, W., Triawan, A. (2016). Analisis Perbandingan Algoritma C4.5 dan Algoritma Naïve Bayes untuk Prediksi Kelulusan

Mahasiswa (Studi Kasus: Prodi Teknik Informatika Universitas Muhammadiyah Jember. Diakses dari http://repository.unmuhjember.ac.id /623/1/JURNAL.pdf.

Handayani, P., Nurlelah, E., Raharjo, M., Ramdani, P. M. (2019). Prediksi Penyakit Liver dengan Menggunakan Metode, Vol. 4, No. 1, Hal 55–59.

Inggris "Hepatic iodothyronine 5” deiodinase : The role of selenium" (PDF). Division of Biochemical Sciences, Rowett Research Institute, University Department of Clinical Chemistry; John R. ARTHUR, Fergus NICOL dan Geoffrey J. BECKETT. Diakses tanggal 2020-06-27.

Joko Suntoro. (2018). Data Mining: Algoritme dan Impelementasi Menggunakan Bahasa Pemrograman PHP. Diakses dari https://osf.io/qwbek/download.

Kementrian Kesehatan. (2015). Situasi dan Analisis Penyakit Tiroid. Diakses darihttps://pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/infodatin-tiroid.pdf. Lamfom HA. (2008). Thyroid disorders in Makkah Saudi Arabia. Ozean J Appl Sci. Vol. 1, Hal. 55.

Maimon, Rokach. (2010). Data Mining and knowledge Discovery Handbook. New York: Springer.

Melisa Sandrianti. (2022). Prodia, Waspadai Gejala Gangguan Tiroid. Diakses dari https://prodia.co.id/id/kegiatanpromosi/pressreleasedetails/waspadai-gejala-gangguan-tiroid.

Mutalazimah, Mulyono B, Murti B, Azwar S. (2013). Karakteristik demografi pada wanita usia subur dengan gangguan fungsi tiroid. Jurnal Kesehatan. Vol. 6, Hal. 123-33.

Patil, T. R., Sherekar, M., S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science and Applications, Vol. 6, No. 2, Hal 256-261.

Putra,A.,Ernawati. dan Erlansari,A. (2017). Sistem Pakar Diagnosa Penyakit Tiroid Menggunakan Metode Naive Bayes Berbasis Android. Jurnal Rekursif, Vol. 5, Nomor 3, ISSN 2303 – 0755.

Rachmat, B., Gafar, A. A., Fajriani, N., Ramdani, U., Uyun, F. R., Purnamasari, Y., Ransi, N. (2017). Implemetasi K-Means Clustering pada RapidMiner untuk Analisis Daerah Rawan Kecelakaan. April, Hal. 58–62.

Riyanto, U. (2019). Analisis Perbandingan Algoritma Naive Bayes Dan Support Vector Machine Dalam Mengklasifikasikan Jumlah Pembaca Artikel Online. JIKA (Jurnal Informatika), Vol. 2, No. 2, Hal. 62–72. Diakses dari https://doi.org/10.31000/.v2i2.1521

Rufiyanto, A., M. Rochcham., Abdul Rohman. (2020). Penerapan Algorirma C4.5 untuk Prediksi Kepuasan Mahasiswa. Yogyakarta: CV Budi Utama.

Saleh, Alfa. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, Citec Journal, Vol. 2, No. 3, Hal 209-210.

Schlitter, N., & Laessig, J. (2013). Distributed Data Analytics using RapidMiner and BOINC Distributed Data Analytics using RapidMiner and BOINC. August.

Silwattananusarn, T., & Tuamsuk, K. (2012). Data Mining and Its Application For Knowledge Management : A literature Riview From 2007 to 2012. Vol.2, Thailand: IJDKP.