

**PENINGKATAN KINERJA ALGORITMA C4.5 DENGAN METODE
AVERAGE GAIN**

TESIS

TITIN QOWIDHO

177038031



**PROGRAM STUDI S2 TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA
MEDAN
2020**

**PENINGKATAN KINERJA ALGORITMA C4.5 DENGAN METODE
AVERAGE GAIN**

TESIS

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah
Magister Teknik Informatika

TITIN QOWIDHO

177038031



**PROGRAM STUDI S2 TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA
MEDAN
2020**

PERSETUJUAN

Judul : PENINGKATAN ALGORITMA C4.5 DENGAN
METODE AVERAGE GAIN
Kategori : TESIS
Nama : TITIN QOWIDHO
Nomor Induk Mahasiswa : 177038031
Program Studi : MAGISTER (S2) TEKNIK INFORMATIKA
Fakultas : ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA

Komisi Pembimbing :

Pembimbing I



Prof. Dr. Muhammad Zarlis

Pembimbing II

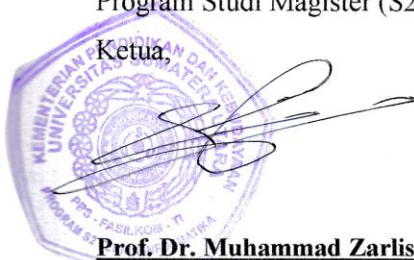


Dr. Erna Budhiarti Nababan, M.IT

Diketahui/disetujui oleh

Program Studi Magister (S2) Teknik Informatika

Ketua,



Prof. Dr. Muhammad Zarlis
195707011986011003

PERNYATAAN

PENINGKATAN ALGORITMA C4.5 DENGAN METODE AVERAGE GAIN

TESIS

Saya mengakui bahwa tesis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 9 November 2020

TITIN QOWIDHO

NIM. 177038031

**PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH
UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademi Universitas Sumatera Utara, saya yang bertandatangan di
bahwa ini :

Nama	: Titin Qowidho
NIM	: 177038031
Program Studi	: Magister (S2) Teknik Informatika
Jenis Karya Ilmia	: Tesis

Dengan pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada
Universitas Sumatera Utara Hak Bebas Royalti Non-Eksklusif (*Non-Exclusive Royalti
Free Right*) atas tesis saya yang berjudul:

**PENINGKATAN ALGORITMA C4.5 DENGAN METODE
AVERAGE GAIN**

Beserta perangkat yang ada (jika di perlukan). Dengan Hak Bebas Royalti Non-
Eksklusif ini mengelola dalam bentuk database, merawat dan mempublikasikan tesis
saya tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai
penulis dan sebagai pemenang dan/ atau sebagai pemilik hak cipta

Demikian pernyataan ini dibuat dengan sebenarnya

Medan, 9 November 2020

TITIN QOWIDHO
NIM.177038031

Telah di uji pada
Tanggal: 06 Agustus 2020

PANITIA PENGUJI TESIS

Ketua : Prof. Dr. Muhammad Zarlis

Angota :1. _Dr. Erna Budhiarti Nababan, M.IT
2. Dr. Syahril Efendi, S.Si., M.IT
3. Prof. Dr. Tulus

RIWAYAT HIDUP

DATA PRIBADI

Nama lengkap : Titin Qowidho
Tempat dan Tanggal Lahir : Tasik Raja, 16 Oktober
Alamat Rumah : Dsn.V PKS Blankahan Kec. Kuala Kab. Langkat
Telepon/HP : 0822-7215-4851
E-mail : titinqowidho123@gmail com

DATA PENDIDIKAN

SD	SDN 118319 TASI K RAJA	TAMAT : 2003
SMP	SMP YPHB KUALA	TAMAT : 2006
SMK	SMK YPHB KUALA	TAMAT : 2009
S1	Sitem Informatika SMTIK KAPUTAMA BINJAI	TAMAT : 2013
S2	Teknik Informatika USU Medan	TAMAT : 2020

UCAPAN TERIMA KASIH

Alhamdulillah, segala pugi dan syukur saya ucapkan kehadiran Tuhan yang maha esa atas berkat rahmatnya, saya dapat menyelesaikan Tesis ini dalam kurun waktu yang telah di tetapkan.

Ucapan terima kasih juga saya sampaikan kepada pihak-pihak yang telah membantu saya selama penulisan Tesis ini, sehingga Tesis ini dapat diselesaikan dengan baik. Pada kesempatan kali ini saya ingin mengucapkan terima kasih yang sebesar-besarnya kepada :

1. Bapak Prof. Dr. Runtung Sitepu, SH, M.Hum selaku Rektor Universitas Sumatera Utara.
2. Bapak Prof. Dr. Opim Salim Sitompul, M.Sc selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi.
3. Bapak Prof. Dr. Muhammad Zarlis selaku Ketua Program Studi Magister Teknik Informatika, sekaligus selaku Pembimbing Pertama yang telah membimbing saya sehingga tesis ini dapat diselesaikan dengan baik.
4. Bapak Dr. Syahril Efendi, S.Si, M.IT selaku Sekretaris Program Studi Magister Teknik Informatika, sekaligus penguji Pertama yang telah memberikan saran dan masukan serta arahan yang baik dalam penyelesaian tesis ini. beserta seluruh Staf Pengajar Program Studi Magister Teknik Informatika Program Pasca sarjana Fakultas Ilmu Komputer Universitas Sumatera Utara.
5. Ibu Dr. Erna Budhiarti Nababan selaku Pembimbing Kedua yang telah membimbing saya sehingga tesis ini dapat diselesaikan dengan baik.
6. Bapak Prof. Dr. Tulus , Vor. Dipl. Math., M. Si., Ph. D selaku Penguji Kedua yang telah memberikan saran dan masukan serta arahan yang baik dalam penyelesaian tesis ini.

7. Orang tua laki-laki saya Abidin Safrizal Situmeang dan Orang tua Perempuan saya Nurhayati Harahap, yang telah mendukung saya dan mendoakan saya sehingga tesis ini terselesaikan dengan baik.
8. Semua pihak yang terlibat langsung ataupun tidak langsung yang tidak dapat saya sebutkan satu persatu yang telah membantu saya dalam menyelesaikan Tesis ini.

Saya menyadari masih banyak kekurangan dalam penulisan Tesis ini, oleh karena itu saya mengharapkan kritik dan saran dari para pembaca sebagai masukan bagi penelitian ini. Agar penelitian ini dapat bermanfaat lebih baik lagi bagi saya ataupun bagi peneliti selanjutnya.

