p-ISSN: 1979-5661
e-ISSN: 2622-321X

Klasifikasi Kebakaran Hutan Menggunakan Algoritma C4.5 dan Rough Set

Arif Budiman[a1]
[a]Fakultas Ekonomi, Universitas Muhammadiyah Berau
Tanjung Redeb, Indonesia
[1]arif_budiman@umberau.ac.id

Abstract

In recent years there have been large-scale forest fires in forested areas of the world. Forest fires are a major environmental problem that has big impact on wildlife, human health, economic. One solution can be taken is using classification algorithm to predict forest fires based on historical forest fire data.

In this research using C4.5 Algorithm combined with Rough Set as feature selection to classify forests fire. Evaluate performance based on created model using confusion matrix to calculate accuracy value.

The results show the C4.5 algorithm with Rough Set as feature selection was found accuracy 98.36%. The use of Rough Set as feature selection can reduce irrelevant attributes effectively.

Keywords: *classification,forest fire, c4.5 algoritm, rough set*

Abstrak

Beberapa tahun terakhir telah terjadi kebakaran hutan dalam skala yang luas di kawasan berhutan di seluruh dunia. Kebakaran hutan merupakan masalah lingkungan utama yang miliki dampak yang besar terhadap kesehatan manusia, ekosistem satwa liar dan kondisi ekonomi oleh karena sebab itu perlunya dilakukan suatu pecegahan agar masalah tersebut dapat diatasi. Salah satu langkah dapat dilakukan yaitu menggunakan algoritma klasifikasi untuk melakukan prediksi kebakaran hutan berdasarkan riwayat data insiden kebakaran hutan.

Penelitian ini menggunakan Algorima C4.5 yang dikombinasikan dengan Rough Set sebagai seleksi fitur untuk klasifikasi kebakaran hutan. Mengevaluasi performa terhadap model yang telah dibuat digunakan confusion matrix untuk menghitung nilai akurasi.

Hasil pengujian menunjukan algoritma C4.5 dengan Rough Set sebagai seleksi fitur menghasilkan nilai akurasi 98.36%. Penggunaan Rough Set sebagai seleksi mampu mengurangi atribut yang tidak relevan secara efektif.

Kata kunci*: klasifikasi, kebakaran hutan, algorima c4.5, rough set*

1. PENDAHULUAN

Beberapa tahun terakhir telah terjadi kebakaran hutan dalam skala yang luas di kawasan berhutan di seluruh dunia[1]. Kebakaran hutan merupakan masalah lingkungan utama yang miliki dampak yang besar terhadap kesehatan manusia, ekosistem satwa liar dan kondisi ekonomi[2]. Kebakaran hutan juga sering terjadi karena pembukaan lahan liar dengan cara membakar hutan untuk digunakan sebagai lahan perkebunan atau pertanian [3].

Mengingat dampak yang dapat ditimbulkan akibat kebakaran hutan sangat besar apalagi hal tersebut dapat diperparah saat memasuki musim kemarau maka perlulah dilakukan suatu pecegahan agar masalah tersebut dapat diatasi. Salah satu langkah dapat dilakukan yaitu menggunakan algoritma klasifikasi untuk melakukan prediksi kebakaran hutan berdasarkan riwayat data insiden kebakaran hutan[1][3].

Algoritma C4.5 adalah algoritma klasifikasi yang digunakan untuk menghasilkan pohon keputusan. Pohon keputusan yang terbentuk dapat digunakan untuk prediksi, misalnya memprediksi minat dari peserta didik dalam menentukan jenis sekolah[4]. Algoritma C4.5 terbukti menunjukan tingkat akurasi paling baik dibandingkan algoritma klasifikasi lainnya seperti naive

