JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 1, JANUARI, 2024, Pp. 70 – 78
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Algoritma *Decision Tree* untuk Prediksi Deteksi Penyakit Kanker Payudara

Ayu Dian Fitri Mellina [(1)*], Suhartono [(2)], M. Ainul Yaqin [(3)]
Teknik Informatika, Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim, Malang
e-mail : 17650087@student.uin-malang.ac.id, {suhartono,yaqinov}@ti.uin-malang.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 27 Juni 2023, direvisi 12 September 2023, diterima 6 Oktober 2023, dan dipublikasikan 25 Januari 2024.

Abstract

Cancer is a deadly disease that is difficult to cure. Early cancer detection can be done through laboratory tests to identify the cancer type. Breast cancer is a type of cancer with initial symptoms in the form of a lump. Data mining and classification methods, such as decision trees with ID3 and C5.0 algorithms, are used to categorize breast cancer. The dataset used is Breast Cancer Coimbra, which was downloaded from UCI Machine Learning in 2018. ID3 has limitations in handling unstructured data and continuous attributes, while C5.0 is better. Both algorithms produce tree models with different levels of accuracy. This study shows that the C5.0 algorithm has the best classification results with 80% accuracy, 84.2% precision, 80% recall, and 80% F1 score. 80% accuracy shows the system's classification ability, so the C5.0 model can be used to predict breast cancer.

Keywords: Breast Cancer, Classification, Prediction, Decision Tree, Machine Learning

Abstrak

Kanker merupakan penyakit mematikan yang sulit untuk disembuhkan. Deteksi dini pada kanker dapat dilakukan melalui uji laboratorium yang dapat mengidentifikasi jenis kanker. Kanker payudara merupakan salah satu jenis kanker ganas dengan gejala awal berupa benjolan. *Data mining* dan metode klasifikasi, seperti *decision tree* dengan algoritma ID3 dan C5.0, digunakan untuk mengkategorikan kanker payudara. *Dataset* yang digunakan adalah Breast Cancer Coimbra yang diunduh dari UCI Machine Learning tahun 2018. ID3 memiliki keterbatasan dalam menangani data tidak terstruktur dan atribut kontinu, sementara C5.0 lebih baik dalam hal tersebut. Kedua algoritma menghasilkan model pohon dengan tingkat keakuratan yang berbeda. Penelitian ini menunjukkan bahwa algoritma C5.0 memiliki hasil klasifikasi terbaik dengan akurasi 80%, presisi 84,2%, *recall* 80%, dan F1 *score* 80%. Akurasi 80% menunjukkan kemampuan sistem dalam klasifikasi, sehingga model C5.0 dapat digunakan untuk memprediksi kanker payudara.

Kata Kunci: Kanker Payudara, Klasifikasi, Prediksi, *Decision Tree, Machine Learning*

1. PENDAHULUAN

Kanker payudara termasuk pada golongan penyakit kanker ganas yang mana kasusnya banyak dijumpai di kalangan wanita. Penyakit ini dapat menyerang wanita pada usia berapapun, tetapi resiko terkenanya penyakit ini meningkat dengan bertambahnya usia. Penyakit ini juga dapat menyerang pria, meskipun hal ini sangat jarang terjadi. Jumlah kasus penyakit kanker payudara di seluruh dunia semakin bertambah setiap tahunnya. Salah satu gejala awal pada kanker payudara adalah munculnya benjolan kecil yang semakin lama akan bertambah semakin besar. Perubahan atau mutasi pada DNA sel payudara merupakan penyebab awal timbulnya kanker payudara. Mutasi gen biasanya terjadi karena diwariskan dari generasi sebelumnya, akan tetapi mutasi ini dapat juga terjadi tanpa penyebab yang pasti. Perempuan dengan resiko terkena kanker lebih besar adalah perempuan yang mengalami siklus menstruasi lebih banyak daripada perempuan normal lainnya, terlambat menopause, serta menstruasi dini. Peningkatan jumlah kasus kanker payudara di seluruh dunia juga disebabkan oleh perubahan pola gaya hidup yang tidak sehat serta kurangnya aktivitas fisik (Musa & Aliyu, 2020).