Medan, 09 November 2020

Penulis

Titin Qowidho

ABSTRAK

Zhang (2012) mengusulkan dua metode pruning. Metode pertama disebut *heterogeneous-cost sensitive learning* (HCSL) dengan memodifikasi *average gain split* atribut (Mitchell, 1997) yang dikalikan dengan selisih misklasifikasi (*misclassification cost* dari atribut sebelum di-split dan setelah di-split). Metode pruning kedua adalah menggunakan nilai ambang (*threshold pruning*). Algoritma C4.5 masih mempunyai kelemahan dalam melakukan prediksi atau klasifikasi data apabila kelas-kelas yang digunakan dalam jumlah yang banyak dapat menyebabkan meningkatnya waktu pengambilan keputusan. Maka dibutuhkan satu pendekatan untuk meningkatkan kinerja terhadap algoritma C4.5 dengan split atribut yang dipilih yang menggunakan penerapan nilai *average gain* guna membantu memprediksi screening test yang akan dilalui oleh pengidap penyakit kanker khususnya kanker serviks sehingga memperoleh pengobatan yang tepat dan cepat. pengujian yang telah dilakukan yang menggunakan dataser *Cervical Cancer* pada metode C4.5 yang memiliki tingkat akurasi sebesar 90.37% , dengan tingkat kesalahan pengklasifikasian dengan nilai 9.63%. Sedangkan klasifikasi model C4.5 Average Gain memiliki akurasi sebesar 93.90%, dengan tingkat kesalahan pengklasifikasian sebesar 6.10%. Pada dataser *Kanker Rahim* pada metode C4.5 yang memiliki tingkat akurasi sebesar 95.61% , dengan tingkat kesalahan pengklasifikasian dengan nilai 4.38%. Sedangkan klasifikasi model C4.5 Average Gain memiliki akurasi sebesar 98.61%, dengan tingkat kesalahan pengklasifikasian sebesar 1.4%. Perbedaan pada penelitian ini disebabkan oleh jumlah dari atribut yang berbeda, semakin banyak atribut yang diuji maka menghasilkan tingkat akurasi yang lebih rendah dari atribut yang sedikit, maka dataset Kanker Rahim memiliki akurasi yang lebih tinggi dibandingkan dengan dataset *Cervical Cancer*.

Kata Kunci : Decision Tree, C4.5, Average Gain, Akurasi.

ABSTRACT

Zhang (2012) proposed two pruning methods. The first method is called heterogeneous-cost sensitive learning (HCSL) by modifying the average gain split attribute (Mitchell, 1997) which is multiplied by the difference in misclassification (misclassification cost of attributes before being split and after being split). The second pruning method is to use a threshold value. (thresholdpruning). The C4.5 algorithm still has weaknesses in predicting or classifying data if a large number of classes are used which can lead to increased decision-making time. So an approach is needed to improve the performance of the C4.5 algorithm with the selected split attribute that uses the application of the average gain value to help predict the screening test that people with cancer, especially cervical cancer, will pass so that they get the right and fast treatment. Tests that have been carried out using the Cervical Cancer dataset on the C4.5 method have an accuracy rate of 90.37%, with a classification error rate of 9.63%. While the classification model C4.5 Average Gain has an accuracy of 93.90%, with a classification error rate of 6.10%. In the uterine cancer dataset using C4.5 method, it has an accuracy rate of 95.61%, with a classification error rate of 4.38%. While the classification model C4.5 Average Gain has an accuracy of 98.61%, with a classification error rate of 1.4%. The difference in this study is caused by the number of different attributes, the more attributes tested, the lower the accuracy rate of the few attributes, the cervical cancer dataset has a higher accuracy than the Cervical Cancer dataset.

Keywords: Decision Tree, C4.5, Average Gain, Accuracy.

DAFTAR ISI

PERSETUJUAN	i
PERNYATAAN ORISINALITAS	ii
PERNYATAAN PERSETUJUAN PUBLIKASI	iii
PANITIA PENGUJI TESIS	iv
RIWAYAT HIDUP	v
UCAPAN TERIMA KASIH	vi
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiii
 BAB I. PENDAHULUAN	 1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
 BAB II. LANDASAN TEORI	 5
2.1 Model Pohon Keputusan (Decision Tree).....	5
2.2 Model Decision Tree C4.5	7
2.3 Average Gain	9
2.4 Teknik Klasifikasi	10
2.5 Penelitian Terdahulu yang Relevan	11
 BAB III. METODE PENELITIAN	 14
3.1 Data Yang Digunakan	14
3.2 Peningkatan Kinerja Algoritma C4.5 dan Metode Average Gain	17

3.3 Data Preprocessing	17
3.3.1 Missing Value	17
3.3.2 Proses Pembentukan Model Klasifikasi Decision Tree C4.5	20
3.3.3 Pengujian Akurasi Menggunakan Confusion Matrix....	27
3.4 Software dan Tools yang Digunakan	28
BAB IV. HASIL DAN KESIMPULAN.....	29
4.1 Hasil Pengujian	29
4.1.1 Persiapan Data Awal (Data Preprocessing)	29
4.1.2 Pemberian Kategor i	31
4.1.3 Hasil Nilai Entropy	32
4.1.4 Hasil Nilai Information Gain	36
4.1.5 Hasil Nilai Split Info	39
4.2 Pengujian Akurasi Menggunakan Confusion Matrix	41
4.3 Kesimpulan Pengujian	47
BAB V. KESIMPULAN DAN SARAN	49
5.1 Kesimpulan	49
5.2 Saran	49
DAFTAR PUSTAKA	
LAMPIRAN	

DAFTAR GAMBAR

Gambar II.1	: Skematis Decision Tree	6
Gambar II.2	: Pemetaan Atribut Dari Dataset dan Model Klasifikasi Decision Tree	7
Gambar II.3	: Proses Klasifikasi (Sumber: Han, et al. 2012)	11
Gambar III.1	: Tahapan Penerapan	16
Gambar IV.1	: Proses Remove Data Duplicate (Rapidminer)	29
Gambar IV.2	: Proses Normalize Data (Rapidminer)	29
Gambar IV.3	: Grafik Perbandingan Akurasi dan Error	48

DAFTAR TABEL

Tabel II.1	: Penelitian Terdahulu	12
Tabel III.1	: Dataset UCI Cervical Cancer	14
Tabel III.2	: Tabel Sebelum Mising Volue	18
Tabel III.3	: Data Sesudah Mising Volue	19
Tabel III.4	: Data Uji Dataset Cervical Cancer	20
Tabel III.5	: Pemberian Kategori Dataset Cervical Cancer	20
Tabel III.6	: Data Uji Dataset Kanker Rahim	21
Tabel III.7	: Confusion Matrix	27
Tabel IV.1	: Hasil Data Preprocessing (Cervical Cancer)	30
Tabel IV.2	: Hasil Data Preprocessing (Kanker Rahim)	30
Tabel IV.3	: Data Uji Dataset (Cervical Cancer)	31
Tabel IV.4	: Hasil Pemberian Kategori (Cervical Cancer)	31
Tabel IV.5	: Hasil Pemberian Kategori (Kanker Rahim)	32
Tabel IV.6	: Nilai Entropy Per Atribut (Cervical Cancer)	32
Tabel IV.7	: Nilai Entropy Per Atribut (Kanker Rahim)	35
Tabel IV.8	: Hasil Perolehan Information Gain (Cervical Cancer)	36
Tabel IV.9	: Hasil Perolehan Information Gain (Kanker Rahim)	38
Tabel IV.10	: Hasil Nilai Split Info (Cervical Cancer)	39
Tabel IV.11	: Hasil Nilai Split Info (Kanker Rahim)	40
Tabel IV.12	: Rules Model Klasifikasi Decision Tree C4.5 (Average Gain)	40
Tabel IV.13	: Hasil Pengujian Model Klasifikasi C4.5 (Cervical Cancer)	41
Tabel IV.14	: Hasil Pengujian Model Klasifikasi C4.5 (Kanker Rahim)	43
Tabel IV.15	: Confusion Matrix C4.5 (Cervical Cancer)	45
Tabel IV.16	: Confusion Matrix C4.5 Average Gain (Cervical Cancer)	45
Tabel IV.17	: Confusion Matrix C4.5 (Kanker Rahim)	46
Tabel IV.18	: Confusion Matrix C4.5 Average Gain (Kanker Rahim)	46
Tabel IV.19	: Performance Matrix Klasifikasi Decision Tree C4.5	48

