Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Klasifikasi Penyakit Kanker Serviks Berdasarkan Kebiasaan dan Rekam Medis dengan Metode C4.5

Kemal Taufiq Hidayah [a,*], *Budi Arifitama* [b], *Silvester Dian Handy Permana* [c]

[a,b,c]*Program Studi Teknik Informatika Universitas Trilogi, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 04 Maret 2024
Revisi Akhir: 30 April2024
Diterbitkan *Online*: 02 Mei 2023=4

KATA KUNCI

Kanker Serviks,
Kebiasaan,
Tes Schiller,
Matriks Kebingungan,
Metode C4.5.

KORESPONDENSI

E-mail: Kemal.taufiq@trilogi.ac.id

A B S T R A C T

Kanker serviks adalah salah satu penyakit yang paling sering ditemui dan dapat menyebabkan kematian pada Wanita di seluruh dunia. Di Indonesia, jumlah kematian akibat kanker serviks terus meningkat setiap tahun, sebgaian besar disebabkan oleh diagnosis dan skrining yang terlambat. berbagai faktor yang disebabkan oleh kanker serviks seperti kebiasaan yang dilakukan ialah, berganti-ganti pasangan seksual, merokok atau pasif merokok, memiliki infeksi kelamin, memiliki riwayat kanker dan sebagainya. untuk mendeteksi adanya kanker serviks atau tidak, dapat dilakukan dengan cara pemeriksaan tes IVA (inspeksi visual asam asetat) atau yang disebut dengan tes *schiller*. Metode klasifikasi ialah bagian dari Teknik data mining untuk melakukan prediksi, Dalam penelitian ini, ingin meningkatkan akurasi dengan menggunakan metode C4.5 untuk melakukan klasifikasi penyakit kanker serviks berdasarkan kebiasaan pasien. Dua belas atribut dan satu atribut dari hasil pengujian digunakan dalam proses klasifikasi. Dataset tersebut terdiri dari 1080 entri, yang akan dibagi menjadi 864 data dan 216 data pelatihan. Data ini diperoleh dari website UCI *Repository*. Penelitian ini menghasilkan akurasi sebesar 94.10%, presisi sebesar 95.57%, *recall* sebesar 96.33% dan AUC (*Area Under Curve*) sebesar 0.987 yang diukur menggunakan *matrix confussion* atau matriks kebingungan dengan alat *rapidminer*.

1. PENDAHULUAN

Kanker leher Rahim disebut dengan kanker serviks, yang merupakan salah satu penyakit paling ditakuti oleh para Perempuan, penyakit ini menduduki peringkat 1 untuk kategori penyakit kanker di dunia dan diikuti oleh kanker payudara. Di Indonesia sendiri menempati urutan 2 kategori penyakit terbanyak bagi Wanita. Terdapat 132.000 kasus kanker di Indonesia dan mendapatkan jumlah penyebab penyakit kanker diposisi ke 8 se Asia Tenggara, dan total kasus kanker serviks di Asia mencapai lebih dari 150.000 kasus rerata 13,9 kematian sampai dengan 100.000 jiwa. Setiap tahun, terdapat 18.297 kasus kematian di Indonesia yang disebabkan oleh kanker serviks. Dengan kata lain, setiap hari ada 50 wanita Indonesia yang kehilangan nyawa mereka akibat penyakit ini[1].

Kanker serviks dimulai pada sel-sel dan pertumbuhan yang tidak teratur dan bagian sel yang tidak teratur ini, dapat menyebabkan melemahkan bagian dalam tubuh yang secara langsung pertumbuhan pada jaringan dan sel di tubuh (metastasis). Pertumbuhan yang tidak teratur dapat menyebabkan kerusakan DNA, yang menyebabkan mutasi pada gen dalam mengontrol pembelahan sel lainnya. Ketika sudah parah, sel tumbuh menjadi tumor ganas yang menyerang jaringan serviks. Penyebab utama kanker serviks adalah HPV(*human papillomavirus*) ada sejumlah faktor lain yang juga bisa mempengaruhi[2].

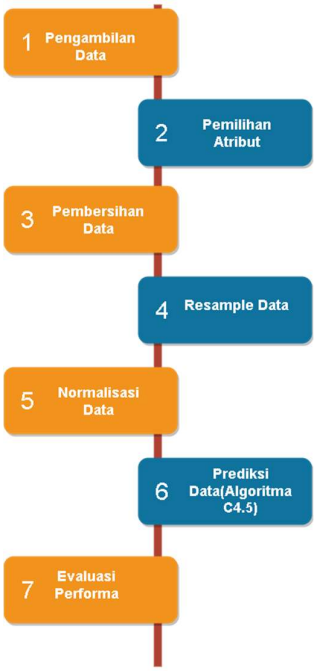
Kanker serviks dapat terdeteksi melalui berbagai jenis uji, termasuk uji *Hinselmann*, *Schiller*, *Citology*, dan *Biopsy*. *Hienselmann* yang dikenal sebagai kolposkopi, adalah prosedur kedokteran teruntuk mengamati pada bagian serviks, vagina, dan vulva menggunakan alat khusus yang disebut kolposkop. Tes *Schiller* pada dasarnya, adalah uji medis di mana larutan yodium

https://doi.org/10.25077/TEKNOSI.v10i1.2024.36-44

dioleskan ke serviks untuk mendiagnosis kanker serviks. Setelah pengolesan, perubahan warna jaringan menjadi coklat menunjukkan bahwa jaringan tersebut normal, sementara warna putih atau kuning menandakan ketidaknormalan. *Citology* yang dikenal sebagai *pap smear*, melibatkan pengambilan sampel sel pada leher rahim. Sampel jaringan sel yang ditangkap kemudian diperiksa di bawah mikroskop yang menentukan apakah jaringan sel tersebut normal, kanker dini, atau yang sudah bermutasi menjadi kanker. *Biopsy* adalah Tindakan bedah di mana beberapa kecil jaringan diambil dari serviks [3], [4].