Terdapat beberapa atribut yang menjadi acuan dalam mendeteksi jenis kanker pada penyakit kanker payudara. Atribut tersebut membentuk sebuah pola yang kemudian dikategorikan sesuai dengan kelas yang sudah ada. Dalam menentukan jenis kanker payudara yang dialami oleh pasien di rumah sakit, pihak laboratorium rumah sakit tentunya membutuhkan waktu yang cukup lama untuk menganalisis hasil diagnosa mengenai jenis kanker tersebut. Hasil data laboratorium tidak sepenuhnya memberikan hasil yang konkrit, tentunya diperlukan hipotesis yang dapat memperkuat hasil laboratorium tersebut. Hipotesis disini bertujuan agar dapat membantu dokter menentukan jenis kanker pasien sehingga dapat ditangani dengan segera. Terdapat banyak cara atau metode dalam menentukan pola pada data mengenai penyakit kanker, salah satu cara tersebut adalah menggunakan metode *data mining*. Pada proses *data mining* terdapat beberapa metode di dalamnya, salah satunya adalah metode prediktif yang memiliki teknik yang dapat digunakan, yakni regresi dan klasifikasi (Sunjana, 2010). Dalam penelitian ini, proses klasifikasi dapat diterapkan untuk mengolah hasil data uji laboratorium dengan mengkategorikannya menjadi dua kategori, yakni kategori jinak (*benign*) dan ganas (*malignant*).

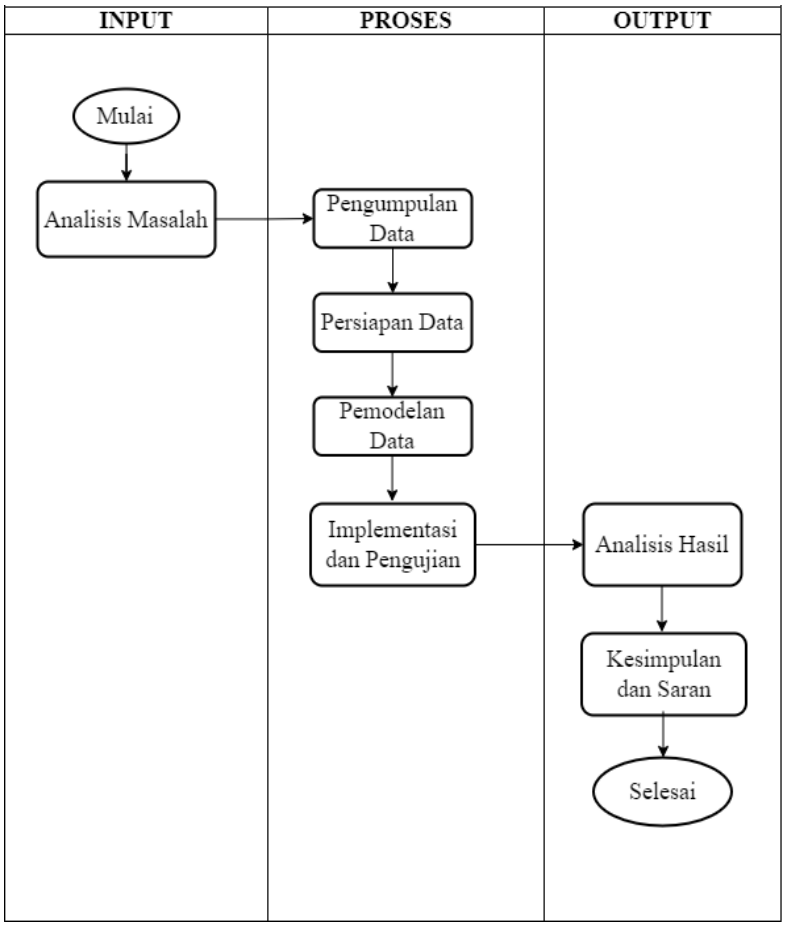
Klasifikasi yang digunakan pada penelitian ini menggunakan metode *decision tree*. *Decision tree* merupakan sebuah metode sistem prediksi yang strukturnya menyerupai pohon bercabang atau biasa juga disebut dengan struktur hierarki sehingga metode ini cocok untuk diterapkan pada permasalah penelitian ini yang di dalamnya menggambarkan sebuah persoalan dan mencari atau membutuhkan sebuah solusi dari persoalan tersebut (Wahyudin, 2009). Kedua algoritma Iterative Dichotomiser-3 (ID3) dan C5.0 merupakan salah satu algoritma *decision tree* yang dapat digunakan untuk tujuan tersebut (Wei, 2011).