BAB I

PENDAHULUAN

1.1. Latar Belakang

Klasifikasi objek data yang didasari oleh objek yang sudah ditentukan dalam sebuah data. ada banyak algoritma klasifikasi tetapi *Decision tree* yang paling sering digunakan. (Seema, 2012). Pengklasifikasian dari *Decision Tree* yaitu salah satu jenis pengelompokan yaitu diagram alur dengan struktur pohon, dengan setiap node didalam memperlihatkan tes dari setiap atribut, masing-masing cabang mewakili hasil dari tes, dan setiap simpul daun mewakili kelas. Model untuk mengklasifikasi sebuah catatan untuk temukan jalur akar daun untuk mengukur uji atribut dan atribut daun tersebut merupakan hasil klasifikasi yang digunakan Decision Tree. (Qin-yun, 2016).

Pohon keputusan adalah struktur pohon seperti bagan alur, di mana masing-masing simpul internal mewakili tes pada atribut, setiap cabang mewakili hasil pengujian, label kelas diwakili oleh setiap simpul daun (atau simpul terminal). Diberikan tuple X, nilai atribut tuple diuji terhadap pohon keputusan. Path dilacak dari root ke node leaf yang menampung prediksi kelas untuk tuple. Sangat mudah untuk mengubah pohon keputusan menjadi aturan klasifikasi. Pembelajaran pohon keputusan menggunakan pohon keputusan sebagai model prediksi yang memetakan pengamatan tentang kesimpulan tentang nilai target. Ini adalah salah satu pendekatan pemodelan prediktif yang digunakan dalam statistik, penambangan data, dan pembelajaran mesin. Model pohon tempat variabel target dapat mengambil dataset hingga nilai-nilai disebut pohon klasifikasi, dalam struktur pohon ini, daun mewakili label kelas dan cabang mewakili konjungsi fitur yang mengarah ke label kelas tersebut. (Sharma, 2016).

Pada penelitian (Mesarić, 2016) yang bertujuan untuk membuat model yang berhasil mengklasifikasikan siswa menjadi salah satu dari dua kategori, tergantung

pada keberhasilan mereka di akhir akademik tahun pertama mereka, dan menemukan variabel bermakna yang mempengaruhi kesuksesan mereka. Model ini didasarkan pada informasi mengenai keberhasilan siswa di sekolah menengah dan kursus mereka setelah menyelesaikan tahun pertama studi mereka, serta peringkat preferensi yang ditugaskan untuk fakultas yang diamati, dan upaya untuk mengklasifikasikan siswa ke dalam salah satu dari dua kategori sesuai dengan keberhasilan akademik mereka. Membuat model diperlukan dalam pengumpulan data pada semua mahasiswa sarjana yang terdaftar untuk memasuki tahun kedua mereka di Fakultas Ekonomi, Universitas Osijek, serta data pada penyelesaian ujian negara.

Algoritma C4.5 merupakan salah satu dari metode *Decision Tree* dalam melakukan proses pengklasifikasian menggunakan konsep *information entropy*. Algoritma C4.5 menggunakan criteria split dari ID3, *Gain Ratio* merupakan modifikasi dari metode tersebut. Algoritma ID3 menggunakan *Information Gain (IG)* untuk criteria split atribut, sedangkan Algoritma C4.5 dengan *Gain Ratio (GR)*, dimana nilai akar(*root*) berasal dari gain yang tinggi. Langkah dari proses algoritma C4.5 yaitu dengan melakukan perhitungan dari nilai Entropy. Dengan masing-masing atribut dilakukan perhitungan nilai Gain Ratio, kemudian atribut yang memiliki nilai Gain Ratio yang tinggi akan dipilih menjadi akar dan yang rendah akan menjadi cabang, kemudian menghitung kembali nilai Gain Ratio dari tiap atribut dengan tidak menggunakan atribut yang terpilih menjadi akar dari proses sebelumnya, selanjutnya proses dilakukan sampai menghasilkan nilai Gain 0 pada atribut yang tersisa.

Dalam pendekatan decision tree, pruning merupakan proses untuk menghilangkan beberapa cabang (*node*) yang tidak diperlukan. Node yang tidak dibutuhkan akan menyebabkan noisy data yang kurang relevan (Zhang, 2012). Menyebabkan Decision Tree yang besar yang dinamakan over fitting (Wang, et al. 2012).

Zhang (2012) mengusulkan dua metode pruning. Metode pruning kedua adalah menggunakan nilai ambang (*threshold pruning*). Kedua metode pruning tersebut diuji menggunakan model pengklasifikasian Decision Tree dengan kriteria split dari algoritma ID3 dan C4.5 (*heterogenous cost*) terhadap enam dataset dan menyimpulkan bahwa kedua metode pruning yang diusulkan tersebut dapat digunakan untuk model pengklasifikasian decision tree.

Sivapriya & Nadira (2013) mengusulkan reduksi fitur *Principal Component Analysis* (PCA) dan seleksi fitur *gain-ratio* algoritma decision tree C4.5 untuk meningkatkan klasifikasi SVM. Dengan menggunakannya sebuah data citra MRI (*Magnetic Resonance Images*) dari jaringan saraf manusia untuk diklasifikasikan melalui pengujian *neuro psychological*. Metode ini memiliki performa klasifikasi SVM dengan tingkat akurasi sebesar 97%.

Kavitha & Kannan (2016) mengusulkan reduksi fitur *Principal Component Analysis* dan seleksi fitur menggunakan algoritma *Wrapper filter* serta kriteria split dari algoritma C4.5 untuk digunakan sebagai metode untuk menghasilkan fitur-fitur yang relevan dari dataset UCI *Heart dataset*. Metode yang diajukan dapat menghasilkan fitur-fitur yang relevan sehingga meningkatkan efisiensi dan akurasi.

1.2. Rumusan Masalah

Algoritma C4.5 masih mempunyai kelemahan dalam melakukan prediksi atau klasifikasi data apabila kelas-kelas yang digunakan dalam jumlah yang banyak dapat menyebabkan meningkatnya waktu pengambilan keputusan. Maka dibutuhkan satu pendekatan untuk meningkatkan kinerja terhadap algoritma C4.5 dengan split atribut yang dipilih yang menggunakan penerapan nilai average gain guna membantu memprediksi screening test yang akan dilalui oleh pengidap penyakit kanker khususnya kanker serviks sehingga memperoleh pengobatan yang tepat dan cepat.