p-ISSN: 1979-5661
e-ISSN: 2622-321X

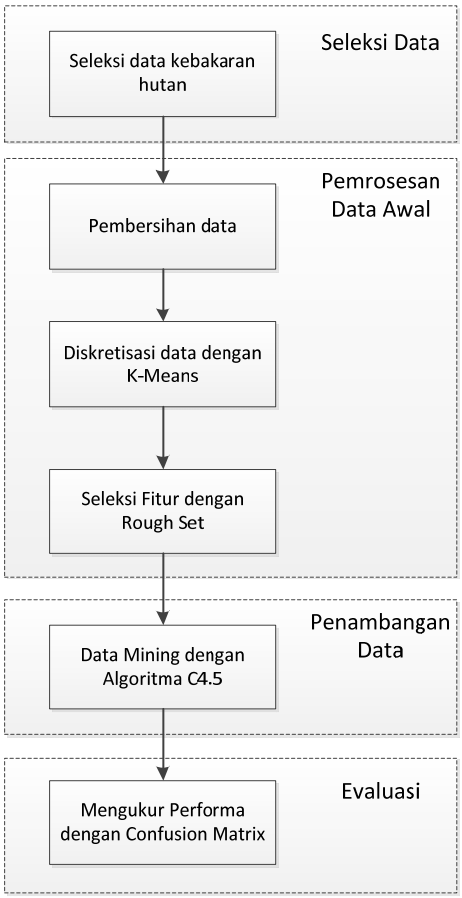
bayes dan neural network[5]. Untuk dapat memperoleh hasil klasifikasi yang akurat membutuhkan preprocessing yang baik. Metode diskretisasi data dan seleksi fitur dipilih karena memiliki peran yang sangat penting didalam tahapan preprocessing data[6]

. Diskritisasi data adalah pengelompokan atribut yang memiliki nilai kontinyu dengan cara membagi rentang tersebut menjadi satu set interval terbatas secara terpisah dan kemudian diasosiasikan dengan label, untuk disriktisasi data dalam penelitian ini mengunakan k-means untuk mengelompokan atribut yang memiliki data kontinyu[7][8]. Seleksi fitur digunakan untuk memperoleh atribut optimal yang akan digunakan untuk klasifikasi dengan cara mengurangi atribut yang tidak relevan terhadap kelas. Metode seleksi fitur berkontribusi meningkatkan algoritma klasifikasi[6]. Seleksi fitur menggunakan Rough Set dapat meningkatkan nilai akurasi dari algoritma klasifikasi yang digunakan[9]

Berdasarkan masalah tersebut penulis menggunakan Algorima C4.5 untuk klasifikasi kebakaran hutan yang dikombinasikan dengan Rough Set sebagai seleksi fitur.

2. METODE PENELITIAN

Penelitian ini menggunakan algoritma C4.5 yang dikombinasikan dengan Rough Set sebagai seleksi fitur. Didalam tahapan preprocessing data digunakan algoritma K-means untuk merubah data bertipe kontinyu menjadi data diskrit. Untuk mengevaluasi model yang telah dibuat digunakan confusion martix untuk menghitung tingkat akurasi. Alur penelitian ini ditampilkan pada gambar 1



Gambar 1. Alur Penelitian

p-ISSN: 1979-5661
e-ISSN: 2622-321X

2.1. Pembersihan Data

Pembersihan data adalah tahapan membuang data berganda agar tidak adanya duplikasi data, membuang data yang tidak konsisten, dan memperbaiki kesalahan penulisan yang ditemukan pada data.

2.2. Diskritisasi Data

Diskritisasi data adalah metode yang digunakan untuk mengurangi jumlah nilai yang berbeda untuk atribut yang memiliki nilai kontinyu dan membagi rentang tersebut menjadi satu set interval terbatas secara terpisah dan kemudian mengasosiasikan interval ini dengan label, sehingga mengurangi kebutuhan memori sistem dan meningkatkan efisiensi algoritma[10]. Pada penenlitian ini menggunakan k-means untuk mengelompokan data kontinyu menjadi data diskrit.

K-Means adalah metode klasterisasi data berbasis partisi yang menggunakan k sebagai jumlah dari klaster data. metode ini membagi data menjadi beberapa kelompok yang memiliki suatu nilai kedekatan. Klasterisasi data menggunakan k-means dengan cara perhitungan secara terus-menerus terhadap pusat centroid pada masing-masing klaster sampai tidak ada perubahan data yang terjadi.[7][8]. Menghitung nilai centroid pada masing-masing klaster dengan persamaan (1)

$$c_i = \min + \frac{(i-1)*(\max - \min)}{n} + \frac{(\max - \min)}{2*n} \quad (1)$$

Keterangan
ci : centoroid dari class i
max : nilai tertinggi dari class data kontinyu
min : nilai terendah dari class data kontinyu
n : jumlah dari kelas diskrit

selanjutnya dilakukan perhitungan jarak menggunakan euclidean distance. Hasil pengelompokan data berdasarkan jarak terpendek antara data dan centroid yang dikelompokan pada masing-masing klaster. Untuk menghitung euclidean distance menggunakan persamaan 2

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{ij}\right)^2} \quad (2)$$

Keterangan
dij : jarak antara objek i dan objek j
p : dimensi data
xik : kordinat dari objek i dalam dimensi k
xij : kordinat dari objek j dalam dimensi k

Apabila masih terdapat perpindahan data dalam klaster atau nilai centroid yang berubah maka perlu dilakukan perhitung kembali hingga tidak adanya perubahan data atau menghasilkan data yang konvergen.