Kedua algoritma Iterative Dichotomiser-3 (ID3) dan C5.0 merupakan algoritma *decision tree* yang dapat digunakan untuk deteksi penyakit kanker payudara. Algoritma Iterative Dichotomiser-3 (ID3) merupakan algoritma yang pertama kali dikembangkan oleh *decision tree* yang memiliki kemampuan yang cukup baik dalam menangani data yang terstruktur. Namun, algoritma ini tidak dapat menangani data yang tidak terstruktur dan tidak dapat menangani atribut yang bernilai kontinu. Sedangkan, C5.0 merupakan pengembangan dari algoritma Iterative Dichotomiser-3 (ID3) yang memiliki kemampuan yang lebih baik dalam menangani data yang tidak terstruktur dan dapat menangani atribut yang bernilai kontinu. Algoritma ini juga memiliki kemampuan untuk membangun model yang lebih akurat dan memiliki fitur pruning yang dapat membantu mengurangi overfitting pada model. Kedua algoritma tersebut dapat menghasilkan model pohon (*tree*) yang berbeda dengan dataset yang sama. Model yang dihasilkan dari kedua algoritma tersebut tentunya memilki tingkat keakuratan yang berbeda.

Latar belakang penelitian ini adalah untuk mengembangkan suatu sistem prediksi menggunakan metode *decision tree* serta membandingkan model yang dihasilkan oleh algoritma yang ada pada *decision tree* yakni Iterative Dichotomiser-3 (ID3) dan C5.0 untuk deteksi penyakit kanker payudara. Sistem ini diharapkan dapat membantu dokter dalam mengambil keputusan yang lebih akurat dan tepat waktu dalam menegakkan diagnosis kanker payudara.

2. METODE PENELITIAN

Dalam menjalankan program yang dibangun, penelitian ini menggunakan bahasa pemrograman R, serta menggunakan beberapa *library* yang disediakan oleh R untuk melakukan proses *machine learning.* Alur penelitian dilakukan dengan memulai dari menganalisis masalah, melakukan pengumpulan data, kemudian mempersiapkan data yang akan digunakan, dan dilanjutkan dengan pemodelan data dengan metode yang sudah ada. Langkah selanjutnya adalah melakukan implementasi dan pengujian, sehingga bisa didapatkan hasil dan dapat ditarik menjadi sebuah kesimpulan. Adapun alur penelitian ditunjukan oleh Gambar 1.