1.3. Batasan Masalah

Penelitian ini mempertimbangkan beberapa batas masalah yaitu:

1. Pada penelitian ini, mengetahui kinerja penggunaan *decision tree* C4.5 maka digunakan data observasi berupa dataset UCI *Cervical Cancer* yang berasal dari (<http://archive.ics.uci.edu/ml>).
2. Penelitian ini berfokus pada target *Biopsy* seperti yang direkomendasikan oleh tinjauan literatur.
3. Penelitian ini hanya membahas tingkat kinerja antara decision tree C4.5 dengan induksi *gain ratio* dan decision tree C4.5 konvensional.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk melakukan peningkatan kinerja metode klasifikasi decision tree C4.5 dengan menggunakan metode average gain sebagai split atribut.

1.5. Manfaat Penelitian

Melalui penelitian ini maka akan memperoleh hasil analisis mengenai pengelompokan data, klasifikasi data, prediksi pada algoritma Decision Tree C4.5 menggunakan metode average gain sebagai split atribut, serta mengetahui pencapaian kinerja yang diperoleh dengan penerapan algoritma Decision Tree C4.5 menggunakan metode average gain sebagai split atribut. Penulis juga mengharapkan manfaat penelitian ini untuk membantu para medis dalam meningkatkan kemampuan screening test khususnya dalam kanker serviks. Juga menambah wawasan dan dalam menganalisa suatu algoritma khususnya Decision Tree C4.5.

BAB II

LANDASAN TEORI

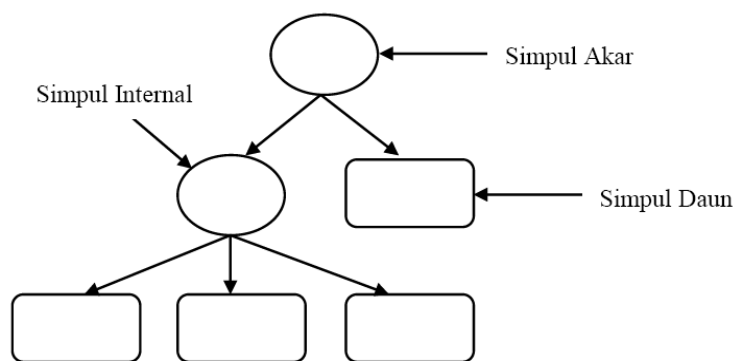
2.1 Model Pohon Keputusan (*Decision Tree*)

Menurut (Alpaydin, 2016) Decision Tree metode struktur data yang memiliki sifat hirarkikal yang telah diterapkan pada strategi *divide and conquer* dan metode *efficient nonparametric* dan telah digunakan dalam penyelesaian metode klasifikasi dan metode regresi. *Decision tree* juga termasuk algoritma machine learning dengan teknik klasifikasi yang bersifat *supervised learning* dengan cara pembentukan dari pohon keputusan yang diproses dari sebuah data. *Decision tree* sangat menyerupai sebuah pohon terbalik, dengan posisi paling atas yaitu akar, dan posisi bawah yaitu daun (Abellan & Castellano 2016).

Pada pohon keputusan ini terdapat *field* dan *record* dalam bentuk tabel. Pembentukan tree disebut sebagai atribut dalam suatu parameter sesuai dengan criteria yang terdapat pada data. Manfaat dari pohon keputusan ini merupakan sebuah teknik dalam mensesederhanakan sebuah proses dalam pengambilan keputusan yang sulit menjadi lebih sederhana agar dapat menemukan solusi dalam permasalahan yang ada.

Decision tree terdiri dari beberapa aturan yang dibagi menjadi beberapa populasi heterogen yang lebih sedikit, yang lebih homogen yang terlihat dari tujuan variabelnya, yang bertujuan mengelompokkan model pohon keputusan yang terarah ke hitungan probabilitas dari masing-masing record terhadap kategori dalam mengklasifikasikan pengelompokkan pada kelas yang serupa. Pada pohon keputusan yang menjadi salah satu metode klasifikasi yang representasi struktur pohon yang setiap simpul internalnya adalah atribut, kemudian simpul terminal yaitu label class, serta simpul yang atas adalah akar.

Decision tree memiliki beberapa elemen yang terdapat dari satu atau lebih variabel atribut dan satu atribut kelas. Pohon klasifikasi yang digunakan akan memprediksi nilai pada atribut kelas dengan dipertimbangkannya nilai yang terdapat pada tiap-tiap variabel atribut yang di uji (Abellan, 2016)



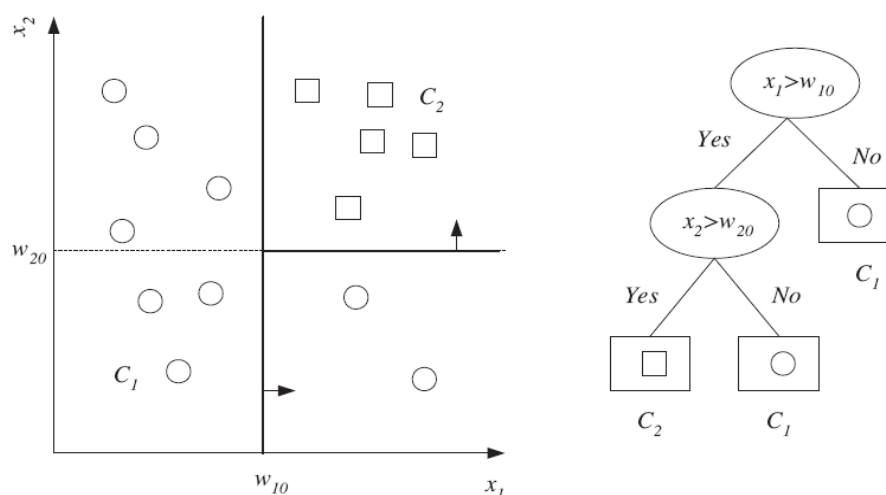
Gambar 2.1. Skematis *Decision Tree*

Dibawah ini merupakan penjelasan tentang 3 macam simpul yang berada pada pohon keputusan dari gambar diatas:

- Simpul paling atas merupakan simpul akar yang tak memiliki masukan dan hanya memiliki keluaran yang lebih dari 1
- Simpul cabang dari simpul akar merupakan simpul internal yang hanya mempunyai 1 masukan dan memiliki min. 2 masukan.
- Simpul yang akhir merupakan simpul daun yang hanya mempunyai satu masukan dan tidak memiliki keluaran tetapi dinamakan dengan simpul terminal.

Fakta yang diubah dalam metode keputusan yang besar dengan menjadikan pohon keputusan dengan menggunakan aturan. Kemudian aturan ini dimengerti dan juga dapat dipahami dengan bahasa yang terstruktur dalam pencarian record pada kategori tertentu.