Kanker serviks merupakan bentuk kanker yang muncul di leher rahim, yang merupakan bagian bawah rahim yang menghubungkan ke vagina. Penyakit ini disebabkan oleh infeksi virus *human papillomavirus* yang menular melalui hubungan seksual. Beberapa faktor resiko yang meningkatkan kemungkinan seseorang terkena kanker serviks termasuk aktivitas seksual yang dimulai pada usia muda, memiliki banyak pasangan seksual, kebiasaan merokok, dan sistem kekebalan tubuh lemah. Manfaat penelitian ini bagi pemangku kepentingan seperti Masyarakat umum, tenaga Kesehatan dan pemerintah, antara lain meningkatkan kesadaran akan pentingnya deteksi dini, mengurangi angka kematian akibat kanker serviks, meningkatkan aksesbilitas terhadap layanan Kesehatan yang berkaitan dengan kanker serviks, dan meningkatkan kualitas hidup pasien yang terkena kanker serviks melalui peyediaan perawatan dan dukungkan yang memadai.

Pohon keputusan atau yang disebut *decision tree*, sebagai unsur dari metode klasifikasi dalam data *mining*, mengadopsi metode klasifikasi yang menerapkan sistem wujud pohon. Pendekatan ini memiliki popularitas yang tinggi karena kemudahan interpretasinya oleh manusia. Data mining, sebagai suatu proses, bertujuan untuk mengeksplorasi ilmu dan penjelasan baru dari sejumlah besar data yang tersimpan di Gudang data[5], [6]. Dalam membangun model prediksi, algoritma C4.5 berperan penting dalam proses klasifikasi data yang telah dikategorikan. Algoritma ini memanfaatkan teknik pembelajaran *decision tree* untuk menghasilkan model klasifikasi yang efektif. C4.5 bekerja dengan menemukan atribut yang paling informatif dalam memisahkan data ke dalam kelas-kelas berbeda. Penentuan atribut ini dilakukan dengan metode "*Information Gain*" yang mengukur seberapa banyak informasi yang diberikan oleh atribut tersebut untuk memisahkan data ke dalam kelas-kelasnya.



Gambar 1. Metode Penelitian

Terdapat 4 penelitian terdahulu yang menjadi rujukan pada penelitian ini yaitu, penelitian[7] ini menyimpulkan bahwa menggunakan metode C.45 membuahkan hasil performa yang lebih baik dibandingkan *naïve bayes* dan *k-Nearest Neighbour* dengan mengasilkan akurasi sebesar 97% untuk C4.5 namun nilai performa pada naïve bayes 90% dan k-NN 95%. Penelitian [8] dengan mengklasifikasi penyakit kanker serviks Tingkat awal yang menyimpulkan bahwa algoritma C4.5 lebih baik dengan akurasi sebesar 98%. Penelitian [9] mengklasifikasi penyakit TBC dengan C4.5 menghasilkan akurasi sebesar 84,56% dan kurva AUC mendapatkan 0,938, kemudian penelitian [10] yang mengimplementasikan C4.5 untuk karakter siswa SD menghasilkan akurasi sebesar 90,08%.

Untuk Metode penelitian yang digunakan ini menggunakan dari penilitian [11] yang menerapkan teknik klasifikasi penyakit kanker serviks dengan menggunakan *Multilayer Percepton*, *BayesNet* dan *k-Nearest Neighbor*, penelitian tersebut berfokus pada 1 jenis atribut target yaitu tes *Biopsy*. Namun, penelitian ini akan menggunakan atribut *schiller* sebagai target dan dilakukan *resample* data pada atribut *schiller* lalu dilakukan klasifikasi menggunakan metode C4.5 yang berdasarkan penelitian

[7]diperoleh bahwa metode C4.5 lebih baik daripada menggunakan metode *naïve bayes* dan k-NN. Tujuan dari penelitian ini adalah mengetahui performa dari klasifikasi metode C4.5 pada penyakit kanker serviks.

2. METODE

Metode penelitian ini menjelaskan metode dan alur pada penelitian ini, salah satu yang menjadi tumpuan pada penerapan penelitian ini sampai menjadi terstruktur, berikut pada gambar 1 merupakan alur metode penelitian ini.

2.1. Pengambilan Data

Pada pengambilan data dalam Penelitian ini, menggunakan dataset yang digunakan oleh penelitian sebelumnya yaitu [12] Penelitian tersebut bertujuan untuk memprediksi penyakit kanker serviks bersandarkan kebiasaan pasien. Yang memakai data koresponden pasien dari rumah sakit, dan data sudah konfirmasi oleh para ahli. yang dilakukan adalah mengambil data dari Website UCI Machine Learning Repository yang dapat diunduh https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors yang berjudul "Cervical cancer (Risk Factors)" dengan berjumlah 858 dataset pasien yang terdiri dari 36 atribut pada Rumah sakit Universitas Caracas, Venezuela dan bentuk data yang didapati adalah tipe *file* .csv.

2.2. Pemilihan atribut

Setelah melakukan pengambilan data, proses selanjutnya memilih atribut, memilih atribut yang dikerjakan adalah menunjuk atribut yang tidak ambigu (isi nilai atribut yang tidak terlalu banyak missing values). Pada data ini terdapat empat target atribut untuk mendiagnosa kanker serviks yaitu hasil tes Biopsy, tes Citology, tes Hinselmann, dan tes Schiller. Penelitian ini menggunakan hasil tes Schiller yang diisi dua label positif (1) dan negatif (0) untuk target atribut dan dalam 13 atribut yang ada, 13 di antaranya dipilih untuk digunakan dalam tahap pembersihan data berikutnya. Berikut tabel 1 dibawah ini merupakan atribut yang dipakai

