Analisis Perbandingan Klasifikasi Algoritma CART dengan Algoritma C 4.5 Pada Kasus Penderita Kanker Payudara

Fitria Melani[1,*], Sulastri[2]

[1,2] Fakultas Teknologi Informasi dan Industri, Program Studi Sistem Informasi, Universitas Stikubank (UNISBANK), Semarang, Indonesia
Email: [1,*] fitriamelani@mhs.unisbank.ac.id, [2] sulastri@edu.unisbank.ac.id

Abstrak—Kanker menjadi salah satu jenis penyakit berbahaya penyebab terjadinya kematian. Dari jumlah keseluruhan kasus kanker di dunia, jenis kanker yang banyak diderita manusia dengan presentase kasus mencapai 11,7% tercatat sebagai kanker payudara. Hal ini dapat terjadi dikarenakan kanker payudara tidak didiagnosa lebih awal. Maka dari itu, jika penyakit kanker payudara dapat diidentifikasi lebih cepat, maka resiko yang mungkin terjadi dapat diminimalisir. Seiring dengan kemajuan teknologi saat ini, data – data pasien kanker payudara dapat diolah dan dimanfaatkan untuk menemukan informasi yang berguna bagi kehidupan masyarakat. Dalam melakukan pengolahan data ada beragam cara yang dapat digunakan, contohnya dengan menggunakan teknik pengolahan data mining. Data mining memiliki bermacam – macam metode yang dapat diterapkan bergantung dengan tujuan dalam penggunaannya, Klasifikasi menjadi salah satu metode yang paling sering dipergunakan dalam teknik data mining. Dalam data mining teknik klasifikasi memiliki beragam model algoritma dengan tingkat kinerja yang bervariasi. Permasalahan dalam penelitian ini berfokus tentang bagaimana cara dalam melakukan analisis perbandingan model algoritma klasifikasi pada dataset kanker payudara yang diambil dari *platform* Kaggle.com. Penelitian ini bertujuan membandingkan algoritma CART dan C 4.5 untuk mendapatkan hasil performa yang optimal. Implementasi pada penelitian ini menggunakan bahasa pemrograman R dengan *software* Rstudio. Dalam 6 kali percobaan dengan probabilitas dataset yang berbeda menghasilkan model pohon keputusan dengan nilai *Accuracy, Recall, Precision* dan *Error rate* yang berbeda. Dari percobaan yang dilakukan, didapatkan rata-rata tingkat performa algoritma CART sebesar 72 %, sedangkan algoritma C 4.5 sebesar 73%, sementara itu variabel yang paling berpengaruh dalam kematian pasien kanker payudara adalah *Survival Months*. Berdasarkan hasil analisis perbandingan yang telah dilakukan dapat diketahui bahwa tingkat performa dari algoritma C 4.5 lebih baik dan stabil jika dibandingkan dengan tingkat performa dari algoritma CART.

Kata Kunci: Data Mining, Klasifikasi, Algoritma CART, Algoritma C 4.5, Kanker Payudara

Abstract—Cancer is a type of dangerous disease that causes death. Of the total number of cancer cases in the world, the type of cancer that affects many humans with a percentage of cases reaching 11.7% is recorded as breast cancer. This can happen because breast cancer cells are not diagnosed sooner. Therefore, if breast cancer can be identified more quickly, then the risks that might occur can be minimized. Along with current technological advances, breast cancer patient data can be processed and used to find useful information for people's lives. In carrying out data processing there are various ways that can be used, for example by using data mining processing techniques. Data mining has various methods that can be applied depending on the purpose of its use. Classification is one of the most frequently used methods in data mining techniques. In data mining classification techniques have various algorithm models with varying levels of performance. The problem in this study focuses on how to carry out a comparative analysis of the classification algorithm model on breast cancer datasets taken from the Kaggle.com platform. This study aims to compare the CART and C 4.5 algorithms to obtain optimal performance results. Implementation in this study uses the R programming language with Rstudio software. In 6 trials with different dataset probabilities resulted in a decision tree model with different Accuracy, Recall, Precision and Error rate values. From the experiments conducted, it was found that the average performance level of the CART algorithm was 72%, while the C 4.5 algorithm was 73%, while the most influential variable in the death of breast cancer patients was Survival Months. Based on the results of the comparative analysis that has been done, it can be seen that the performance level of the C 4.5 algorithm is better and more stable when compared to the performance level of the CART algorithm.

Keywords: Data Mining, Classification, CART Algorithm, C 4.5 Algorithm, Breast Cancer

1. PENDAHULUAN

Kanker menjadi salah satu jenis penyakit berbahaya penyebab terjadinya kematian pada manusia di seluruh dunia [1]. Menurut data *Global burden of cancer* (Globocan) dari *International agency for research on cancer* yang didirikan oleh *World health organization* (WHO) mencatat jumlah kasus pasien penderita kanker di seluruh dunia pada tahun 2020 menyentuh angka hingga 19,3 juta kasus, dimana jumlah kematian yang disebabkan oleh penyakit kanker tercatat mencapai 10 juta jiwa. Hal itu diperkirakan akan meningkat di tahun 2040 dengan peningkatan jumlah kasus sebesar 30,2 juta. Dari jumlah keseluruhan kasus kanker di dunia, jenis kanker yang paling banyak diderita oleh manusia yaitu kanker payudara dengan presentase kasus yang terjadi sebesar 11,7%. [2].

Perkembangan zaman yang semakin maju ditandai dengan pesatnya pertumbuhan teknologi informasi yang ada, hal ini dapat dimanfaatkan sebaik mungkin untuk mengatasi masalah – masalah yang sedang dihadapi. Teknologi informasi sendiri telah memberikan banyak informasi serta data yang bermanfaat dan dapat diolah agar nantinya menghasilkan sebuah *knowledge* atau pengetahuan baru untuk kehidupan manusia. Teknologi informasi dapat menghasilakn data dari berbagai bidang, salah satunya pada bidang kesehatan. Data yang dihasilan, salah satunya berupa data penyakit seperti kanker payudara. Pengolahan data dengan memanfaatkan data tersebut dapat dilakukan untuk menemukan informasi lebih lanjut terkait kasus kanker payudara. Hal ini dapat dimanfaatkan untuk membantu dalam pengobatan kanker payudara ataupun untuk langkah pencegahan agar tidak terkena penyakit kanker payudara tersebut [3]. Penelitian – penelitian dalam hal penanganan dan pencegahan kanker payudara telah banyak dilakukan, tujuannya sangat beragam ada yang membuat sebuah sistem pendeteksian sel kanker payudara, ada pula yang hanya sekedar untuk melakukan pengujian algoritma yang digunakan, itu semua bergantung dengan keahlian dari setiap peneliti [4].