2.3. Seleksi Fitur

Seleksi fitur sering digunakan dalam tahap preprocessing data. metode ini mengidentifikasi relevansi dari atribut yang digunakan terhadap class data dan membuang semua atribut yang tidak relevan. Tujuan seleksi fitur adalah mengurangi dimensi data agar algoritma data mining dapat berjalan lebih cepat[8][11]. Pada penelitian ini menggunakan Rough Set untuk memilih atribut terbaik yang akan digunakan untuk klasifikasi.

Rough Set telah banyak diterapkan dalam berbagai bidang seperti untuk seleksi fitur, klasifikasi, pencarian pola [12]. Data didalam Rough Set direpresentasikan kedalam bentuk tabel, dimana baris dalam tabel merepresentasikan objek dan kolom merepresentasikan atribut dari

p-ISSN: 1979-5661
e-ISSN: 2622-321X

objek tersebut. Tabel yang terbentuk dinamakan information system. Information system digambarkan menggunakan persamaan 3

$$IS = (U, A) \quad (3)$$

Keterangan :
U : Universe
A : Atribut

Setelah information system terbentuk, selanjutnya dilakukan perhitungan Discernibility Matrix untuk membandingkan isi sebuah atribut antara suatu objek dengan objek lainnya. Dalam perbandingan data apabila memiliki nilai yang sama maka tidak akan menghasilkan nilai, sebaliknya jika memiliki nilai yang berbeda akan menghasilkan suatu nilai. Langkah terakhir adalah melakukan reduksi untuk menseleksi atribut minimal dari sekumpulan atribut kondisi dengan menggunakan Prime Implicant fungsi Boolean. Kumpulan dari Prime Implicant tersebut kemudian menghasilkan set reduksi yang akan digunakan sebagai seleksi fitur[12]

2.4. Penambangan Data

Algoritma C4.5 merupakan pengembangan dari alogritma ID3, namun yang membedakannya dengan ID3 adalah penggunaan gain ratio untuk melakukan pemisahan bukan menggunakan information gain seperti pada algoritma ID3. Algoritma C4.5 digunakan untuk membentuk pohon keputusan. Dalam Pohon keputusan yang terbentuk terdiri dari akar, cabang dan daun[7][10]. Cara kerja pohon keputusan adalah dengan melakukan penelusuran dari akar menuju cabang hingga menemukan class dari objek tersebut [5][8]. Langkah awal yang dilakukan adalah menghitung nilai entropy dari dataset kebakaran hutan menggunakan permasaan 4.

$$Entropy(S) = \sum_{i=1}^{n} -p_i * \log_2 p_i \quad (4)$$

Keterangan :
S : himpuanan kasus
n : jumlah partisi s
pi : proporsi S_i terhadap S

Setelah nilai entropi ditemukan, Selanjutnya dilakukan perhitungan nilai information gain dari dataset kebakaran hutan menggunakan persamaan 5.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (5)$$

Keterangan :
n : jumlah partisi atribut
$|S_i|$: proporsi S_i terhadap S
|S| : jumlah kasus dalam S

Nilai information gain yang telah terbentuk selanjutnya digunakan untuk mencari nilai split information dengan menggunakan persamaan 6

$$SplitInformation(S, A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \log_2 \frac{|S_i|}{|S|} \quad (6)$$

Perhitungan terakhir untuk mencari nilai gain ratio dengan cara membagi information gain dengan split information. nilai gain ratio yang telah terbentuk kemudian dibandingan untuk mencari nilai gain ratio tertinggi yang akan digunakan sebagai akar. gain ratio dihitung menggunakan persamaan 7.

p-ISSN: 1979-5661
e-ISSN: 2622-321X

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (7)$$

Setelah akar ditemukan selanjutnya, proses perhitungan kemudian dilanjutkan untuk mencari cabang dan daun hingga menghasilkan sebuah pohon keputusan

2.5. Evaluasi

Dalam penelitian ini menggunakan confusion matrix untuk evaluasi performa dari model yang telah dibuat. Nilai pada Confusion matrix deperolah berdasarkan hasil perbandingan data aktul dan data prediksi yang memiliki nilai benar ataupun salah. Hasil evaluasi confusion matrix ditampilkan pada tabel 1