Tabel 1. Pemilihan atribut

Nama atribut	Tipe data	Jenis atribut
Age	*Integer*	*Regular*
Number of sexual partners	*Integer*	*Regular*
First Sexual intercourse	*Integer*	*Regular*
Num of Pregnancies	*Integer*	*Regular*
Smokes	*Boolean*	*Regular*
Smokes (years)	*Integer*	*Regular*
Smokes(packs/year)	*Integer*	*Regular*
Hormonal Contraceptives	*Boolean*	*Regular*
Hormonal Contraceptives (years)	*Integer*	*Regular*
IUD	*Boolean*	*Regular*
IUD (years)	*Integer*	*Regular*
STDs	*Boolean*	*Regular*
Schiller	*boolean*	*Target*

Berikut penjelasan pada setiap atribut di tabel 1 yang dipakai pada penelitian ini:

a. *Age*
   Atribut ini merupakan usia pada pasien, dengan tipe data *integer* yang Dimana satuan pada atribut ini adalah tahuncontoh: 16 tahun,32 tahun, dsb.
b. *Number of sexual partners*
   Atribut ini merupakan jumlah berapa kali melakukan hubungan seksual pada pasien, dengan tipe data *integer*. Pada atribut ini memiliki satuan yaitu orang contoh: 2 orang, 8 orang, dsb
c. *First sexual intercourse*
   Pada atribut ini menjelaskan bahwa pada umur berapa pasien melakukan pertamakalinya hubungan seksual, yang Dimana atribut ini tipe datanya adalah *integer* dengan satuan yaitu tahun contoh; 15 tahun,21 tahun, dsb.
d. *Num of pregnancies*
   Atribut ini merupakan jumlah kehamilan pada pasien yang bertipe data *integer* dan satuan pada atribut ini adalah kehamilan contoh; 0kehamilan, 3 kehamilan,dsb.
e. *Smokes*
   Atribut ini merupakan apakah pasien merokok atau tidak, pada tipe data atribut ini adalah *Boolean* yang berisikan dengan nilai 0 (tidak) dan 1 (iya)
f. *Smokes(years)*
   Pada atribut ini menjelaskan bahwa berapa tahun pasien sudah merokok, tipe data pada atribut ini adalah *integer* dan satuan pada atribut ini adalah tahun contoh; 0 tahun, 6 tahun dsb.
g. *Smokes(packs/years)*
   Atribut ini menjelaskan bahwasannya berapa jumlah bungkus rokok pertahunnya pada pasien dengan tipe datanya adalah *integer* dan satuan pada atribut ini adalah bungkus pertahun dengan sebagai contoh; 15 bungkus pertahun, 100 bungkus pertahun, dsb.
h. *Hormonal Contraceptives*
   Atribut ini menjelaskan bahwa apakah pasien pernah menggunakan alat hormonal kontrasepsi, pada atribut ini tipe datanya adalah *Boolean* yang berisikan 0(tidak) dan 1(iya).
i. *Hormonal Contraceptives (years)*
   Atribut ini menjelaskan berapa lama menggunakan alat kontrasepsi hormonal, pada atribut ini bertipekan *integer* dengan satuan tahun sebagai contoh; 1 tahun, 5 tahun, dsb.
j. IUD
   Atribut ini menjelaskan apakah pasien pernah menggunakan alat kontrasepsi dalam rahim, yang di mana atribut ini bertipekan *Boolean* dengan berisikan 0 (tidak) dan 1 (iya).
k. IUD (*years*)
   Atribut ini merupakan berapa lama pasien menggunakan alat kontrasepsi dalam rahim, atribut ini bertipekan *integer* dengan satuan tahun sebegai contoh; 2 tahun,0 tahun, dsb.
l. STDs
   Atribut ini menjelaskan apakah pasien memiliki infeksi saluran kelamin atau tidak, atribut ini tipe datanya adalah *Boolean* yang berisikan 0(tidak) dan 1 (iya).

https://doi.org/10.25077/TEKNOSI.v10i1.2024.36-44

m. *Schiller*
Atribut ini merupakan tes diagnosa kanker serviks pada pasien, yang di mana atribut ini sebagai target atribut. Dan tipe data pada atribut ini adalah *Boolean* yang berisikan 0(tidak) dan 1(iya)

2.3. Pembersihan Data

Proses pembersihan data melibatkan penghapusan beberapa baris yang memiliki nilai kosong [13]. Tidak lengkapnya data, akan menyebabkan hasil prediksi menjadi keambiguan. Setelah proses pembersihan, sebanyak 858 dan 13 atribut yang siap digunakan dalam prediksi.

2.4. Resample Data

Sebelum ke tahap proses memprediksi data, terdapat ketidakseimbangan pada atribut target *schiller*, hasil atribut tersebut mencakup 784 label negatif kanker (0) dan 74 label positif kanker (1). Karena ketidakseimbangan dalam data, penelitian ini menggunakan resampling data. Pada label positif akan di-resampling menjadi 296 data Dan total semua data menjadi 1.080 data. Berikut data yang sudah bersih dan diresample pada tabel 2 dibawah ini.

Tabel 2. *Resample* data

NO	***Age***	***Number of sexual partners***	***First sexual intercourse***	***Num of pregnancies***	***Smokes***	***…***	***Schiller***
1	18	4	15	1	0.0	…	1
2	15	1	14	1	0.0	…	1
3	34	1	17	1	0.0	…	1
4	52	5	16	4	1.0	..	1
….	….	…..	……	…..	…..	…	…..
1080	35	2	20	2	0.0	…	1

Pada tabel 2 di atas, melakukan *oversampling* pada atribut *schiller* dengan label (1) sebanyak 3 kali, proses *resample* ini berguna untuk menyeimbangkan pada data label di atribut target *schiller.*

2.5. Normalisasi Data

Langkah Selanjutnya dari penelitian ini adalah mengkategorikan data, dalam Langkah ini sangat membantu mengurangi ambiguitas dan ketidakpastian dalam data dengan mengelompokkan ke dalam kategori yang lebih spesifik dan terdefinisi. Dalam data yang diperoleh, setiap atribut data ini memiliki nilai yang sifatnya terlalu banyak numerik sehingga data ini diperlukan pengelompokkan agar lebih mudah dipahami [14], dan berikut ini dalam tabel 3 di bawah ini yang berisikan range nilai data pada setiap atributnya.