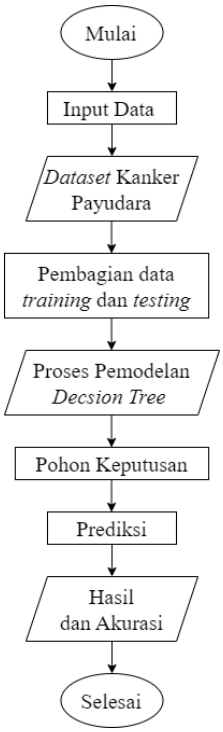
Gambar 1 Prosedur Penelitiain

2.1 Pengumpulan Data

Data yang digunakan pada penelitian ini diperoleh dari *website* resmi UCI *Machine Learning Repository*. Data tersebut merupakan *dataset breast cancer* coimbra tahun 2018 (Patrcio et al., 2018). Total jumlah keseluruhan *dataset* yang digunakan sebanyak 116 data, dengan perincian 52 data merupakan kelas *benign* (jinak) serta 64 data merupakan kelas *malign* (ganas). *Missing value* pada keseluruhan data yang digunakan berjumlah 0 atau tidak ada. Pada *dataset* tersebut terdapat 9 atribut dan 1 kelas klasifikasi, 9 atribut tersebut yaitu umur, glukosa, resistin, *Homeostatic Model Assessment* (HOMA), insulin, leptin, *Body Mass Index* (BMI), adiponectin, dan MCP-1.

2.2 Desain Sistem

Gambaran umum desain sistem prediksi deteksi penyakit kanker payudara menggunakan metode *decision tree* yakni metode Iterative Dichotomiser-3 (ID3) dan algoritma C5.0 yang dijabarkan pada *flowchart* Gambar 2. Alur dari desain sistem yang dibangun dimulai dari penginputan *dataset* kanker payudara, kemudian pembagian *dataset* menjadi dua bagian yakni data *training* dan *testing.* Kemudian proses *learning* dengan menerapkan model *decision tree* dengan menggunakan algoritma ID3 dan C5.0, proses pembentukan pohon keputusan, prediksi, kemudian hasil dan akurasi dari sistem yang dibangun.



Gambar 2 Desain Sistem

2.3 Pembagian *Dataset*

Pada langkah ini, *dataset* yang digunakan adalah data terstruktur sebanyak 116 *item*. *Dataset* terdiri dari usia, BMI, kadar glukosa dalam darah, insulin, HOMA, Leptin, Adiponectin, Resistin, dan MCP.1. *Dataset* pelatihan diklasifikasikan menjadi dua, yaitu kanker jinak dan kanker ganas. *Dataset* tersebut dibagi menjadi dua bagian. Bagian pertama digunakan sebagai data pelatihan (*training*) untuk membuat model klasifikasi, sementara bagian kedua digunakan sebagai data uji (*testing*) untuk mengevaluasi model yang telah dibuat. Pembagian *dataset* pada penelitian ini dilakukan dengan jumlah persentase yang beragam untuk mengetahui variasi nilai akurasi yang yang ada pada kedua metode *decision tree.* Adapun pembagian persentase pengujian data dijelaskan pada Tabel 1.

Tabel 1 Skenario Pengujian

Skenario	Jumlah Data Latih	Jumlah Data Uji	Keterangan
1	80%	20%	Dataset sebanyak 80% akan menjadi data latih (data training), sedangkan 20% sisanya akan menjadi data uji (data testing).
2	75%	25%	Dataset sebanyak 75% akan menjadi data latih (data training), sedangkan 25% sisanya akan menjadi data uji (data testing).
3	70%	30%	Dataset sebanyak 70% akan menjadi data latih (data training), sedangkan 30% sisanya akan menjadi data uji (data testing).
4	50%	50%	Dataset sebanyak 50% akan menjadi data latih (data training), sedangkan 50% sisanya akan menjadi data uji (data testing).
5	25%	75%	Dataset sebanyak 25% akan menjadi data latih (data training), sedangkan 75% sisanya akan menjadi data uji (data testing).