Tabel 1. Confusion Matrix

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

berdasarkan tabel confussion matrix yang terbentuk selanjutnya dapat dilakukan perhitungan akurasi terhadap algoritma klasifikasi yang digunakan dengan menggunakan persamaan 8

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (8)$$

Keterangan :
TP (True Positive) : Jumlah class aktual positif yang diprediksi benar
TN (True Negative) : Jumlah class aktual negatif yang diprediksi benar
FP (False Positive) : Jumlah class aktual negatif yang diprediksi salah
FN (False Negative) : Jumlah class aktual positif yang diprediksi salah

3. HASIL DAN PEMBAHASAN

3.1. Seleksi Data

Dalam penelitian ini menggunakan dataset kebakaran hutan yang diperoleh dari UCI Machine Learning Respository dengan nama “Algerian Forest Fire”, dataset yang dikumpulkan terdiri dari 244 record dengan 11 kategori data yaitu, Temprature (Temp), Relative Humidity (RH), Wind Speed (WS), Rain, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), Fire Weather Index (FWI), data tersebut dikelompokan menjadi 2 kategori yaitu fire dan not fire. Berikut adalah tabel deksripsi dataset kebakaran hutan ditampilkan pada tabel 2

Tabel 2. Deskripsi Dataset

Nama Atribut	Rentang Nilai
Temprature (C°)	22-42
Relative Humidity (%)	21-90
Wind Speed (km/jam)	6-29
Rain (mm)	0-16.8
Fine Fuel Moisture Code	28.6-96
Duff Moisture Code	0.7-65.9
Drought Code	6.9-220.4
Initial Spread Index	0-19
Buildup Index	1.1-68
Fire Weather Index	0-31.1
Kelas	Fire atau Not Fire

p-ISSN: 1979-5661
e-ISSN: 2622-321X

3.2. Pemrosesan Data Awal

3.2.1. Pembersihan Data

Pembersihan data perlu dilakukan karena masih terdapat data yang tidak lengkap dan terdapat kesalahan dalam penulisan data. Untuk mendapatkan hasil akurasi yang lebih baik dalam proses klasifikasi maka diperlukan penanganan pada nilai atribut yang tidak ada. penulis menggunakan nilai rata untuk menangani missing value pada data yang akan digunakan.

3.2.2. Diskritisasi data

Diskritisasi data dilakukan untuk merubah data kontinyu menjadi data diskrit agar dapat digunakan nantinya saat proses data mining. Diskretisasi dalam penelitian ini menggunakan k-means dimana tiap data kontinyu pada masing-masing atribut dibagi sebanyak 3 klaster. Berikut adalah dataset kebakaran hutan sebelum proses diskritisasi menggunakan k-means seperti ditampilkan pada tabel 3

Tabel 3. Dataset sebelum proses diskritisasi

Temp	RH	WS	Rain	....	BUI	FWI
29	57	18	0	....	3.4	0.5
29	61	13	1.3	....	3.9	0.4
26	82	22	13.1	....	2.7	0.1
25	89	13	2.5	....	1.7	0
....n	....	....	....	....	....	....
24	64	15	0.2	....	4.8	0.5

Datset sebelum proses diskritisasi data memiliki nilai yang beragam pada masing-masing atribut. Hal ini dapat menyulitkan disaat proses klasifikasi menggunakan algoritma C4.5 nantinya oleh karena itu perlu dilakukan penyederhanaan data menggunkan k-means. Berikut ini adalah dataset kebakaran hutan setelah proses diskritisasi menggunakan k-means seperti yang ditunjukan pada tabel 4

Tabel 4. Dataset setelah proses diskritisasi

Temp	RH	WS	Rain	....	BUI	FWI
Cluster2	Cluster1	Cluster1	Cluster0	....	Cluster2	Cluster2
Cluster2	Cluster1	Cluster2	Cluster0	....	Cluster2	Cluster2
Cluster2	Cluster2	Cluster0	Cluster2	....	Cluster2	Cluster2
Cluster2	Cluster2	Cluster2	Cluster1	....	Cluster2	Cluster2
....n	....	....	....	....	....	....
Cluster2	Cluster1	Cluster2	Cluster0	....	Cluster2	Cluster2

Dataset tersebut telah diolah menggunakan k-means dengan jumlah k=3 dimana tiap data memiliki kedekan dengan pusat centroid pada masing-masing klaster.