Dalam mengeksplorasi dengan menggunakan data decision tree dalam prosesnya dengan penemuan hubungan bersembunyi dari beberapa calon variable input dengan target. Karena dalam decision tree terdapat eksplorasi data dan pemodelan, decision tree juga sangat bagus dalam proses awal dalam pembuatan model walaupun sebagai model akhir pada teknik lain.



Gambar 2.2. Pemetaan Atribut Dari Dataset dan Model Klasifikasi Decision Tree.

(Sumber: Alpaydin, 2010)

Pada gambar diatas yang menjelaskan tentang peta atribut dalam sebuah dataset untuk membuat sebuah pemodelan pada klasifikasi decision tree. Node yang berbentuk oval adalah simpul akar (*root/internal node*) dan node berbentuk persegi adalah simpul daun/*subset (leaf node)*. Hasil perhitungan probabilitas dari record pada atribut x_1 lebih besar dari record pada atribut w_{10} sehingga dipilih sebagai simpul akar dan memiliki 2 simpul percabangan dengan kondisi nilai *Yes* dan *No*. Simpul cabang yang bernilai *No* diklasifikasikan kedalam subset C_1 sedangkan simpul cabang yang bernilai *Yes* diklasifikasikan kedalam subset C_2 namun belum termasuk ke dalam kelas yang sama sehingga dilakukan kembali perhitungan probabilitas secara rekursif untuk membentuk pohon subset dan menghasilkan kembali 2 simpul percabangan yang memiliki nilai *Yes* dan *No*, semua subset C_1 dan C_2 masuk ke dalam kelas yang sama, maka setelah dilakukan split pertumbuhan tree dihentikan.

2.2. Model Decision Tree C4.5

Dalam pendekatan heuristic C45 menggunakan 2 pendekatan dalam pengujian yang dilakukan dengan peringkat probabilitas sebagai berikut:

1. Information gain yang meminimalkan nilai total entropy dari subset yang terjadi bias pada saat pengujian menggunakan data numeric.

2. Gain ratio melakukan pembagian information gain dengan menggunakan informasi entropy dari tiap-tiap atributnya.

Algoritma C4.5 merupakan sebuah jenis lain dari metode decision tree yang mirip dengan struktur *flowchart*, pada masing-masing internal *nodenya* dinyatakan sebagai atribut pengujian. Cabang yang mewakili *output* dengan pengujian node-node (*leafnode*) dengan melakukan penentuan class. Bagian paling atas node dari pohon merupakan node root. Tahapan dari algoritma C4.5 yaitu berikut ini: (Mitchell, 1997)

- (1) Perhitungan dari nilai *Entropy* pada tiap-tiap atribut :

$$\text{Entropy}(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (2.1)$$

Dimana:

S = Himpunan Kasus

n = Jumlah Partisi S

p_i = Proporsi Subset dari S pada partisi ke- i

- (2) Perhitungan dari nilai *Information Gain* pada tiap-tiap atribut:

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i) \quad (2.2)$$

Dimana:

S = Himpunan Kasus

A = Atribut Subset

n = Jumlah banyaknya data dari Atribut A

$|S_i|$ = Ukuran Subset dari Himpunan Kasus yang dimiliki atribut A pada data ke- i

$|S|$ = Ukuran Jumlah masalah pada Himpunan Kasus

- (3) Perhitungan dari nilai *Split Information* untuk tiap-tiap atribut :

$$\text{SplitInfo}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (2.3)$$

Dimana:

D = Keseluruhan Dataset

A = Atribut Subset

v = Jumlah Partisi Atribut A

$|D_j|$ = Ukuran Subset dari Dataset yang dimiliki atribut A partisi ke- j

$|D|$ = Ukuran Jumlah Kasus dalam Dataset

(4) Menghitung nilai *Gain Ratio* untuk masing – masing atribut :

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{SplitInfo\ (A)} \quad (2.4)$$

2.3. Average Gain

a. Metode *Information Gain*

Pada metode ini yang dikenalkan seorang Quinlan yang didasari oleh model ID3. Metode ini cocok dengan dataset yang variable diskrit tapi tidak cocok dengan data yang memiliki missing value. Metode ini dilakukan dengan memilih split atribut yang dinamakan Gain. Penyampaian informasi bergantung dengan robabilitas yang diukur dengan menggunakan bits sebagai minus algoritma 2. Dengan contoh lain $\log_2(1/8) = 3\ bits$. agar didapatkannya hasil yang diharap yang berkaitan dengan class yang ada, maka dilakukan proses penjumlahan dengan semua class dengan bagian frekuensi S yang dapat dilihat dibawah ini:

$$Info\ (S) = - \sum_{j=1}^k \frac{freq(cj.s)}{|S|} \times Info_2\left(\frac{freq(cj.s)}{|S|}\right) bits \quad (2.5)$$

Dalam penerapan sebuah permasalahan, info (T) di ukur pada rata – rata dari sebuah informasi teridentifikasi class pada sebuah kasus T. yang disebut dengan *Entropy* (S).

Dilakukan perbandingan terhadap data yang mirip setelah T selesai dipartisi yang sama dengan nilai n dari hasil percobaan pada X. nilai yang digunakan untuk menetapkan menggunakan pembobotan dari jumlah subset.

$$Info_x(x) = \sum_i^n = 1 \frac{|T_i|}{|T|} \times info(T_i) \quad (2.6)$$

Adapun keterangan persamaan diatas adalah sebagai berikut ini:

- n = jumlah data dari *subset*
- T = jumlah atribut dari dataset
- T_i = *subset* pada atribut dari dataset

Nilai informai yang diinginkan disebut dengan *Entropy*, adapun nilai *Entropy* dapat dilihat pada persamaan berikut:

$$Gain(x) = info\ (S) - Info\ x\ (T) \quad (2.7)$$

Perhitungan dari persamaan diatas memiliki keseuaian antara partisi T dan tes X, sehingga dapat dipilih atribut mana yang memiliki nilai information gain yang besar.

b. Metode *Info Gain Ratio*

Metode Info Gain Ratio merupakan pengembangan dari ID3 yang disebut dengan C4.5, yang mana dalam pemilihan *split* atributnya menggunakan metode *Info Gain Ratio* yang digantikan dari *Info Gain*. Metode C4.5 ini mampu bekerja dari variable yang berlanjut dan *missing value*.

Info Gain Ratio memiliki persamaan yaitu:

$$\text{gain ratio}(X) = \text{gain}(X) / \text{split info}(X) \quad (2.8)$$

Split info memiliki persamaan rumus berikut:

$$\text{Split info}(x) = - \sum_{i=1}^n \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right) \quad (2.9)$$

2.4. Teknik Klasifikasi

Proses dalam menganalisa sebuah data dalam menemukan model yang dapat menguraikan atau mengelompokkan data-data kelas yang penting yang digunakan untuk memprediksi kelas dari objek yang tidak diketahui kelasnya merupakan proses yang dilakukan sebagai teknik klasifikasi. Sehingga modelnya ditemukan dengan cara analisa data yang diuji atau yang terdapat pada objek data yang diketahui kelasnya (Han, 2012). Model – model tersebut berupa algoritma klasifikasi yang pada umumnya sering digunakan dalam menganalisa data yang termasuk K-NN, *Genetic Algorithm*, *Rule Base*, C4.5, Naïve Bayesian dan lain sebagainya.