Tabel 3. Range nilai atribut

Nama Atribut	**Range Nilai atribut**
Age	16-80 Tahun
Number of Sexual partners	1-28 pasangan seksual
First Sexual Intercourse	10-32 tahun
Number of Pregnancies	0-11 kehamil
Smokes	0/1
Smokes (Years)	0-37 tahun
Smokes (pack/years)	0-37 bungkus per tahunnya.
Hormonal contraceptives	0/1
Hormonal contraceptives (years)	0 – 22 tahun
IUD	0/1
IUD (years)	0-19 tahun
STDs	0/1
Schiller	0/1

Pada tabel 3 di atas bahwa nilai range pada setiap atribut memerlukan pengelompokkan, oleh karena itu data tersebut dikelompokkan menjadi seperti berikut ini:

a. Age
Pada nama atribut *age* berubah menjadi "Umur", kemudian nilai dari atribut Umur dikelompokkan menjadi:
- Nilai < 16, maka menjadi "Remaja Awal"
- Nilai dari 17-25, maka menjadi "Remaja Akhir"
- Nilai dari 26-35, maka menjadi "Dewasa Awal"
- Nilai dari 36-45, maka menjadi "Dewasa Akhir"
- Nilai dari 45-55, maka mejadi "Lansia Awal"
- Nilai dari 56-65, maka menjadi "Lansia Akhir"
- Nilai > 66, maka Menjadi "Manula"

b. Number of Sexual partners
Pada nama atribut Number of Sexual berubah menjadi "Jumlah Pasangan Seksual", kemudian nilai dari Number of Sexual dikelompokkan menjadi:
- Nilai = 1, maka menjadi "Monogami"
- Nilai > 1, maka menjadi "Poliamori"

c. First Sexual Intercourse
Pada nama atribut First Sexual Intercouse berubah menjadi "Umur pertamakali hubungan seks", kemudian nilai-nilai dari First Sexual Intercouse dikelompokkkan menjadi:
- Nilai < 16, maka menjadi "Remaja Awal"
- Nilai dari 17-25, maka menjadi "Remaja Akhir"
- Nilai dari 26-35, maka menjadi "Dewasa Awal"
- Nilai dari 36-45, maka menjadi "Dewasa Akhir"
- Nilai dari 45-55, maka mejadi "Lansia Awal"
- Nilai dari 56-65, maka menjadi "Lansia Akhir"
- Nilai > 66, maka Menjadi "Manula"

d. Number of Pregnancies
Pada nama atribut Number of Pregnancies berubah menjadi "Jumlah Kehamilan", kemudian nilai dari smokes dikelompokkan menjadi:
- Nilai = 0, maka menjadi "Tidak Hamil"
- Nilai <= 4, maka menjadi "Normal"
- Nilai > 5, maka menjadi "Banyak"

e. Smokes
Pada nama atribut Smokes Berubah menjadi "merokok", nilai atribut dari smokes dikelompokkan menjadi:

- Nilai = 0, maka menjadi "Tidak"
- Nilai = 1, menjadi "Iya"

f. Smokes (pack/years)

Pada nama atribut *Smokes (Pack/years)* berubah menjadi "Jumlah bungkus rokok per tahun", nilai atribut dari "Smokes (Years/pack)" dikelompokkan menjadi:
- nilai = 0, maka "Tidak Merokok"
- nilai <= 10, maka "Sedikit"
- nilai <=20, maka "Cukup Banyak"
- nilai > 20, maka "Banyak"

g. Hormonal contraceptives

Pada nama atribut *hormonal contraceptives* berubah menjadi "alat kontrasepsi hormon", nilai atribut dari "*hormonal contraceptives*" dikelompokkan menjadi:
- nilai = 0, maka "Tidak Memakai"
- nilai =1, maka "Memakai"

h. Hormonal contraceptives (years)

Pada nama atribut *Hormonal contraceptives (years)* Berubah menjadi "lama pemakaian alat kontrasepsi hormon", Nilai-nilai atribut ini dikelompokkan menjadi:
- Nilai = 0, maka "Tidak Pernah"
- nilai <= 2, maka "Baru"
- nilai <= 5, maka "Menengah"
- nilai > 5, maka "Lama"

i. IUD

Pada nama atribut IUD berubah menjadi "alat kontrasepsi rahim", nilai atribut ini dikelompokkan mejadi:
- nilai = 0, maka "Tidak Memakai"
- nilai = 1, maka "Memakai"

j. IUD (years)

Pada nama atribut *IUD (years)* berubah menjadi "lama pemakaian alat kontrasepsi rahim", nilai-nilai atribut ini dikelompokkan menjadi:
- Nilai = 0, maka "Tidak Pernah"
- Nilai <= 2, maka "Baru"
- nilai <= 5, maka "Menengah"
- nilai > 5, maka "Lama"

k. STDs

Pada nama atribut *STDs* berubah mejadi "penyakit menular seksual", nilai atribut ini dikelompokkan menjadi:
- nilai = 0, maka "Tidak"
- nilai = 1, maka "Iya"

l. Schiller

Pada nilai dari atribut ini dikelompokkan menjadi:
- nilai = 0, maka "Negatif"
- nilai = 1, maka "Positif"