3.2.3. Seleksi Fitur

Seleksi fitur dalam penelitian ini menggunakan Rough Set dimana dataset yang terbentuk setelah proses diskritisasi data direpresentasikan menggunakan persamaan $U = \{x1, x2, ..., xm\}$ dan $A = \{a1, a2, \ldots,\}$ untuk menghasilkan sebuah tabel information system untuk menggambarkan objek dan atribut. Information system ditampilkan pada tabel 5

Tabel 5. Information system

U	*a1*	*a2*	*a3*	*a4*	*am*	*a9*	*a10*
x1	2	1	1	0	....	2	2
x2	2	1	2	0	....	2	2
X3	2	2	0	2	....	2	2
X4	2	2	2	1	....	2	2
Xm	....	....	....	....	....	....	....
x245	2	1	2	0	....	2	2

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Setelah tabel Information system terbentuk. Selanjutnya dilakukan perbandingan antara objek menggunakan discernibility matrix. Tabel discernibility matrix ditampilkan pada tabel 6

Tabel 6. Discernibility matrix

	x1	*x2*	*x3*	*x4*	*xm*	*x244*
x1	0	e	b,c,d,e	b,c,d,e	....	c
x2	e	0	b,c,d,e	b,d,e	....	0
x3	b,c,d,e	b,c,d,e	0	c,d	....	b,c,d,e
x4	b,c,d,e	b,d,e	c,d	0	....	b,d,e
xm	....	....	....	....	....	....
x244	c	0	b,c,d,e	b,d,e	....	0

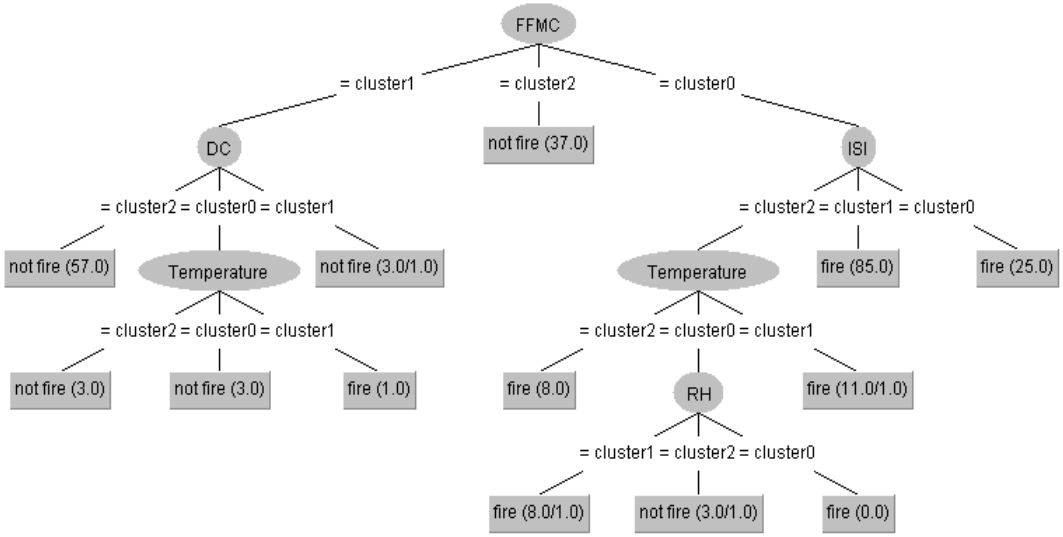
Setelah didapatkan hasil dari Discernibility matrix, tahapan selanjutnya adalah melakukan reduksi atribut. Proses reduksi digunakan untuk memilih atribut yang akan digunakan sebagai fitur pilihan. Berikut adalah beberapa kombinasi fitur terpilih yang ditampilkan pada tabel 7

Tabel 7. Seleksi Fitur

No	Fitur terpilih	Jumlah Atribut
1	Temperature, RH, WS, FFMC, DC, ISI	6
2	Temperature, RH, FFMC, DC, ISI	5
3	Temperature, RH, WS, FFMC, ISI, BUI	6
4	Temperature, WS, FFMC, ISI, BUI	5
5	Temperature, RH, FFMC, DMC, DC, ISI	6

3.3. Penambangan Data

Untuk dapat menghasilkan prediksi, algoritma C4.5 digunakan untuk menghasilkan pohon keputusan. Untuk memudahkan analisa data kebakaran hutan digunakan alat bantu WEKA. Berikut adalah pohon keputusan yang terbentuk berdasarkan hasil seleksi atribut Temperature, RH, WS, FFMC, DC, ISI yang ditampilkan pada gambar 2.