Teknik klasifikasi didasarkan pada empat komponen utama yaitu:

1. *Class label attribute*.

Dalam proses ini melakukan pengkategorian untuk mempresentasikan label yang ada pada objek data. Misalnya: resiko penyakit, kresit maupun jenis pinjaman, dan lain sebagainya.

2. *Predictor*

Variable ini mempresentasikan karakteristik pada atribut data. Misalnya: merokok atau tidak, pergi kesekolah atau tidak, dan lain sebagainya.

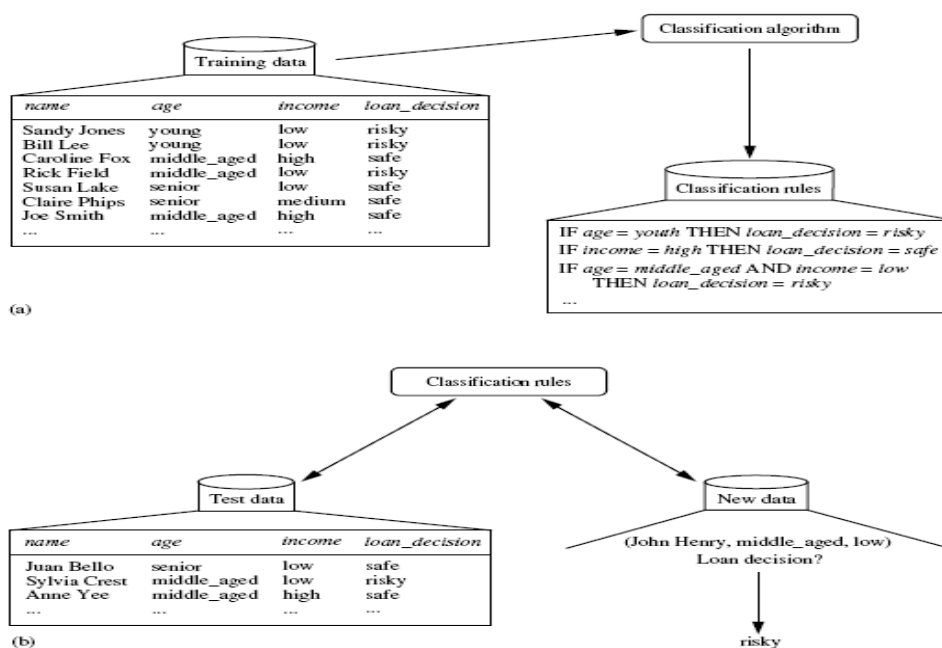
3. Training Dataset

Sebuah dataset yang bernilai dari kelas komponen dan *predictor* digunakan dalam penentuan kelas yang cocok yang dasarnya yaitu *predictor*.

4. Testing Dataset

Dalam pengklasifikasian pada model predictor menggunakan data baru yang hasilnya dari pengukuran akurasi klasifikasi dari metode evaluasi.

Pengklasifikasian data dapat diilustrasikan pada Gambar 2.4 berikut ini.



Gambar 2.3 Proses Klasifikasi (Sumber: Han, et al. 2012)

Pada Gambar 2.3. diatas menjelaskan proses pembelajaran dengan data uji yang dianalisa dengan menerapkan algoritma klasifikasi. Atribut keputusan menjadi sebuah label kelas dan model klasifikasi dipresentasikan dalam bentuk rule klasifikasi. Sedangkan proses klasifikasi selanjutnya yang dipakai dalam pengestimasi keakurasian dari aturan klasifikasi yang dapat dihasilkan. Jika akurasi yang dihasilkan maka aturan dapat diperoleh dari data yang baru. (Han, et al. 2012).

2.5. Penelitian Terdahulu yang Relevan

Pada penelitian ini juga pernah dilakukan dalam penelitian sebelumnya dapat dilihat sebagai berikut:

Tabel 2.1 Penelitian Terdahulu

No.	Nama Peneliti dan Tahun	Metode	Hasil Penelitian
1.	Zhang (2012)	Penerapan berbagai jenis algoritma decision tree telah banyak dilakukan, kajian terhadap penelitian-penelitian tersebut sangat bervariasi sesuai dengan bidang permasalahan yang diteliti oleh peneliti lain, peneliti mengenai penerapan algoritma decision tree pada Optimasi Algoritma C4.5 melalui metode split atribut dan <i>pruning</i> yang disebut: <i>heterogeneous cost sensitive learning (HCSL)</i> dan <i>Theshold pruning</i> untuk mereduksi permasalahan misklasifikasi dan <i>over-fitting</i> .	Kombinasi metode <i>splitting</i> atribut dan <i>pruning</i> menjadikan performa parameter <i>missing rate</i> sebesar 20% dari dataset dan misklasifikasi antara 100 sampai 600.
2.	Sivapriya & Nadira (2013)	Penelitian silang (<i>hybrid</i>) yaitu: reduksi fitur PCA dan seleksi fitur gain-ratio Decision Tree C4.5 untuk meningkatkan klasifikasi Support Vector Machine (SVM) bagi data medis. Data yang digunakan adalah citra MRI (Magnetic Resonance Images) untuk diklasifikasikan melalui neuropsychological test.	Metode silang ini memiliki nilai kontribusi untuk performa akurasi klasifikasi SVM sebesar 97%.
3.	Hussain (2015)	Pendekatan <i>Principal Component Analysis (PCA)</i> sebagai metode seleksi fitur untuk mereduksi indikator-indikator yang berhubungan dengan prediksi tingkat ketahanan hidup pasien terinfeksi kanker payudara. Data yang digunakan berasal dari <i>SEER dataset</i> sebanyak 684.394 rekam medis pasien.	Metode seleksi fitur PCA yang diusulkan memperoleh akurasi sebesar 92%.
4.	Dang (2016)	Optimasi teknik prediksi penyakit tumor dengan kombinasi Metode <i>Principal Component Analysis</i> untuk ekstraksi fitur dari data DNA microarray, metode Decision Tree (ID3) untuk seleksi fitur tanpa metode <i>Pruning</i> dan metode <i>Multi-Layer Perceptron (MLP)</i> .	Menerapkan metode reduksi fitur untuk teknik prediksi penyakit dengan kombinasi ekstraksi fitur <i>Principal Component Analysis</i> , ID3 serta MLP

Tabel 2.1 Penelitian Terdahulu (lanjutan)

5.	Kavitha & Kannan (2016)	Penelitian silang (hybrid) yaitu: reduksi fitur <i>Principal Component Analysis</i> dan seleksi fitur subset menggunakan algoritma Wrapper filter dan algoritma decision tree C4.5. Data yang digunakan adalah UCI Heart dataset yang terdiri dari 500 rekam medis pasien dan 15 atribut/indikator	Penelitian ini menonjolkan hasil reduksi fitur dari PCA tanpa memperhitungkan hasil akurasi dari pengklasifikasi algoritma decision tree C4.5.
----	-------------------------	--	--