Hasil pengelompokkan data ini dapat dilihat pada lampiran di tabel i

2.6. Prediksi Data (Algoritma C4.5)

Prediksi pada data ini akan menggunakan metode C4.5. Atribut dipilih untuk dijadikan akar, dan cabang dibuat dari setiap titik awal, yaitu akar. Kasus dibagi disetiap akar dan proses ini diulangi sampai setiap cabang memiliki nilai kelas yang sama. Dalam algoritma C4.5, perhitungan melibatkan pengukuran nilai *entropy* dan *gain* untuk menentukan titik awal akar, akar dan daun[15]. Proses menghitung C4.5 dimulai dengan menghitung *entropy* dari masing-masing kelas dan atribut. Entropi merupakan ukuran ketidakpastian dari suatu data. Berikut adalah perhitungan rumus *entropy:*

$$E(S) = -\sum_{i=1}^{c} P_i log_2(P_i) \quad (1)$$

Setelah menghitung entropi, Langkah selanjutnya adalah menghitung perolehan informasi (*gain*) dari masing-masing atribut. *Gain* merupakan ukuran seberapa besar informasi yang dapat diperoleh dari suatu atribut untuk mengklasifikasikan data. Atribut dengan *gain* tertinggi akan menjadi akar pohon keputusan. Berikut adalah perhitungan rumus *gain*:

$$Gain = E(total) - \sum_{i=i}^{k}(\frac{Juml \quad data\ pada\ subset\ i}{total\ data}\ X\ E(Subset\ i)) \quad (2)$$

2.7. Evaluasi Performa

Pada tahapan evaluasi klasifikasi yang sudah dilakukan hasilnya akan digunakan untuk menguji dan mengukur keakuratan kinerja dari metode C4.5. Tahap evaluasi dapat dilihat menggunakan tabel *confusion matrix*, pada tahapan ini menggunakan evaluasi pada tools *rapidminer* yang menghitung *accuracy*, *precision recall* dan kurva ROC (*Receiver Operating Characteristics*) dengan nilai AUC (*Area Under Curve*), tabel tersebut akan memperlihatkan akurasi klasifikasi dengan membandingkan data yang sudah mempunyai kelas dan kelas prediksi berikut tabel 4 di bawah ini merupakan tabel confussion matrix.

Tabel 4. Matriks kebingungan

	Prediksi		
		Positif	Negatif
Data Testing	Positif	*True Positive*	*False Negative*
	Negatif	*False Positive*	*True Negative*

Tabel 4 merupakan *confusion matrix* 2x2 terdiri dari kelas yang terdapat pada data *testing* dan prediksi kelas. *Confusion matrix* akan memberikan penilaian tentang kualitas metode C4.5 yang telah dibuat. Pada confusion matrix akan memberikan juga informasi tentang TP (*True Positive*), FP (*False Positive*), TN (*True Negative*), dan FN (*False Negative*) yang dapat dihitung sebagai *accuracy, precision* dan *Recall*. Oleh karena itu, ini sangat penting karena biasanya hasil klasifikasi tidak dapat dijelaskan dengan baik hanya dengan satu perhitungan. Berikut adalah rumus *accuracy, precission* dan *recall*:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(Total\ Data)} x100\% \quad (3)$$

$$Precision = \frac{True\ Positive\ (TP)}{False\ Positive\ (FP) + True\ Positive\ (TP)} x100\% \quad (4)$$

$$Recall = \frac{Recall = Tru \quad Positive\ (TP)}{False\ Negative\ (FN) + True\ Positive\ (TP)} x100\% \quad (5)$$

https://doi.org/10.25077/TEKNOSI.v10i1.2024.36-44

3. HASIL DAN PEMBAHASAN

Berikut ini adalah penelitian terdahulu yang berkaitan dengan pentingnya menggunakan algoritma C4.5 untuk klasifikasi penyakit kanker. Pada penelitian [16] yang berjudul 'Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu Indonesia Pintar (KIP)' bertujuan untuk menyeleksi sistem penerimaan KIP di sekolah SMPN 38 Jakarta dengan total data 102 dan 9 variabel atribut memiliki Tingkat akurasi sebesar 90%. Selanjutnya dari penelitian [17] ini bertujuan mengklasifikasikan kanker seviks dengan menggunakan metode C4.5, *logistic function* dan *ZeroR* yang sebelumnya dilakukan *preprocessing* dengan *instance selection* dengan *naïve bayes.* Akurasi yang diperoleh sebesar 99,69% menggunakan metode C4.5 untuk *logistic function* dan *ZeroR* memperoleh akurasi sebesar 99,38% dan 95,20%.

Kemudian penelitian dari [18] ini bertujuan untuk seleksi penerimaan mitra penjualan yang memiliki dataset sebesar 107 dengan menggunakan metode C4.5 memiliki Tingkat akurasi sebesar 96% menggunakan hasil *ten-fold cross validation.*

Dan yang terakhir dari penelitian [19] yang bertujuan untuk klasifikasi prestasi belajar mahasiswa pada pandemi dengan menggunakan metode C4.5 didapatkan akurasi sebesar 97,5% dengan menggunakan *tools rapidminer.*

Pada hasil penelitian ini akan menjelaskan perhitungan C4.5 pada total *sample* data dengan total data 1080 dan menjelaskan performa penelitian ini menggunakan metode C4.5 dengan *tools rapidminer*.

3.1. Perhitungan C4.5

Langkah awal perhitungan ini mencari entropi total atribut *schiller* dengan berjumlah data 1080, negatif 784 dan positif 296 dengan persamaan (1) berikut tabel 5 adalah perhitungan entropi total data.

Tabel 5. Entropi total

	Jumlah	Negatif	Positif	ENTROPHY
Total	1080	784	296	0,8472501

Pada perhitungan tabel 5 di atas dapat menggunakan persamaan (1) untuk mencari entropi total dengan cara berikut:

Entropy Total = $-\sum_{i=1}^{c} P_i log_2(P_i)$

Entropy Total = $\left(\left(-\frac{784}{1080} log_2 * \left(\frac{784}{1080}\right)\right) + \left(\left(-\frac{296}{1080} log_2 * \left(\frac{296}{1080}\right)\right)\right)$

Entropy Total = 0,8472501

Kemudian setelah menghitung entropi totalnya, Langkah selanjutnya adalah menghitung entropi pada setiap atributnya. Di bawah ini adalah tabel 6 perhitungan entropi atribut umur.