Data dapat diolah dengan berbagai teknik, salah satunya dengan penggunaan teknik data mining. Data mining yaitu sebuah istilah untuk menyebut suatu kegiatan menemukan infromasi tersembunyi dalam sebuah basis data. Data mining adalah proses pengolahan data semi otomatis yang memanfaatkan teknik statistik, ilmu matematika, kecerdasan buatan dan *machine learning* dalam mengekstraksi dan menandai informasi pengetahuan potensial yang berguna dan dapat bermafaat dan disimpan dalam sebuah database besar. Turban dalam [5].

Teknik klasifikasi merupakan salah satu contoh teknik yang ada dalam data mining. Dengan menggunakan teknik klasifikasi, dimungkinkan untuk menemukan rules atau informasi berdasarkan sekumpulan variabel dengan kelas target tertentu dalam data yang cukup besar. Kamber & Han dalam [6]. Teknik klasifikasi bertujuan untuk melakukan prediksi pada sebuah kelas target dimana prediksi dilakukan dengan akurat dan memanfaatkan variabel – variabel prediktor yang relevan. Klasifikasi menjadi salah satu teknik dalam data mining yang cukup pesat perkembangannya [6]. Dimana hal itu ditandai dengan banyaknya penelitian yang mengimplementasikan teknik klasifikasi data mining pada berbagai bidang misalnya pada bidang kesehatan, dimana pada penelitian ini akan menggunakan dataset kanker payudara.

Teknik klasifikasi memiliki beragam model algoritma dengan tingkat kinerja yang bervariasi. Harper dalam [6] mengungkapkan bahwa setiap model algoritma klasifikasi mempunyai nilai kinerja yang berbeda tergantung pada kasus ataupun dataset yang dipergunakan. Kemudian pendapat tersebut didukung dengan penelitian lain yang dilakukan oleh Rahman dan Afrozi dalam [4] yang mengatakan bahwa suatu model algoritma yang sesuai diterapkan pada suatu kasus belum tentu sesuai jika diterapkan pada kasus lain. Temurtas, Yumusak dan Temurtas dalam [4] menyatakan di dalam penelitiannya bahwa jenis algoritma klasifikasi mempunyai nilai kinerja yang bervariasi dan sangat bergantung dengan model algoritma yang dipergunakan.

Selain penelitian yang telah disebutkan diatas, penelitian lain yang dijadikan sebagai bahan acuan yaitu yang dilakukan oleh [3] dalam penelitian ini dilakukan pengklasifikasian dengan algoritma Naïve bayes menggunakan data pasien penderita kanker payudara yang diperoleh dari portal data terbuka *University Medical Center*. Hasil yang diperoleh menunjukan nilai akurasi terhadap pasien yang dinyatakan kambuh atau tidak kambuh yaitu sebesar 71,43%. Dimana dalam penelitian ini disebutkan bahwa penyebab nilai akurasi yang rendah dapat terjadi dikarenakan kekurang kompleksitasan data yang digunakan.

Sementara itu [7] melakukan perbandingan tingkat akurasi antara algoritma C 4.5 dengan algoritma CART dalam melakukan prediksi kategori IPK mahasiswa, yang mendapatkan hasil akurasi yang berbeda dari 2 kategori yang diteliti. Pertama akurasi dalam memprediksi kategori IPK dengan data non numerik mendapatkan nilai akurasi yang sama antara kedua algoritma yaitu sebesar 86,86%. Sedangkan dalam memprediksi kategori IPK data numerik didapatkan hasil yang berbeda antara kedua algoritma, dimana algoritma CART mendapatkan skor akurasi yang lebih tinggi dibandingkan dengan skor algoritma C 4.5 yaitu sebesar 63,16% berbanding dengan 61,54%.

Pentingnya analisis perbandingan algoritma untuk menemukan sebuah model algoritma yang memiliki tingkat performa terbaik yang dapat diterapkan pada suatu dataset tertentu. Dalam hal ini, peneliti tertarik untuk melakukan uji coba perbandingan tingkat performa teknik klasifikasi antara algoritma CART dengan algoritma C 4.5 dengan memanfaatkan data kasus pasien penderita kanker payudara. Tujuannya untuk melihat algoritma

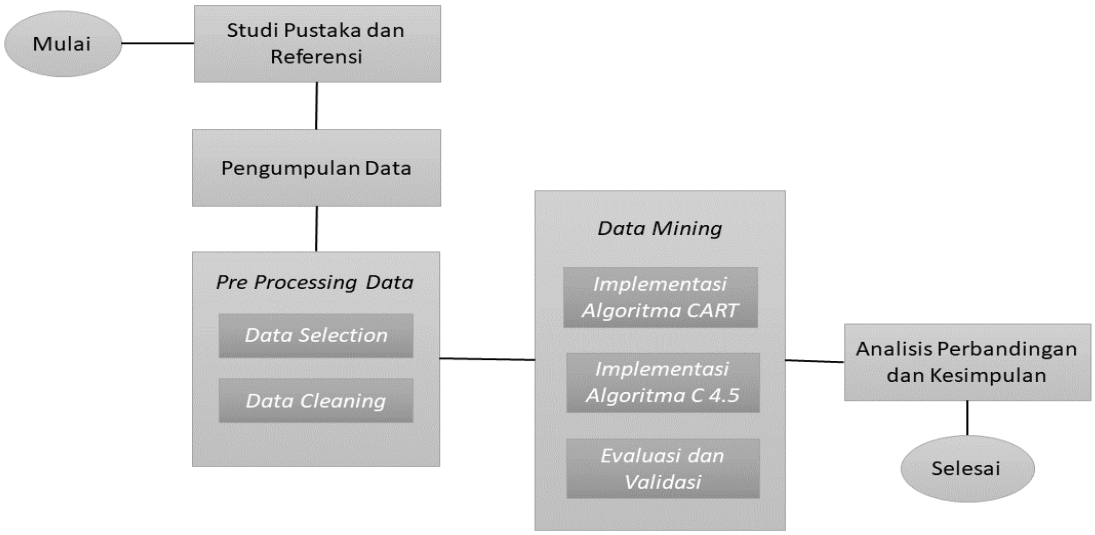
mana yang memiliki tingkat performa paling optimal dan tingkat keakuratan paling tinggi dalam mengklasifikasikan data kanker payudara.