Gambar 2. Pohon keputusan

Pohon keputusan yang diperoleh tersebut kemudian dapat digunakan untuk memprediksi kebakaran hutan.

3.4. Evaluasi

Setelah pohon keputusan terbentuk. Selanjutnya dilakukan pengujian menggunakan confusion matrix untuk mengetahui ketepatan klasifikasi berdasarkan model yang telah terbentuk. Evaluasi confusion matrix ditunjukan pada tabel 8

p-ISSN: 1979-5661
e-ISSN: 2622-321X

Tabel 8. Confusion Matrix

Klasifikasi	Fire	Not Fire
Fire	136	2
Not Fire	4	102

Berikut adalah perbandingan algoritma C4.5 standar dan Algoritma C4.5 dengan menggunakan Rough Set untuk seleksi fitur. Berikut adalah perbandingan algoritma ditampilkan pada tabel 9

Tabel 9. Perbandingan algoritma

Algoritma	Akurasi
C4.5	98.36%
C4.5 dan Rough Set	98.36%

4. KESIMPULAN

Berdasarkan hasil pengujian menggunakan algorima C4.5 dengan seleksi fitur menggunakan Rough Set untuk klasifikasi kebakaran hutan. Hasil penelitian menunjukan algoritma C4.5 dan Rough Set menghasilkan nilai akurasi 98.36%. Penggunaan Rough Set sebagai seleksi fitur tidak menghasilkan peningkatan nilai akurasi, namun mampu mengurangi atribut yang tidak relevan secara efektif.

.

References

[1] D. A. Wood, “Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight,” *Artif. Intell. Agric.*, vol. 5, pp. 24–42, 2021.

[2] F. U. Robert Kurniawan, “PERBANDINGAN ALGORITMA LSDBC DAN DBSCAN PADA PEMETAAN DAERAH RAWAN KEBAKARAN HUTAN (Studi Kasus di Pulau Sumatera, Kalimantan, Sulawesi, dan Papua),” *J. Apl. Stat. Komputasi Stat.*, vol. 12.2.2020, no. 2086–4132, pp. 25–30, 2020.

[3] D. F. Pramesti, Lahan, M. Tanzil Furqon, and C. Dewi, “Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017.

[4] S. Narulita, A. T. Oktaga, and I. Susanti, “PENGUJIAN AKURASI MODEL PREDIKSI MENGGUNAKAN METODE DATA MINING CLASSIFICATION DECISION TREE ALGORITMA C4 . 5,” vol. 1, pp. 15–29, 2021.

[5] A. Mukminin and D. Riana, “Komparasi Algoritma C4 . 5 , Naïve Bayes Dan Neural Network Untuk Klasifikasi Tanah,” vol. 4, no. 1, pp. 21–31, 2017.

[6] N. Cahyani and M. A. Muslim, “Increasing Accuracy of C4.5 Algorithm by Applying Discretization and Correlation-based Feature Selection for Chronic Kidney Disease Diagnosis,” *J. Telecommun. Electron. Comput. Eng.*, vol. 12, no. April, pp. 25–32, 2020.

[7] A. Budiman, “Cronic Kidney Disease Prediction Using C4. 5 Algorithm and K-Means,” *JASIKA (Jurnal Apl. Sist. Inf. dan …*, vol. 1, no. 1, pp. 76–82, 2020.

[8] D. A. Effendy, K. Kusrini, and S. Sudarmawan, “Classification of intrusion detection system (IDS) based on computer network,” *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, pp. 90–94, 2018.

[9] A. Prabowo, “Analisis Akurasi Algoritma Naive Bayes Dengan Seleksi Fitur Rough Set Pada Klasifikasi Data,” 2021.

[10] T. B. Nugroho and E. Sugiharti, “The Improvement of C4 . 5 Algorithm Accuracy in Predicting Forest Fires Using Discretization and AdaBoost,” vol. 3, no. April, pp. 43–52, 2021.

[11] R. Rinawati, “Penentuan Penilaian Kredit Menggunakan Metode Naive Bayes Berbasis Particle Swarm Optimization,” *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 1, no. 1, p. 48, 2017.

[12] M. Rifai, S. Musdalifah, S. Matematika, and J. Matematika, “Klasifikasi pasien kanker payudara menggunakan metode rough set,” vol. 16, pp. 207–220, 2019.