Tabel 6. Entropi atribut umur

Atribut		Jumlah	Negatif	Positif	entropi
Umur	Remaja Awal	23	23	0	0
	Remaja Akhir	415	407	8	0,1373623
	Dewasa Awal	356	284	72	0,7263998
	Dewasa Akhir	260	204	56	0,7516499
	Lansia Awal	21	9	12	0,9852281
	Lansia Akhir	1	1	0	0
	Manula	4	4	0	0

Pada perhitungan tabel 6 menggunakan persamaan (1) untuk mencari sub entropi pada atribut umur

Entropy remaja awal = 0

Entropy remaja akhir = $\left(\left(-\frac{407}{415} log_2 * \left(\frac{407}{415}\right)\right) + \left(\left(-\frac{8}{415} log_2 * \left(\frac{8}{415}\right)\right) = 0{,}1373623\right.$

Entropy dewasa awal = $\left(\left(-\frac{284}{356} log_2 * \left(\frac{284}{356}\right)\right) + \left(\left(-\frac{72}{356} log_2 * \left(\frac{72}{356}\right)\right) = 0{,}7263998\right.$

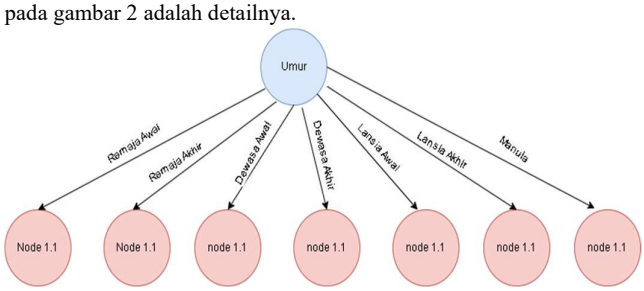
Entropy dewasa akhir = $\left(\left(-\frac{204}{260} log_2 * \left(\frac{204}{260}\right)\right) + \left(\left(-\frac{56}{260} log_2 * \left(\frac{56}{260}\right)\right) = 0{,}7516499\right.$

Entropy Lansia awal = $\left(\left(-\frac{9}{21} log_2 * \left(\frac{9}{21}\right)\right) + \left(\left(-\frac{12}{21} log_2 * \left(\frac{12}{21}\right)\right) = 0{,}9852281\right.$

Langkah selanjutnya adalah menghitung menghitung semua entropi dari atribut dan sub atribut dari atribut jumlah pasangan seksual sampai atribut penyakit menular seksual. Ketika perhitungan entropi total dan setiap entropi atribut selesai, langkah selanjutnya adalah menghitung nilai *gain* pada setiap atributnya dengan menggunakan persamaan (2). Berikut di bawah ini adalah perhitungan *gain* yang nilai *gain* paling tinggi:

a. *gain* umur = 0,8472501 - ((23/1080)*(0) – ((415/1080)* (0,1373623) – ((356/1080)*(0,7263998) – ((260/1080)*(0,7516499) – ((21/1080)*(0,9852281) – ((1/1080)*(0) – ((4/1080)*(0) = **0,354914537**

Setelah perhitungan entropy dan gain untuk semua data, simpul akar ditentukan berdsarkan gain tertinggi yang terlihat dari semua data. Nilai gain tertinggi ditemukan pada atribut umur. Berikut pada gambar 2 adalah detailnya.



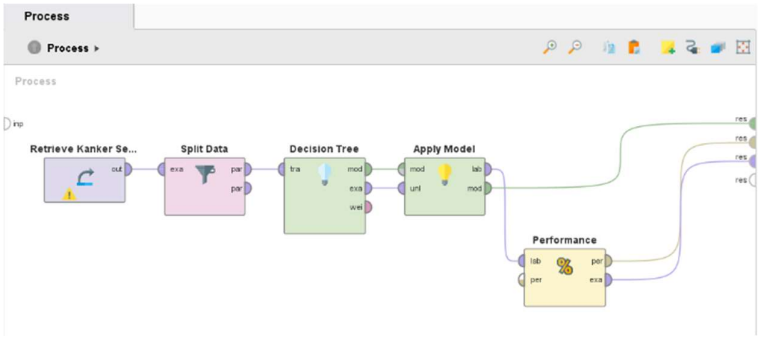
Gambar 2. Akar pertama pohon Keputusan

Pada gambar 2 menjelaskan bahwa, akar pertama dari perhitungan entropi dan *gain* total data adalah atribut umur, kemudian setelah menghitung entropi dan *gain* pada total data maka proses menghitung entropi dan *gain* diulangi sampai nilai gain dan entropi sama. Pada perhitungan C4.5 dari total data dan semua atribut didapatkan 74 *rule* yang dapat memperkirakan positif atau negatf pada pasien yang berdasarkan faktor kebiasaan dengan menggunakan tes *schiller.*

3.2. Performa C4.5 dengan Rapidminer

Implementasi ini menggunakan tools rapidminer untuk mengetahui berapa besar keakuratan pada penelitian ini. Hasil Penelitian ini merupakan dari implementasi algoritma C4.5 untuk

membuat rule dan model *decision tree* dengan menggunakan *rapidminer* 10.1.3. Dengan adanya pengujian penelitian ini akan diketahui hasil yang diperoleh dengan hasil secara komputerisasi. Penelitian ini akan memberikan hasil penggunaan rapidminer untuk implementasi C4.5 dan hasil pada penelitian ini yang menunjukkan alur proses implementasi rapidminer dari gambar 3.