.

2. METODE PENELITIAN

2.1 Tahap Penelitian

Teknik klasifikasi dipiih untuk digunakan dalam penelitian ini, dimana model algoritma yang akan diimplementasikan yaitu algoritma CART (*Classification and Regression Trees*) dan algoritma C 4.5. kedua algoritma tersebut masuk kedalam jenis teknik klasifikasi pohon keputusan (*Decision Tree*). Algoritma CART membangun pohon keputusan dengan menyeleksi percabangan yang paling optimum dari masing – masing node. Febti Eka dalam [8]. Sedangkan algoritma C 4.5 merupakan algoritma yang telah banyak diketahui dan dipergunakan dalam data mining khususnya teknik klasifikas, dimana algoritma ini umum digunakan pada data – data yang mempunyai variabel dengan tipe numerik ataupun kategorial. [9]. Hasil dari setiap metode kemudian dievalusai dan divalidasi dengan menggunakan *Confusion Matrix* untuk menentukan tingkat akurasi model. Tahapan dalam penelitian ini dapat digambarkan seperti pada Gambar berikut ini.



Gambar 1. Tahap Penelitian

2.2 Pengumpulan Data

Tahap awal pengumpulan data menggunakan sumber data sekunder, dengan data yang dipergunakan dalam penelitian ini yaitu diperoleh dari *platform public dataset* Kaggle.com dengan alamat web https://www.kaggle.com/datasets/reihanenamdari/breast-cancer. Data yang dikumpulkan yaitu Dataset pasien kanker payudara yang diperoleh dari program SEER NCI November 2017 yang memberikan informasi tentang statistik kanker berbasis populasi. Total *record* data yang terkumpul sejumlah 4024 *record*, yang terdiri dari 3408 *record* berisi data pasien kanker payudara yang berstatus masih "Hidup" dan 616 *record* berisi data pasien kanker payudara yang berstatus telah "Meninggal". Atribut atau variabel yang digunakan ada 15 variabel *predictor* dan 1 variabel *class*, detail variabel yang dipergunakan dapat dilihat pada gambar 2.

Variabel Predictor	Variabel *Class*
Age	Status (Alive & Dead)
Race	
Marital Status	
T Stage	
N Stage	
6th Stage	
Differentiate	
Grade	
A Stage	
Tumor Size	
Estrogen Status	
Progesterone Status	
Regional Node Examines	
Regional Node Positive	
Survival Months	

Gambar 2. *Variabel Predictor dan Variabel Class*

2.3 Pengolahan Data Awal

Agar memperoleh sebuah data dengan kulitas yang baik, ada beberapa cara yang dapat dijalankan pada tahap *preprocessing data* yaitu dengan beberapa tahap berikut :

2.3.1 *Data Selection*

Pada proses *data selection* dilakukannya tahap pemilihan variabel yang diperlukan untuk proses klasifikasi. Pada tahap *data selection*, variabel yang tidak diperlukan dalam proses klasifikasi akan dihapus atau dihilangkan [6].

Age	Race	Marital Status	T Stage	N Stage	6th Stage	differentiate	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Regional Node Positive	Survival Months	Status
68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	4	Positive	Positive	24	1	60	Alive
50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	35	Positive	Positive	14	5	62	Alive
58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	63	Positive	Positive	14	7	75	Alive
58	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	18	Positive	Positive	2	1	84	Alive
47	White	Married	T2	N1	IIB	Poorly differentiated	3	Regional	41	Positive	Positive	3	1	50	Alive
51	White	Single	T1	N1	IIA	Moderately differentiated	2	Regional	20	Positive	Positive	18	2	89	Alive
51	White	Married	T1	N1	IIA	Well differentiated	1	Regional	8	Positive	Positive	11	1	54	Alive
40	White	Married	T2	N1	IIB	Moderately differentiated	2	Regional	30	Positive	Positive	9	1	14	Dead
40	White	Divorced	T4	N3	IIIC	Poorly differentiated	3	Regional	103	Positive	Positive	20	18	70	Alive
69	White	Married	T4	N3	IIIC	Well differentiated	1	Distant	32	Positive	Positive	21	12	92	Alive

Gambar 3. Data Awal Kasus Kanker Payudara

Variabel yang di hilangkan dan dianggap tidak berpengaruh pada proses klasifikasi yaitu variabel *Race* dan *Differentiate.*

2.3.2 *Data Cleaning*

Pada pembersihan data awal dilakukan tahap pemeriksaan data yang tidak konstan atau memperbaiki kesalahan pada data. Dalam data kasus penderita kanker payudara terdapat beberapa data yang kosong (*missing value*) dimana data tersebut tidak sesuai dengan data lainnya (*inkonsisten*), pada penelitian ini ditemukan data *inkonsisten* pada variabel *marital status*, maka pada proses *cleaning* data tersebut akan dihilangkan. Sehingga diperoleh *record data* sebanyak 3979 dari data sebelumnya yang berjumlah sebanyak 4024 *record data.* Data tersebut sudah siap untuk digunakan untuk proses klasifikasi, sebagaimana ditunjukan pada gambar potongan dataset kanker payudara berikut :