2.4 *Modeling*

Proses pemodelan pada algoritma *decision tree* Iterative Dichotomiser-3 (ID3) dilakukan dengan cara pembentukan pohon klasifikasi menggunakan dua langkah. Langkah pertama yang dilakukan adalah dengan cara menentukan nilai *entropy*, kemudian dilanjut dengan langkah kedua yakni menghitung nilai *information gain* pada tiap variabel (Pribadi et al., 2018). *Entropy* pada proses ini berfungsi untuk mengukur node yang digunakan sebagai parameter pada sampel data. Perhitungan nilai *entropy* pada algoritma ID3 yakni sebagaimana pada Pers. (1). Di mana S merupakan data sampel yang digunakan sebagai training, P_+ adalah probablitas sampel S dengan kelas positif, dan P_- adalah probablitas sampel S dengan kelas positif.

$$Entropy(S) = -P_+ log_2 P_+ - P_- log_2 P_- \quad (1)$$

Pengurangan *entropy* pada algoritma ID3 disebut dengan *information gain*. Pembagian sampel S terhadap atribut X dihitung dengan menggunakan rumus *information gain* yakni sebagaimana pada Pers. (2). Di mana X adalah atribut, V yaitu nilai yang memungkin untuk atribut X, dan $value(X)$ himpunan yang mungkin untuk atribut (X).

$$Gain(S, X) = Entropy(S) - \sum_{V\ \epsilon\ value(X)} \frac{|S_V|}{|S|}\ Entropy(S_V) \quad (2)$$

Setelah *information gain* pada semua atribut dihitung, kemudian dipilih nilai *information gain* tertinggi untuk dijadikan *root* pada suatu pohon keputusan. Hal ini dilakukan seterusnya hingga parameter pada tiap-tiap atribut terklasifikasi dengan sempurna. Proses perhitungan tersebut memiliki kesamaan dengan algoritma ID-3 yakni mulai dari perhitungan *entropy* dan *information gain*, akan tetapi pada algoritma C5.0 atribut dengan *gain ratio* tertinggi akan dipilih sebagai *root node* (Wei, 2011). Adapun rumus perhitungan *gain ratio* yakni sebagaimana pada Pers. (3).

$$Gain\ Ratio = \frac{Gain(S, X)}{\sum_{i=1}^{m} Entropy(S_i)} \quad (3)$$

2.5 Evaluasi Sistem

Evaluasi sistem dilakukan dengan menggunakan *binary class confusion matrix* karena penelitian ini termasuk dalam klasifikasi dua kelas. Evaluasi pada sistem yang dibangun meliputi nilai akurasi, presisi, *recall*, dan F-1 *score.* Akurasi klasifikasi mengacu pada persentase data pengujian yang diklasifikasikan dengan benar oleh model. Jika akurasi klasifikasi dianggap memadai, maka model dapat digunakan untuk mengklasifikasikan set data di masa mendatang yang memiliki label kelas yang belum diketahui (Agarwal, 2013). Presisi atau yang dikenal juga sebagai *precision*, menggambarkan proporsi unit yang diprediksi sebagai positif oleh model yang juga benar-benar positif dalam data yang sebenarnya. Presisi dapat diinterpretasikan sebagai tingkat kesesuaian antara permintaan informasi dan respons terhadap permintaan tersebut (Mayadewi & Rosely, 2015). *Recall* adalah hasil perhitungan yang menunjukan sejauh mana semua data uji yang positif telah diprediksi dengan benar sebagai positif dalam klasifikasi. *Recall* juga dikenal sebagai *True Positive Rate* (TPR), sensitivitas, dan probabilitas deteksi (Grandini et al., 2020). Dalam klasifikasi *binary class* di mana setiap observasi hanya memiliki satu label, skor F1 yang dihitung dengan metode mikro (*micro-averaged* F1) sama dengan akurasi klasifikasi secara keseluruhan (Zhang et al., 2015). Rumus dari akurasi, presisi, *recall* dan F1 *score* ditunjukan oleh Pers. (4) sampai (7).

$$accuracy = \frac{TP}{Total\ data\ testing} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score = \left(\frac{2 \times Recall \times Precision}{Recall + Precision}\right) \quad (7)$$