Gambar 3. Proses *rapidminer*

Pada gambar 3 di atas merupkan alur proses implementasi *rapidminer* dengan metode C4.5 dari bagian proses *retrive* data sampai bagian proses *performance*. Pada *process retrive* data nilai *output* disambungkan ke *example input* proses *split data*, dalam parameter *split* data disiikan *partitions* dengan ratio 0.8 dan 0.2 yang artinya 80% *training* data dan 20% data *testing*. Kemudian tarik *output split* data ke proses *decision tree* dalam parameter *decision tree* pilih *criterion gain ratio* dan *maximal depth* diisikan 13. Lalu tarik output mod dan exa ke proses *apply* model pada *output apply* model mod ditarik ke res dan output lab tarik ke input lab proses *performance* dan pada output *performance* per dan exa tarik ke res untuk menghasilkan nilai performa pada C4.5 dan menghasilkan pohon Keputusan. Berikut ini adalah tabel 18 merupakan matriks kebingungan di rapidminer yang diperoleh.

Tabel 18. Matriks kebingungan *rapidminer*

	true Positif	***true Negatif***	***class precision***
pred. Positif	209	23	90.09%
pred. Negatif	28	604	95.57%
class recall	88.19%	96.33%	

Pada tabel 18 di atas adalah tabel matriks kebingungan yang di mana menghasilkan performa seperti akurasi, presisi dan *recall*, Berikut adalah perhitungan akurasi, presisi, dan *recall* pada implementasi di *rapidminer*. Perhitungan akurasi, presisi dan *recall* menggunakan persamaan (3), (4) dan (5) dengan sebagai berikut ini:

$$Accuracy = \frac{(True\ Positive + Tr\quad Negative)}{(Total\ Data)} x100\%$$

$$Accuracy = \frac{(209+604)}{864} x100\% = 94{,}10\%$$

$$Precision = \frac{True\ Positive\ (TP)}{False\ Positive\ (FP) + True\ Positive\ (TP)} x100\%$$

$$Precision = \frac{209}{209+2}\ x\ 100\% = 95.57\%$$

$$Recall = \frac{True\ Positive\ (TP)}{False\ Negative\ (FN) + True\ Positive\ (TP)} x100\%$$

$$Recall = \frac{209}{209+2}\ x\ 100\% = 96.33\%$$

Maka didapatkan hasil performa pada penelitian ini dengan implementasi menggunakan *rapidminer* dengan hasil tabel 19.

Tabel 19. Hasil performa implementasi *rapidminer*

Accuracy	**Precision**	**Recall**	**Curve ROC**
94%	95%	96%	0.987

Hasil dari tabel 19 Dalam evaluasi performa, mendapatkan nilai akurasi, presisi, *recall,* dan AUC (*Area Under Curve*). Akurasi yang diperoleh adalah 94.10%, presisi 95.57%, *recall* 96.33% dan AUC 0.987 dengan menggunakan data penelitian [12] yang menggunakan 12 atribut dan yang di*resample* 1080 data dapat dikatakan bahwa penelitian ini sangat baik, karena pada penelitian yang sebelumnya tujuannya adalah untuk menemukan metode *transfer learning* yang lebih efektif dibandingkan dengan metode lain yang ada. Penelitian ini melakukan perbandingan hasil dari nilai gain yang diperoleh saat mengukur signed Area Under the gain Curve (sAUC). sehingga hasil sAUC pada penelitian tersebut yang sangat tidak sebesar pada penelitian ini.

4. KESIMPULAN

Dari hasil penelitian ini, data yang digunakan pada penelitian [12] dan diolah dapat menghasilkan pohon Keputusan yang menghasilkan sebuah rule. Terdiri dari 74 *rule* yang dihasilkan dari implementasi algoritma C4.5. rule ini bisa digunakan dalam sebagai prediksi kanker serviks dapat memperkirakan jumlah positif atau negatifnya kanker serviks menggunakan tes *schiller* dalam berdasarkan kebiasaan dari pasien. Penelitian klasifikasi kanker serviks berdasarkan kebiasaan dengan mengimplementasikan C4.5 ini mendapatkan akurasi sebesar 94.10%, *recall* 96.33% dan *precission* 95.57% dengan kurva

https://doi.org/10.25077/TEKNOSI.v10i1.2024.36-44

AUC sebesar 0.987 dan penelitian ini menunjukkan kedalam klasifikasi sangat baik.

Pada penelitian ini memiliki kekurangan dan kelemahan. yang harus dipertimbangkan untuk pengembangan penelitian selanjutnya:

1. Memerlukan pemilihan atribut yang kompleks dan untuk jenis target atribut memakai semuanya untuk membandingkan hasil akurasi di mana tes kanker serviks yang paling cocok untuk mengklasifikasi kanker serviks pada data tersebut.
2. Penelitian selanjutnya dianjurkan untuk menggunakan algoritma yang beragam seperti *random forest*, *random tree*, *gradient boosted tree* dan ID3 untuk membandingkan mana yang paling bagus dalam dari sisi keakuratan dan kurva AUCnya.

DAFTAR PUSTAKA

[1] S. S. Arifin, A. M. Siregar, and T. Al Mudzakir, "Klasifikasi Penyakit Kanker Serviks Menggunakan Algoritma Support Vector Machine (SVM)," in *Conference on Innovation and Application of Science and Technology (CIASTECH)*, 2021, pp. 521–528.

[2] H. Akbar and S. Sandfreni, "Klasifikasi Kanker Serviks Menggunakan Model Convolutional Neural Network Alexnet," *JIKO (Jurnal Informatika dan Komputer)*, vol. 4, no. 1, pp. 44–51, 2021.

[3] E. Rizkyani, I. Ernawati, and N. Chamidah, "Klasifikasi Multi-Label Menggunakan Metode Multi-Label K-Nearest Neighbor (Ml-KNN) Pada Penyakit Kanker Serviks," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 4, pp. 1281–1293, 2022.

[4] A. Dharma, P. Manalu, G. S. Sinaga, R. Siringoringo, I. S. Palangai, and K. Setiawan, "Deteksi Pola Pasien Kanker Serviks dengan Algoritma Extra Trees dan K-Nearest Neighbor," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 3, no. 1, pp. 32–36, 2020.