Age	Marital Status	T Stage	N Stage	6th Stage	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Regional Node Positive	Survival Months	Status
68	Married	T1	N1	IIA	3	Regional	4	Positive	Positive	24	1	60	Alive
50	Married	T2	N2	IIIA	2	Regional	35	Positive	Positive	14	5	62	Alive
58	Divorced	T3	N3	IIIC	2	Regional	63	Positive	Positive	14	7	75	Alive
58	Married	T1	N1	IIA	3	Regional	18	Positive	Positive	2	1	84	Alive
47	Married	T2	N1	IIB	3	Regional	41	Positive	Positive	3	1	50	Alive
51	Single	T1	N1	IIA	2	Regional	20	Positive	Positive	18	2	89	Alive
51	Married	T1	N1	IIA	1	Regional	8	Positive	Positive	11	1	54	Alive
40	Married	T2	N1	IIB	2	Regional	30	Positive	Positive	9	1	14	Dead
40	Divorced	T4	N3	IIIC	3	Regional	103	Positive	Positive	20	18	70	Alive
69	Married	T4	N3	IIIC	1	Distant	32	Positive	Positive	21	12	92	Alive

Gambar 4. Dataset Kanker Payudara

2.4 Klasifikasi *(Classification)*

Klasifikasi (*Classification)* yaitu kinerja proses pengelompokan dalam kaitannya dengan ciri – ciri tertentu. Wijaya & Ridwan dalam [10]. Teknik pohon keputusan (*Decision Tree*) digunakan dalam proses klasifikasi penelitian ini.

Pohon keputusan (*Decision Tree)* merupakan sebuah bagan alur yang serupa dengan bentuk pohon dengan masing – masing *internal nodes* (simpul dalam) menjadi simbol dari variabel yang akan dites dan *leaf nodes* (daun) yang mewakili kelas terpilih atau persebaran dari kelas. Han & Kamber dalam [10].

2.5 Algoritma CART (*Classification and Regression Trees*)

CART (*Classification and Regression Trees*) masuk kedalam salah satu algoritma dari teknik pengolahan data pohon keputusan (*Decision Tree*). Algoritma CART membangun pohon keputusan dengan menyeleksi percabangan yang paling optimum dari masing –masing node. Febti Eka dalam [8]. Algoritma CART merupakan algoritma dari pohon keputusan yang pasti dalam percabangannya selalu berjumlah dua atau bercabang biner. Leo Breiman, Jerome Friedman, Richard Olshen, dan Charles Stone merupakan para ahli yang pertama kali menemukan konsep perhitungan algoritma CART. Larose dalam [11]. Pohon yang terbentuk dengan menggunakan data training (*training set*) mungkin dapat menempatkan nilai kelas berdasarkan nilai variabel lain atau secara *independen* pada variabel target dari kumpulan data baru. Membangun pohon biner berdasarkan fungsi dari variabel input dengan memisahkan record pada setiap node dikenal sebagai algoritma CART.

Langkah – langkah dalam algoritma CART [11] terdiri dari :

a. Mempersiapkan calon cabang (*Candidate split*) merupakan langkah awal. Untuk menyusun daftar kandidat cabang mutakhir, persiapan menyeluruh dilakukan pada semua variabel prediktor.
b. Langkah kedua, menilai kinerja keseluruhan daftar kandidat cabang mutakhir dengan menghitung keseluruhan nilai kesesuaian.
c. Langkah terakhir, menyeleksi calon cabang yang akan digunakan, dengan menggunakan kriteria berdasarkan nilai kesesuaian tertinggi.

Pada tahap pengembangan pohon keputusan, nilai *impurity* digunakan saat pemilihan *split* atribut. Nilai *impurity* dapat ditentukan dengan menggunakan nilai *Gini Index*. Ningrat & Santosa dalam [10]. *Gini index* diperlukan dalam penentuan titik pembelah terbaik (*splitting optimal point*). Gorunescu dalam [6]. Semakin rendah nilai *Gini index* maka semakin tinggi ukuran kesamaannya. *Gini index* variabel *T* untuk dataset dengan *m class* diformulasikan [6], sebagai berikut :

$$Gini(T) = 1 - \sum_{i=1}^{m} p\frac{2}{i} \quad (1)$$

m merupakan jumlah variabel kelas dalam dataset, p_i merupakan frekuensi dari satu record dalam dataset yang mempunyai atribut kelas T_i, untuk perhitungannya dapat dilakukan dengan cara membagi jumlah variabel dalam *class* T_i dengan jumlah keseluruhan record dalam dataset [10]

Penggunaan indeks gini terendah untuk membagi dataset menjadi dua bagian. Ketika data dipecah terhadap A menjadi dua himpunan bagian, D1 dan D2. *gini index* dapat diformulasikan seperti dibawah ini [6].

$$Gini_A(D) = \frac{|D1|}{|D|} Gini(D1) + \frac{|D2|}{|D|} Gini(D2) \quad (2)$$

2.6 Algoritma C 4.5

Algoritma C 4.5 adalah hasil dari pengembangan algoritma sebelumnya yaitu algoritma *Iterative Dichotomiser* 3 (ID3). Quinlan dalam [12]. Proses pengembangan yang dilakukan pada algoritma C 4.5 menghasilkan beberapa perubahan yang lebih baik diantaranya dapat menangani masalah *missing values* (nilai yang hilang) dalam sebuah dataset, dapat menangani data kontinu (data yang memiliki kemungkinan nilai tidak terbatas) dan melakukan *Prunning* [9]. Algoritma C 4.5 mempunyai input berupa data latih (*training*) serta data sampel. Data latih yaitu data yang dipergunakan sebagai data ajar atau data contoh untuk pembuatan model pohon keputusan yang sebelumnya telah dilakukan pengujian validitasnya. kemudian data sampel adalah suatu dataset yang dipergunakan sebagai tolak ukur dalam model klasifikasi

[10]. Dalam pembuatan model pohon keputusan, algoritma C 4.5 menggunakan perhitungan Gain Ratio. Santosa dalam [13].

Tahapan dalam perhitungan algoritma C 4.5 [12] sebagai berikut :

a. Mencari nilai *entropy*.
b. Melakukan perhitungan nilai *gain*.
c. Setelah menghitung kedua nilai tersebut kemudian dibangunlah pohon keputusan (*Decision tree*).