3. HASIL DAN PEMBAHASAN

Hasil uji coba tingkat akurasi pada skenario pengujian yang dilakukan mulai dari iterasi pertama hingga iterasi kelima dikelompokan dalam Tabel 2. Dari hasil pemaparan pada Tabel 2 tersebut nilai akurasi yang paling optimal didapatkan oleh skenario pengujian dengan pembagian data dengan perbandingan rasio yakni sebesar 70:30 dengan artian 70% dari total keseluruhan data atau sebanyak 81 data digunakan sebagai data latih (*training*), sementara 30% dari total keseluruhan data atau sebanyak 35 data digunakan sebagai data uji (*testing*). Tabel 3 menunjukan lima sampel data *training*, sedangkan Tabel 4 menunjukan lima sampel data *testing*.

Tabel 2 Hasil Akurasi pada Skenario Pengujian

No.	Skenario Pengujian	Akurasi	
		Iterative Dichotomiser-3 (ID3)	C5.0
1	Skenario 1 (80%-20%)	69.57%	75.00%
2	Skenario 2 (75%-25%)	72.41%	68.97%
3	Skenario 3 (70%-30%)	77.14%	80.00%
4	Skenario 4 (50%-50%)	72.41%	63.79%
5	Skenario 5 (25%-75%)	68.57%	70.11%

Tabel 3 Sampel Data *Training*

No.	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
1	29	32,270	84	5,81	1,203	45,619	6,209	24,603	904,981	1
2	66	36,212	101	15,533	3,869	74,706	7,539	22,320	864,968	1
3	86	21,111	92	3,549	0,805	6,699	4,819	10,576	773,92	1
4	69	28,444	108	8,808	2,346	14,748	5,288	16,485	353,568	2
5	51	22,892	103	2,74	0,696	8,016	9,349	11,554	359,232	2

Tabel 4 Sampel Data *Testing*

No.	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
1	68	21,367	77	3,226	0,612	9,882	7,169	12,766	928,22	1
2	49	22,854	92	3,226	0,732	6,831	13,679	10,317	530,41	1
3	34	21,47	78	3,469	0,667	14,57	13,11	6,92	354,6	1
4	29	23,01	82	5,663	1,145	35,59	26,72	4,58	174,8	1
5	25	22,86	82	4,09	0,827	20,45	23,67	5,14	313,73	1

Tabel 5 Hasil Prediksi Klasifikasi ID3

No.	*Classification*	Prediksi ID3
1	1	2
2	1	1
3	1	1
4	1	1
5	1	1

Hasil prediksi dari klasifikasi kemudian ditunjukan oleh *confusion matrix* pada Tabel 6. Seluruh nilai tersebut diperlukan untuk proses perhitungan nilai *accuracy*, *precision*, *recall*, dan *micro* F1 pada *confusion matrix*. Pada Tabel 7 berisi konfusi *matrix* beserta deskripsi dari nilai-nilai

tersebut. Selanjutnya dilakukan perhitungan nilai akurasi, presisi, *recall*, dan F1 *score*, dapat dilakukan dengan mengunakan Pers. (4) sampai Pers. (7).

Tabel 6 Hasil Prediksi Algoritma ID3

		Prediksi ID3	
		1	**2**
Aktual	**1.**	14	6
	2.	2	13

Tabel 7 *Confussion Matrix* Algoritma ID3

		Prediksi ID3	
		1	**2**
Aktual	**1.**	14 (TP)	6 (FN)
	2.	2 (FP)	13 (TN)

$$Accuracy = \frac{14+13}{35} = \frac{27}{35} = 0{,}7714 = 77{,}14\%$$

$$Precision = \frac{14}{14+2} \times 100\% = \frac{14}{16} \times 100\% = 87{,}5\%$$

$$Recall = \frac{14}{14+6} \times 100\% = \frac{14}{20} \times 100\% = 70\%$$

$$\text{F1 } score = \frac{2 \times 70\% \times 87{,}5\%}{70\% + 87{,}5\%} = \frac{12{,}250}{157} = 78\%$$