BAB III

METODE PENELITIAN

3.1 Data Yang Digunakan

Pada penelitian ini, untuk mengetahui kinerja dari metode *decision tree* C4.5 maka digunakan data observasi berupa dataset UCI *Cervical Cancer* yang berasal dari (<http://archive.ics.uci.edu/ml>) dan Dinas Kesehatan Kabupaten Langkat. Data kanker serviks ini melibatkan 858 sampel dengan 32 faktor serta empat label kelas meliputi: *Hinselmann*, *Schiller*, *Cytology* dan *Biopsy* yang telah dipublikasikan oleh Fernandes *et al.* serta terdiri dari 716 sampel data dengan 5 atribut, dan 2 kelas. Penelitian ini berfokus pada target *Biopsy* seperti yang direkomendasikan oleh tinjauan literatur

Adapun deskripsi dari dataset UCI *Cervical Cancer* dapat dilihat pada tabel 3.1 sebagai berikut:

Tabel 3.1 Dataset UCI *Cervical Cancer*

Attributes	Type	Attributes	Type	Attributes	Type	Attributes	Type
Age	Int	Hormonal Contraceptives (years)	Int	STDs:vulvo-perineal condylomatosis	Bool	STDs: HPV	Bool
Number of sexual partnes	Int	Intra Uterine Device (IUD)	Bool	STDs: syphilis	Bool	STDs: number of diagnosis	Int
Age of first sexual Intercourse	Int	IUD (years)	Int	STDs: pelvic inflammatory disease	Bool	STDs: time since first diagnosis	Int
Number of pregnancies	Int	Sexually Transmitted Diseases (STDs)	Bool	STDs: genital herpes	Bool	STDs: time since last diagnosis	Int
Smokes	Bool	STDs (number)	Int	STDs: molluscum contagiosum	Bool	Dx: Cancer	Bool
Smokes (years)	Bool	STDs: condylomatosis	Bool	STDs: AIDS	Bool	Dx:CIN (cervical intraepithelial neoplasia)	Bool
Smokes (packs/year)	Bool	STDs: cervical condylomatosis	Bool	STDs: HIV	Bool	Dx: HPV (Human Papillomavirus)	Bool
Hormonal Contraceptives	Bool	STDs: vaginal condylomatosis	Bool	STDs: Hepatitis B	Bool	Dx (diagnosis diseases)	Bool

Sumber : <http://archive.ics.uci.edu/ml>

Berdasarkan tabel 3.1, Masing-masing atribut memiliki sejumlah *tuple* yang proporsional yang terdiri dari 32 indikator, Berikut ini merupakan representasi dari

Faktor labelnya:

X₁: Age

X₂: Number of sexual partners

X₃: First sexual intercourse

X₄: Number of pregnancies

X₅: Smoke

X₆: Smoke (years)

X₇: Smoke (packs/year)

X₈: Hormonal contraceptives

X₉: Hormonal contraceptives (years)

X₁₀: IUD

X₁₁: IUD (years)

X₁₂: STDs

X₁₃: STDs(number)

X₁₄: STDs: condylomatosis

X₁₅: STDs: cervical condylomatosis

X₁₆: STDs: vaginal condylomatosis

X₁₇: vulvo-perineal condylomatosis

X₁₈: STDs: syphilis

X₁₉: pelvic inflammatory disease

X₂₀: genital herpes

X₂₁: STDs: molluscum contagiosum

X₂₂: STDs:AIDS

X₂₃: STDs:HIV

X₂₄: STDs:Hepatitis B

X₂₅: STDs:HPV

X₂₆: STDs: Number of diagnosis

X₂₇: STDs: Time since first diagnosis

X₂₈: STDs: Time since last diagnosis

X₂₉: Dx : Cancer

X₃₀: Dx : CIN

X₃₁: Dx:HPV

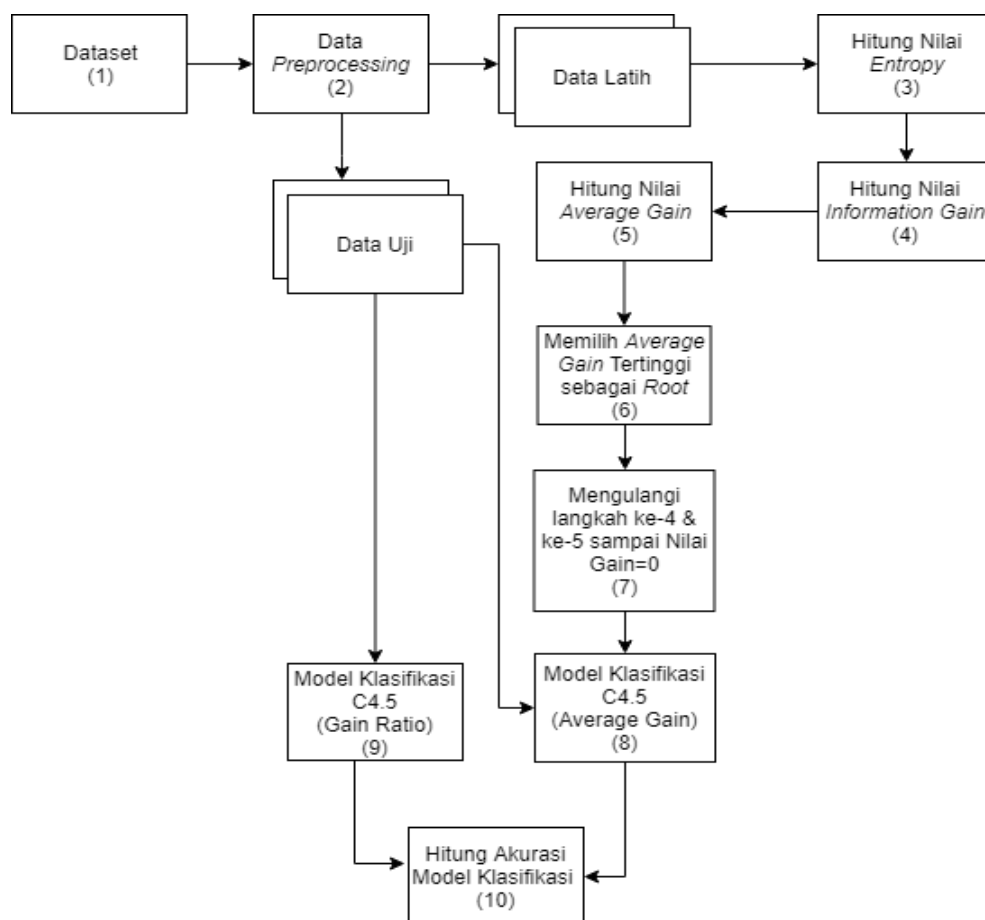
X₃₂: Dx

X₃₃: Class Biopsy

Dataset *Cervical Cancer* tersebut juga mempunyai karakteristik *missing value* dan data yang duplikat sehingga terlebih dahulu dilakukan data *preprocessing* guna meningkatkan tingkat akurasi pengklasifikasiannya melalui partisi data yang akan dibagi menjadi 90% sebagai data *training* dan 10% sebagai data *testing*.