Untuk menghitung nilai *entropy* dan *gain* maka akan digunakan rumus [14] sebagai berikut :

$$Entropy\ (s) = \sum_{i=1}^{n} - pi * log_2\ p_i \quad (3)$$

Keterangan :
S : Himpunan keseluruhan data
A : Variabel
n : Banyaknya bagian S
P_i : Rasio dari S_i atas S

Setelah melakukan perhitungan *entropy* pada setiap kasus, kemudian dilakukan perhitungan *information gain* dengan menggunakan rumus sebagai berikut :

$$Gain\ (S, A) = Entropy\ (s) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (s_i) \quad (4)$$

Keterangan :
S : Total keseluruhan data
A : Variabel
n : Banyaknya bagian variabel A
$|S_i|$: Banyaknya data pada bagian ke-i
|S| : Banyaknya data dalam S

Kemudian cari nilai *Gain ratio*. Untuk menghitung *Gain ratio*, sebelumnya perlu menghitung nilai *Split information* menggunakan rumus [8] sebagai berikut

$$SplitInformation\ (S, A) = - \sum_{i=1}^{n} \frac{Si}{S}\ Log2\ \frac{Si}{S} \quad (5)$$

Keterangan :
S = Ruang sample
A = Variabel
Si = Jumlah sample untuk variabel ke-i

Setelah menghitung nilai *Split Information*, selanjutnya menghitung nilai *Gain ratio* dengan rumus sebagai berikut :

$$Gain\ Ratio = -\frac{Gain\ (S,A)}{SplitInformation\ (S,A)} \quad (6)$$

2.7 *Confusion Matrix*

Confusion Matrix merupakan salah satu cara evaluasi model algoritma pada data mining dengan melakukan perhitungan nilai akurasi atau pengukuran performa dari model algoritma klasifikasi. Terdapat empat kombinasi berbeda yang dihasilkan dari nilai observasi dan nilai prediksi untuk menampilkan hasil dari proses teknik klasifikasi, yaitu *True Positive* (TP) menunjukkan bahwa observasi tersebut positif dan benar diprediksi positif. *False Negative* (FN) menunjukkan bahwa observasi positif namun diprediksi negatif. *False Positive* (FP) menunjukkan bahwa observasi negatif namun diprediksi positif, dan *True Negative* (TN) menunjukkan bahwa observasi negatif dan benar prediksi negatif. Gambar 3 menunjukkan *Confusion Matrix* dengan empat kombinasi tersebut (TP, FN, FP, dan TN) [15].

		Observation Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Gambar 5. *Confusion Matrix*

Dengan menggunakan keempat istilah ini, perhitungan *Accuracy* dilakukan untuk mengetahui kedekatan antara nilai yang diprediksi dan yang diamati, perhitungan *Error rate* dilakukan untuk mengetahui tingkat kesalahan prediksi dari model klasifikasi, lalu perhitungan *Recall* atau *sensitivity* dilakukan untuk menggambarkan keberhasilan model dalam mengambil ulang suatu informasi, kemudian perhitungan *Precision* dilakukan untuk membuktikan tingkat ketepatan atau keakuratan model klasifikasi dalam memberikan informasi yang diinginkan dengan hasil prediksi yang di keluarkan. Sharma dalam [15] dengan rumus perhitungan sebagai berikut :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Error\ rate = \frac{FP+FN}{TP+TN+FP+FN} \quad (10)$$

Keterangan :
TP = Jumlah record data pada observation values positive serta predicted values positive
TN = Jumlah record data pada observation values negative serta predicted values negative
FP = Jumlah record data pada observation values negative serta predicted values positive
FN = Jumlah record data pada observation values positive serta predicted values negative

3. HASIL DAN PEMBAHASAN

3.1 Implementasi Algoritma CART

Pemodelan dengan algoritma CART (*Classification And Regression Trees*) menggunakan bantuan *tools* bahasa pemrograman R dengan *software* Rstudio. Pada penelitian ini akan dilakukan pengujian untuk algoritma CART sebanyak 3 kali dengan masing – masing probabilitas sebagai berikut :

Tabel 1. Pembagian Dataset Pengujian dengan Rstudio

Uji Coba Ke	Jumlah *Record* Data		
	Probabilitas	*Training*	*Testing*
1	70 : 30	2813	1166
2	75 : 25	3003	976
3	80 : 20	3183	796

Pembuatan pohon keputusan dengan algoritma CART pada Rstudio menggunakan *package "rpart"*. Adapun tahapan klasifikasi yaitu :

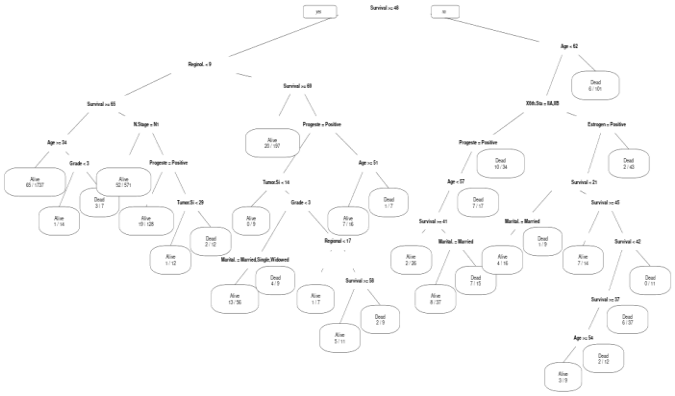
- Memuat data sampel
  data2 <- read.table("kanker.csv" , header = TRUE, sep = ",")
- Pembagian data sampel secara acak
  set.seed(110)
- *Splitting* data secara acak
  a <- sample (2, nrow(data2), replace = TRUE, prob = c(0.75, 0.25))
  train.data75 <- data2 [a ==1,]
  test.data25 <- data2 [a ==2,]
- *Install package* dan memuat *library rpart*
  Install.packages("rpart")
  library(rpart)
- Membuat model *decision tree* dengan mode *splitting gini index*
  Pohontrain1<-rpart(Status ~., data = train.data75, parms = list(split="gini"), cp=0.0003)
- Untuk melihat *summary* dari model yang dibuat
  summary(pohontrain1)
- Menampilkan *rule decision tree* bentuk *split node*
  print(pohontrain1)
- Memvisualisasikan model *decision tree*
  Install.packages("rpart.plot")
  library(rpart.plot)
  prp(pohontrain1 ,faclen = 9, cex = 0.5, extra = 0)
- Pengujian *accuracy* model *decision tree*
  Pohontest1<-predict(pohontrain1,test.data25,type = "class")
  Install.packages("caret")
  library(caret)
  confusionMatrix(pohontest1,test.data25$Status)

Dari 3 kali percobaan yang telah dilakukan dengan menggunakan algoritma CART didapatkan hasil nilai *Accuracy, Recall, Precision* dan *Error Rate* yang ditunjukan pada Tabel 3.