Prediksi selanjutnya menggunakan *decision tree* C5.0 yang hasil klasifikasi dan prediksinya ditunjukan pada Tabel 8 dan 9. Pada Tabel 10 berisi *confusion matrix* pada algoritma C5.0 beserta deskripsi dari nilai-nilai pada tiap prediksi tersebut, maka perhitungan nilai akurasi, presisi, *recall,* dan F1 *score*, dapat dilakukan dengan mengunakan Pers (4) sampai Pers. (7).

Tabel 8 Hasil Prediksi Klasifikasi Algoritma C5.0

No.	***Classification***	**Prediksi C5**
1	1	1
2	1	2
3	1	1
4	1	1
5	1	1

Tabel 9 Hasil Prediksi Algoritma ID3

		Prediksi ID3	
		1	**2**
Aktual	**1.**	16	4
	2.	3	12

Tabel 10 *Confussion Matrix* Algoritma ID3

		Prediksi ID3	
		1	**2**
Aktual	**1.**	16 (TP)	4 (FN)
	2.	3 (FP)	12 (TN)

$$Accuracy = \frac{16+12}{35} = \frac{28}{35} = 0{,}8 = 80\%$$

$$Precision = \frac{16}{16+3} \times 100\% = \frac{16}{19} \times 100\% = 84{,}2\%$$

$$Recall = \frac{16}{16+4} \times 100\% = \frac{16}{20} \times 100\% = 80\%$$

$$\text{F1 } score = \frac{2 \times 84{,}2\% \times 80\%}{84{,}2\% + 80\%} = \frac{13{,}472}{164{,}2} = 82\%$$

Hasil evaluasi yang telah dilakukan pada uji coba skenario pertama hingga kelima dengan menggunakan *ratio* yang berbeda-beda pada data latih (*training*) dan data uji (*testing*), serta tidak saling beririsan. Dapat dilihat bahwasanya algoritma Iterative Dichotomiser-3 (ID3) memiliki tingkat nilai akurasi yang konsisten pada ketiga skenario uji coba, yang berkisar 72.41%-77.14%. Hal ini disebabkan karna algoritma Iterative Dichotomiser-3 (ID3) sulit memprediksi model pohon klasifikasi dengan data uji (*testing*) yang berubah-ubah secara signifikan dan bervariasi seperti pada skenario uji coba kedua sampai keempat, hal inilah yang menyebabkan hasil kinerja pada algoritma tersebut lebih konsisten, karena banyaknya data latih yang digunakan tidak mempengaruhi nilai akurasi. Serta pohon keputusan yang dihasilkan cenderung sederhana dan lebih umum. Sedangkan pada algoritma C5.0, nilai akurasi yang didapatkan lebih bervariasi pada tiap uji coba skenario. Hal ini dikarenakan algoritma C5.0 lebih memiliki kecenderungan untuk mempelajari pola yang sangat spesifik pada data latih (*training*) yang digunakan. Hal inilah yang menyebabkan hasil pohon keputusan pada algoritma C5.0 lebih sederhana jika dibandingkan dengan ID3, karena algoritma C5.0 akan melakukan pemangkasan (*pruning*) terhadap cabang-cabang yang tidak signifikan dari pohon keputusan untuk menghindari mempelajari pola yang terlalu spesifik pada data pelatihan. Algoritma C5.0 memiliki lebih banyak *hyper-parameter* dan aturan (*rule*) yang dapat mempengaruhi hasil akurasi. Jika *ratio* pembagian data yang digunakan berbeda dalam setiap skenario pengujian, maka hasil akurasi C5.0 dapat bervariasi secara signifikan.