[5] I. M. A. O. Gunawan, I. D. A. I. Saraswati, I. D. G. R. Agung, and I. P. E. Putra, "Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4. 5 Dengan Rapidminer," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 2, pp. 73–83, 2023.

[6] D. Jollyta, W. Ramdhan, and M. Zarlis, *Konsep Data Mining Dan Penerapan*. Deepublish, 2020.

[7] T. Arifin, "Metode Data Mining Untuk Klasifikasi Data Sel Nukleus Dan Sel Radang Berdasarkan Analisa Tekstur," *Jurnal Informatika*, vol. 2, no. 2, 2015.

[8] T. G. Pratama, A. Ridwan, and A. Prihandono, "Penerapan Algoritma C4. 5 untuk Klasifikasi Kanker Serviks Tingkat Awal," *Flurecol Journal. Part E: Engineering*, vol. 1, no. 1, pp. 1–6, 2021.

[9] A. Amrin, I. Satriadi, and O. Rosanto, "Algoritma C4. 5 Untuk Diagnosa Penyakit Tuberkulosis," *Jurnal Khatulistiwa Informatika*, vol. 7, no. 2, 2019.

[10] C. Paramita, F. A. Rafrastara, and L. I. Kencana, "Pengembangan Sistem Klasifikasi Karakteristik Siswa Berbasis Website dengan menggunakan Algoritma C4. 5," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 1, pp. 17–21, 2023.

[11] M. F. Unlersen, K. Sabanci, and M. Özcan, "Determining cervical cancer possibility by using machine learning methods," *Int. J. Latest Res. Eng. Technol*, vol. 3, no. 12, pp. 65–71, 2017.

[12] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings 8*, Springer, 2017, pp. 243–250.

[13] A. Sepharni, I. E. Hendrawan, and C. Rozikin, "Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4. 5," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 7, no. 2, pp. 117–126, 2022.

[14] W. Yunus, "Implementasi Algoritma C. 45 Dalam Prediksi Penyakit Kanker," *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, vol. 4, no. 1, pp. 70–76, 2023.

[15] W. Yusnaeni and W. Widiarina, "Penerapan Algoritma C4. 5 Dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes)," *Jurnal Khatulistiwa Informatika*, vol. 8, no. 1, pp. 56–60, 2022.

[16] M. Yunus, H. Ramadhan, D. R. Aji, and A. Yulianto, "Penerapan Metode Data Mining C4. 5 Untuk Pemilihan Penerima Kartu Indonesia Pintar (KIP)," *Paradig.-J. Komput. dan Inform*, vol. 23, no. 2, 2021.

[17] F. K. Fikriah, "Instance Selection dengan Naïve Bayes pada Klasifikasi Kanker Serviks," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 2, pp. 83–91, 2021.

[18] M. F. Arifin and D. Fitrianah, "Penerapan Algoritma Klasifikasi C4. 5 Dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus: PT Atria Artha Persada," *InComTech: Jurnal Telekomunikasi dan Komputer*, vol. 8, no. 2, pp. 87–102, 2018.

[19] K. F. Irnanda, D. Hartama, and A. P. Windarto, "Analisa Klasifikasi C4. 5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi," *Jurnal Media Informatika Budidarma*, vol. 5, no. 1, pp. 327–331, 2021.

NOMENKLATUR

Persamaan (1)

S	dataset
c	jumlah kelas
P_i	proporsi data dalam kelas i

Persamaan (2)

k	jumlah subset yang dihasilkan

BIODATA PENULIS

Kemal Taufiq Hidayah
Kemal Taufiq Hidayah lahir di Jakarta 10 Agustus 2000, merupakan seseorang mahasiswa program studi S1 Teknik Informatika Universitas Trilogi, Jakarta.

Budi Arifitama.,S.T., M.M.S.I
Merupakan peneliti dan dosen di Program Studi Teknik Informatika Universitas Trilogi degan konsentrasi bidang keilmuan Teknologi Multimedia

Silvester Dian Handy Permana, S.T.,M.T.I.
Merupakan peneliti dan dosen di Program Studi Teknik Informatika Universitas Trilogi degan konsentrasi bidang keilmuan Teknologi Sistem Cerdas.

LAMPIRAN

Tabel i normalisasi data

NO	Umur	Jumlah Pasangan Seksual	Umur Pertama kali Hubungan Seks	Jumlah Kehamilan	Merokok	Merokok pertahun	Jumlah bungkus rokok per tahun	Alat kontrasepsi hormno	Pemakaian Alat kontrasepsi hormno	Alat kontrasepsi rahim	Pemakaian alat kontrasepsi rahim	Penyakit menular seksual	Schiller
1	Remaja Akhir	Poliamori	Remaja Awal	Normal	TIdak	Tidak Merokok	Tidak Merokok	Tidak memakai	Tidak Pernah	Tidak Memakai	Tidak Pernah	Tidak	Positif
2	Remaja Akhir	Monogami	Remaja Awal	Normal	TIdak	Tidak Merokok	Tidak Merokok	Tidak memakai	Tidak Pernah	Tidak Memakai	Tidak Pernah	Tidak	Positif
3	Dewasa Awal	Monogami	Remaja Akhir	Normal	TIdak	Tidak Merokok	Tidak Merokok	Tidak memakai	Tidak Pernah	Tidak Memakai	Tidak Pernah	Tidak	Positif
4	Lansia Awal	Poliamori	Remaja Awal	Normal	Iya	Berat	Banyak	Memakai	Menengah	Tidak Memakai	Tidak Pernah	Tidak	Positif
…	…	…	….	…	….	…..	…..	……	……	…..	….	….	….
1080	Dewasa Akhir	Poliamori	Remaja Akhir	Normal	TIdak	Tidak Merokok	Tidak Merokok	Tidak memakai	Tidak Pernah	Tidak Memakai	Tidak Pernah	Tidak	Positif

https://doi.org/10.25077/TEKNOSI.v10i1.2024.36-44