Tabel 2. Nilai Performa Algoritma CART

Probabilitas 70 : 30		Nilai Observasi		*Accuracy*	*Recall*	*Precision*	*Error Rate*
		Alive	*Dead*				
Nilai Prediksi	*Alive*	916	91	**87%**	**94%**	**91%**	**13%**
	Dead	58	101				
Probabilitas 75 : 25		Nilai Observasi		*Accuracy*	*Recall*	*Precision*	*Error Rate*
		Alive	*Dead*				
Nilai Prediksi	*Alive*	802	78	**90%**	**97%**	**91%**	**10%**
	Dead	22	74				
Probabilitas 80 : 20		Nilai Observasi		*Accuracy*	*Recall*	*Precision*	*Error Rate*
		Alive	*Dead*				
Nilai Prediksi	*Alive*	661	56	**91%**	**98%**	**92%**	**9%**
	Dead	12	67				
Rata – rata					**72%**		

Berdasarkan tabel 3 dapat kita ketahui bahwa performa terbaik pada algoritma CART ada pada percobaan 3 dengan probablitas dataset 80 : 20 dengan rata –rata performa algoritma CART yaitu sebesar 72 %. Untuk model pohon keputusan yang terbentuk ditunjukan pada Gambar 4.



Gambar 6. Pohon Keputusan Algoritma CART dengan Probabilitas 80:20

Berdasarkan pada gambar 4 didapatkan model pohon keputusan dengan variabel paling berpengaruh yaitu *Survival Months, Regional Node Positive, dan Age*. *Survival Months* menjadi *root node* pada model pohon keputusan ini, dimana apabila *Survival Months* pasien kanker payudara lebih besar atau sama dengan 48 bulan maka dapat dinyatakan pasien berstatus hidup dengan jumlah *record* terklasifikasi benar sebanyak 216 *record* dengan tetap memperhatikan variable data lainnya. Sedangkan pasien dengan *Survival Months* lebih kecil dari 48 bulan dinyatakan meninggal yaitu sebanyak 119 *record* terklasifikasi benar dengan tetap memperhatikan variabel data lainnya.

3.2 Implementasi Algoritma C 4.5

Pemodelan dengan algoritma C 4.5 menggunakan bantuan *tools* bahasa pemrograman R dengan *software* Rstudio. Pada penelitian ini akan dilakukan pengujian untuk algoritma C 4.5 sebanyak 3 kali dengan masing – masing probabilitas sebagai berikut :

Tabel 3. Pembagian Dataset Pengujian dengan Rstudio

Uji Coba Ke	Jumlah *Record* Data		
	Probabilitas	*Training*	*Testing*
1	70 : 30	2813	1166
2	75 : 25	3003	976
3	80 : 20	3183	796

Pembentukan pohon keputusan dengan algoritma C 4.5 pada Rstudio menggunakan *package "partykit"*. Adapun tahapan klasifikasi yaitu :

- Memuat data sampel
  Data2 <- read.table("kanker.csv" , header = TRUE, sep = ",")
- Pembagian data sampel secara acak
  set.seed(1234)
- *Splitting* data secara acak
  b <- sample (2, nrow(data2), replace = TRUE, prob = c(0.75, 0.25))
  traindata75 <- data2 [b ==1,]
  testdata25 <- data2 [b==2,]
- *Install package* dan mengaktifkan *library partykit*
  Install.packages("partykit")
  library(partykit)
- Pembentukan pohon keputusan
  Model1<-Status ~ Age + Marital.Status + T.Stage + N.Stage + X6th.Stage + Grade + A.Stage + Tumor.Size + Estrogen.Status + Progesterone.Status + Regional.Node.Examined + Reginol.Node.Positive +Survival.Months
  tree1<- ctree(Model1, data = traindata75)
- Menampilkan *rule* model pohon keputusan
  print(tree1)
- Memvisualisasikan *rule* dari model pohon keputusan dengan fungsi **plot grafik dari *partykit***

```
plot(tree1)
```

- Melakukan prediksi dengan data *testing*

```
testpred <-predict(tree1, newdata = testdata25)
table(testpred, testdata25$Status)
```

- pengujian akurasi model pohon keputusan

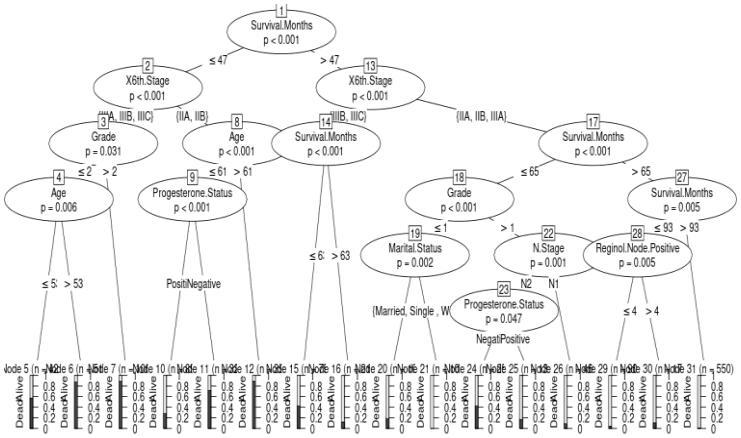
```
tabt1<-table(testpred,testdata25$Status, dnn =
        list("Prediction", "Reference"))
cmt1=confusionMatrix(tabt1)
cmt1
tabt1
akurasi1<-percent(sum(diag(tabt1)/sum(tabt1)))
akurasi1
```

Dari 3 kali percobaan yang telah dilakukan dengan menggunakan algoritma C 4.5 diperoleh hasil nilai *Accuracy*, *Recall*, *Precision* dan *Error Rate* yang ditunjukan pada Tabel 5.