4. KESIMPULAN

Berdasarkan hasil perbandingan akurasi dari dua algoritma pada metode *decision tree* yakni Iterative Dichotomiser-3 (ID3) dan C5.0 secara umum diambil dari semua skenario pengujian yang telah dilakukan, maka algoritma C5.0 dengan perolehan nilai akurasi sebesar 80% mengungguli algortima Iterative Dichotomiser-3 (ID3) dengan hasil nilai akurasi sebesar 77,14%. Algortima C5.0 memiliki keunggulan dalam kompleksitas model yan dihasilkan dengan menggunakan teknik *pruning* sehingga dapat meningkatkan kinerja dan akurasi model. Sementara untuk hasil lain dari Iterative Dichotomiser-3 (ID3) nilai presisi yang didapat sebesar 87,5%, *recall* sebesar 70%, dan F1 *score* sebesar 78%. Sementara pada algoritma C5.0 nilai presisi sebesar 84,2%, *recall* sebesar 80%, dan nilai F1 *score* sebesar 82%. Hasil tersebut menunjukan bahwa nilai akurasi, presisi, *recall*, dan F1 *score* pada C5.0 termasuk ke dalam kategori baik. Sehingga dapat ditarik kesimpulan bahwa pemodelan sistem dengan menggunakan algoritma C5.0 memiliki tingkat akurasi yang lebih baik dibandingkan dengan algoritma Iterative Dichotomiser-3 (ID3).

DAFTAR PUSTAKA

Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. *2013 International Conference on Machine Intelligence and Research Advancement*, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. http://arxiv.org/abs/2008.05756

Mayadewi, P., & Rosely, E. (2015). Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining. *Seminar Nasional Sistem Informasi Indonesia*

(SESINDO) 2015, *2015*. https://is.its.ac.id/pubs/oajis/index.php/home/detail/1582/PREDIKSI-NILAI-PROYEK-AKHIR-MAHASISWA-MENGGUNAKAN-ALGORITMA-KLASIFIKASI-DATA-MINING

Musa, A. A., & Aliyu, U. M. (2020). Application of Machine Learning Techniques in Predicting of Breast Cancer Metastases Using Decision Tree Algorithm, in Sokoto Northwestern Nigeria. *Journal of Data Mining in Genomics & Proteomics*, *11*(1). https://www.walshmedicalmedia.com/open-access/application-of-machine-learning-techniques-in-predicting-of-breast-cancer-metastases-using-decision-tree-algorithm-in-sokoto-north-53078.html

Patrcio, M., Pereira, J., Crisstomo, J., Matafome, P., Seia, R., & Caramelo, F. (2018). *Breast Cancer Coimbra*. UCI Machine Learning Repository. https://doi.org/https://doi.org/10.24432/C52P59

Pribadi, D., Athiry, S., Saputra, R. A., Supiandi, A., Prayudi, D., Nusa, S., & Sukabumi, M. (2018). Sistem Pakar Diagnosa Penyakit Demam Berdarah Dengue Menggunakan Algoritma Iterative Dichotomiser 3 (ID3). *SNIT 2018*, *1*(1), 129–133. https://seminar.bsi.ac.id/snit/index.php/snit-2018/article/view/37

Sunjana. (2010). Aplikasi Mining Data Mahasiswa dengan Metode Klasifikasi Decision Tree. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 1907–5022. https://journal.uii.ac.id/Snati/article/view/1857

Wahyudin. (2009). *Metode Iterative Dichotomizer 3 (ID3) Untuk Penerimaan Mahasiswa Baru*. Universitas Pendidikan Indonesia.

Wei, W. (2011). ID3 Algorithm and C4.5 Algorithm Based on Decision Tree. *Journal of Hubei University of Technology*.

Zhang, D., Wang, J., & Zhao, X. (2015). Estimating the Uncertainty of Average F1 Scores. *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, 317–320. https://doi.org/10.1145/2808194.2809488