Tabel 4. Nilai Performa Algoritma C 4.5

Probabilitas 70 : 30		Nilai Observasi		*Accuracy*	*Recall*	*Precision*	*Error Rate*
		Alive	*Dead*				
Nilai Prediksi	*Alive*	949	88	**90%**	**97%**	**92%**	**10%**
	Dead	25	104				
Probabilitas 75 : 25		Nilai Observasi		*Accuracy*	*Recall*	*Precision*	*Error Rate*
		Alive	*Dead*				
Nilai Prediksi	*Alive*	812	72	**91%**	**99%**	**92%**	**9%**
	Dead	12	80				
Probabilitas 80 : 20		Nilai Observasi		*Accuracy*	*Recall*	*Precision*	*Error Rate*
		Alive	*Dead*				
Nilai Prediksi	*Alive*	668	65	**91%**	**99%**	**91%**	**9%**
	Dead	5	58				
Rata – rata					**73%**		

Berdasarkan table 5 dapat kita ketahui bahwa performa terbaik pada algoritma C 4.5 ada pada percobaan 2 dengan probablitas dataset 75 : 25 dengan rata –rata performa algoritma C 4.5 yaitu sebesar 73 %. Untuk model pohon keputusan yang terbentuk ditunjukan pada Gambar 5.



Gambar 7. Model Pohon Keputusan Algoritma C 4.5 dengan Probabilitas 75 : 25

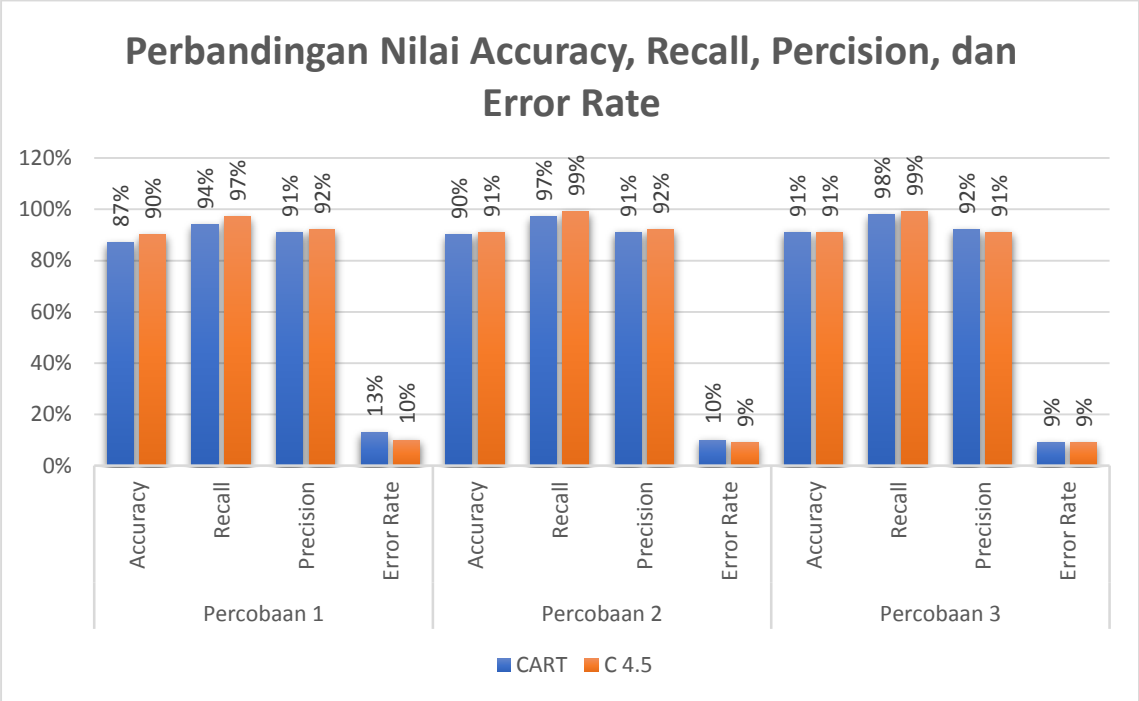
Berdasarkan pada gambar 5 didapatkan model pohon keputusan dengan variabel paling berpengaruh yaitu *Survival Months, 6th Stage* dan *Grade*. *Survival Months* menjadi *root node* dalam model pohon keputusan ini, dimana apabila *Survival Months* pasien kanker payudara lebih besar atau sama dengan 47 bulan maka peluang pasien dinyatakan hidup jauh lebih besar dengan jumlah *record* terklasifikasi benar sebesar 903 *record* pada *node* 29 dengan tetap memperhatikan variabel data lainnya. Sedangkan pasien dengan *Survival Months* lebih kecil dari 47 bulan maka memiliki peluang besar

dinyatakan meninggal dengan jumlah *record* terklasifikasi benar sebesar 100 *record* pada *node* 7 dengan tetap memperhatikan variabel data lainnya.

3.3 Analisis Perbandingan Algoritma CART dengan Algoritma C 4.5

Pada penelitian ini dilakukan 6 kali percobaan untuk melihat bagaimana pohon keputusan terbentuk dengan algoritma CART dan C 4.5, dimana dalam setiap percobaan dilakukan dengan probabilitas dataset masing – masing yaitu sebesar 70:30, 75:25 dan 80:20. Pada 6 kali percobaan ini menghasilkan bentuk dan model pohon keputusan yang beragam. Namun dapat disimpulkan bahwa variabel data yang paling berpengaruh dalam kasus klasifikasi pasien kanker payudara ditempati oleh variabel *Survival.Months.* Variabel ini muncul di setiap percobaan dan menjadi *root node* pada setiap percobaan yang dilakukan.

Untuk melihat perbandingan nilai *Accuracy*, *Recall*, *Precision* dan *Error Rate* berdasarkan uji coba



Gambar 8. Bagan Perbandingan Algoritma CART dengan Algoritma C 4.5

yang telah dilakukan dengan algoritma CART dan C 4.5 dapat dilihat pada Gambar 6.

Dari gambar 6 dapat dilihat bahwa performa dari algoritma C 4.5 cenderung lebih stabil dibandingkan dengan algoritma CART hal itu ditunjukan dengan presentase hasil yang didapatkan dalam setiap percobaan yang dilakukan.

4. KESIMPULAN

Berdasarkan analisis dan pembahasan dari serangkain percobaan yang telah dilaksanakan mengahasilkan kesimpulan seperti dibawah ini:

- Variabel yang paling berpengaruh pada algoritma CART yaitu *Survival Months, Regional Node Positive, dan Age.* Sedangkan pada algoritma C 4.5 variabel yang paling berpengaruh yaitu *Survival Months, 6th Stage* dan *Grade.* Ada satu kesamaan variabel dari kedua algoritma tersebut yaitu pada variabel *root node* dimana di tempati oleh variabel *Survival Months.*
- Untuk perbandingan tingkat performa dalam pengklasifikasian kasus kanker payudara, algoritma C 4.5 mendapatkan hasil nilai *Accuracy*, *Recall* , *Precision* dan *Error rate* yaitu sebesar 91%, 99%, 92%, 9%, dengan rata-rata performa algoritma yaitu 73% Sedangkan algoritma CART mendapatkan hasil nilai *Accuracy*, *Recall* , *Precision* dan *Error rate* yaitu sebesar 91%, 98%, 92%, 9%, dengan rata-rata performa

algoritma yaitu 72%. Dengan demikian dapat ditarik sebuah kesimpulan bahwa algoritma C 4.5 memperoleh tingkat performa yang lebih baik jika dibandingkan dengan algoritma CART. Selain itu jika dilihat dari nilai *accuracy* dari setiap percobaan, algoritma C 4.5 cenderung lebih stabil tingkat akurasinya dibandingkan dengan algoritma CART.

- Pada model tree yang terbentuk nilai *Accuracy*, *Recall* , *Precision* dan *Error rate* sangat bergantung dan dipengaruhi oleh komposisi *record* dalam setiap data *training* dan data *testing.*

UCAPAN TERIMAKASIH

Peneliti mengucapkan terimakasih kepada banyak pihak yang sudah membantu dan mendukung pelaksanaan kegiatan penelitian ini. Puji syukur peneliti panjatkan kepada kehadirat Allah SWT yang senantiasa memberikan pertolongannya sehingga menjadikan penelitian ini berjalan dengan lancar. Kepada kedua orang tua yang selalu memberikan motivasi serta semangat yang luar biasa, peneliti ucapkan terimakasih. Terimakasih kepada dosen pembimbing yang sudah bersedia membantu serta membimbing dalam menyelesaikan kegiatan penelitian ini, dan terakhir terimakasih kepada Univeritas Stikubank (UNISBANK) Semarang yang telah mendukung untuk penelitian ini. Harapannya peneliti maupun paraa pembaca dapat saling mengambil manfaat dan mendapatkan pengetahuan baru dari penelitian telah dilaksanakan.

REFERENCES

[1] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, "Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019, doi: 10.33480/pilar.v15i2.601.

[2] I. S. Syarief, "19,3 Juta Orang di Dunia Menderita Kanker, Paling Banyak Kanker Payudara," *www.suarasurabaya.net*, 2021. https://www.suarasurabaya.net/kelanakota/2021/193-juta-orang-di-dunia-menderita-kanker-paling-banyak-kanker-payudara/?amp

[3] I. Ramadhan and K. Kurniawati, "Data Mining untuk Klasifikasi Penderita Kanker Payudara Berdasarkan Data dari University Medical Center Menggunakan Algoritma Naïve Bayes," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, p. 21, 2020, doi: 10.30865/jurikom.v7i1.1755.

[4] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.

[5] V. Novalia, R. Goejantoro, and Sifriyani, "Perbandingan Metode Klasifikasi Naive Bayes dan K-Nearest Neighbor (Studi Kasus : Status Kerja Penduduk Di Kabupaten Kutai Kartanegara Tahun 2018)," *J. EKSPONENSIAL*, vol. 11, pp. 159–166, 2020.

[6] M. Yusa, E. Utami, and E. Luthfi. Taufiq, "Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3, C4.5, dan CART Pada Dataset Readmisi Pasien Diabetes," *Infosys (Information Syst. J.*, vol. 4, no. 1, pp. 23–34, 2016.

[7] D. Alverina, A. R. Chrismanto, and R. G. Santosa, "Perbandingan Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa," *J. Teknol. dan Sist. Komput.*, vol. 6, no. 2, pp. 76–83, 2018, doi: 10.14710/jtsiskom.6.2.2018.76-83.

[8] E. T. Novalyn, G. Ginting, and H. K. Siburian, "Pemanfaatan Metode Cart Dalam Memprediksi Omset Pakaian Pria Remaja (Studi Kasus : Pt. Matahari Department Store Thamrin Plaza Medan)," *J. Pelita Inform.*, vol. 7, no. 2, pp. 199–206, 2018.

[9] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017, doi: 10.15575/join.v2i1.71.

[10] A. Jananto, S. Sulastri, E. Nur Wahyudi, and S. Sunardi, "Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus Menggunakan Algoritma CART Klasifikasi Data Mining," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 10, no. 1, pp. 71–78, 2021, doi: 10.32736/sisfokom.v10i1.991.

[11] N. Nafi'iyah, "Algoritma Cart Dalam Penentuan Pohon Keputusan," vol. 7, no. 2, 2015.

[12] D. Rahmadani and A. A. Muzafar, "Comparative Analysis of C4 . 5 and CART Algorithms for Classification of Stroke Analisis Perbandingan Algoritma C4 . 5 dan CART untuk," pp. 198–206, 2022.

[13] R. W. Ningrat and B. Santosa, "Pemilihan Diet Nutrien bagi Penderita Hipertensi Menggunakan Metode Klasifikasi Decision Tree (Studi Kasus: RSUD Syarifah Ambami Rato Ebu Bangkalan)," *J. Tek. Its*, vol. VOL.1, no. 1, pp. 536–539, 2012.

[14] Kusrini and E. T. Luthfi, *Algoritma Data Mining*, 1st ed. Yogyakarta: C.V ANDI OFFSET, 2009.

[15] N. P. Wong, F. N. S. Damanik, C. -, E. S. Jaya, and R. Rajaya, "Perbandingan Algoritma C4.5 dan Classification and Regression Tree (CART) Dalam Menyeleksi Calon Karyawan," *J. SIFO Mikroskil*, vol. 20, no. 1, pp. 11–18, 2019, doi: 10.55601/jsm.v20i1.622.