3.2. Peningkatan kinerja algoritma C4.5 dan Metode Average Gain

Adapun tahapan-tahapannya secara garis besar dapat dilihat pada gambar 3.1 :



Gambar 3.1 Tahapan Penerapan

3.3. Data Preprocessing

Dalam melakukan proses *preprocessing*, yaitu *missing value* dan *cleaning data*, pada *missing value* yaitu nilai factor yang numerik diganti menjadi rata-rata nilai (mean) dari faktor didalam kolom yang sama. Kemudian *missing value* yaitu nilai factor yang nominal diganti menjadi nilai yang banyak dari faktor dikolom yang sama, dan kemudian *cleaning process* caranya dengan membuang duplikat data yang menjadi jumlah observasi awalnya banyaknya 858 record yang menjadi 827 record.

3.3.1. Missing Value

Missing Value Missing value adalah informasi yang tidak tersedia untuk sebuah objek (kasus). Missing value terjadi karena informasi untuk sesuatu tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. Missing value pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jika jumlahnya hanya sedikit, misal hanya 1 % dari seluruh data. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah data yang mengandung banyak missing tersebut masih layak diproses lebih lanjut ataukah tidak.

Pengisian missing value menggunakan metode mean yang dapat dilihat sebagai berikut:

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (3.1)$$

Dengan keterangan sebagai berikut:

\bar{x} = Rata-rata hitung (mean)

x = Data

n = jumlah data

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

$$\bar{x} = \frac{\sum x}{n} = \frac{18+4+15+1+0}{32}$$

$$\bar{x} = 1.1875 \approx 1$$

Tabel 3.2. Tabel Sebelum Mising Value

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	X33
18	4	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
15	1	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
34	1	?	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
52	5	16	4	1	37	37	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	1	0	1	0	0
46	3	21	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
42	3	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
51	3	17	6	1	34	3	0	0	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	1
26	1	26	3	0	0	0	1	2	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
45	1	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	1	0	1	1	0
44	3	15	?	1	1	3	0	0	?	?	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
44	3	26	4	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
27	1	17	3	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
45	4	14	6	0	0	0	1	10	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
44	2	25	2	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
43	2	18	5	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
40	3	18	2	0	0	0	1	15	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
41	4	21	3	0	0	0	1	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
43	3	15	8	0	0	0	1	3	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0
...
29	2	20	1	0	0	0	1	1	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?	0	0	0	0	0

Tabel 3.3 Data Sesudah Missing Value

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32	X33
18	4	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
15	1	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
34	1	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
52	5	16	4	1	37	37	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0
46	3	21	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
42	3	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
51	3	17	6	1	34	3	0	0	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
26	1	26	3	0	0	0	1	2	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
45	1	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0
44	3	15	1	1	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
44	3	26	4	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
27	1	17	3	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
45	4	14	6	0	0	0	1	10	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
44	2	25	2	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
43	2	18	5	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
40	3	18	2	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
41	4	21	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
43	3	15	8	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
29	2	20	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

3.3.2 Proses Pembentukan Model Klasifikasi *Decision Tree C4.5*

Dalam proses pembentukan model klasifikasi decision tree C4.5, hasil data *preprocessing* yakni *data cleaning* dari dataset UCI *Cervical Cancer* yang diperoleh sebanyak 827 data observasi kemudian dibagi menjadi 90% data sebagai data latih dan 10% data sebagai data uji. Tabel 3.2 adalah data uji yang digunakan pada pengujian model klasifikasi Decision Tree C4.5:

Tabel 3.4 Data Uji dataset *Cervical Cancer*

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
1	45	1	20	5	0	0	0	0	0	0	...	1
2	43	2	18	4	0	0	0	1	15	0	...	1
3	40	3	15	3	0	0	0	1	3	0	...	1
4	41	4	16	5	0	0	0	1	15	0	...	1
5	33	3	21	6	1	7	2	1	0	0	...	1
6	35	2	18	6	0	0	0	1	1	0	...	1
7	35	1	21	1	0	0	0	1	5	0	...	1
8	35	3	16	5	0	0	0	1	4	1	...	1
9	31	1	20	5	0	0	0	1	2	0	...	1
10	33	5	19	1	1	4	0	1	2	0	...	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
83	26	8	15	1	1	9	1	1	5	1	...	0

Sebelum dilakukan proses perhitungan entropy, maka terlebih dahulu dibentuk kedalam kategori untuk data yang bertipe numerik, Adapun hasil pembentukan kategori seperti berikut:

Tabel 3.5 Pemberian Kategori dataset *Cervical Cancer*

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
1	<20	≤7	≤17	≤4	0	≤13	≤13	0	≤13	0	...	1
2	<20	≤7	≤17	≤4	0	≤13	≤13	0	≤13	0	...	1
3	<40	≤7	≤17	≤4	0	≤13	≤13	0	≤13	0	...	1
4	≥50	≤7	≤17	≤4	1	>26	>26	1	≤13	0	...	1
5	<50	≤7	≤24	≤4	0	≤13	≤13	1	≤26	0	...	1
6	<50	≤7	≤24	≤4	0	≤13	≤13	0	≤13	0	...	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
827	<30	≤7	≤24	≤4	0	≤13	≤13	1	≤13	0	...	0

Tabel 3.6 Data Uji dataset Kanker Rahim

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
1	30	0	0	1	0	0	1
2	29	0	0	0	0	0	0
3	29	0	0	0	0	0	0
4	29	0	0	0	0	0	0
5	24	0	1	0	0	0	1
6	25	0	0	0	0	1	1
7	26	0	0	0	0	0	0
8	26	0	0	0	0	0	0
9	26	0	0	0	0	0	0
10	27	0	1	0	0	0	1
11	28	0	0	0	0	0	0
12	29	0	0	0	0	0	0
13	29	0	0	0	0	0	0
14	24	0	0	0	0	0	0
15	24	0	0	0	0	0	0
16	24	0	1	0	0	0	1
17	25	0	0	0	0	0	0
18	26	0	0	0	1	0	1
19	26	0	0	0	0	0	0
20	24	0	0	0	0	0	0
21	24	0	0	0	0	1	1
22	27	0	1	0	0	0	1
23	26	0	0	0	0	0	0
24	26	0	0	1	0	0	1
25	28	0	0	1	0	0	1
26	28	0	0	0	0	0	0
27	28	0	0	1	0	0	1
28	28	0	0	0	0	0	0
...
716	57	0	0	1	0	0	1

X₁: UmurX₂: NegatifX₃: PositifX₄: Curiga KankerX₅: Kelainan RahimX₆: Kanker Leher RahimX₇: Keluhan Rahim

Berikut merupakan tahapan algoritma C4.5: (Mitchell, 1997)

(1) Perhitungan nilai *Entropy* pada masing – masing atribut :

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (3.2)$$

Dimana:

S = Himpunan Kasus

n = Jumlah Partisi S

p_i = Proporsi Subset dari S pada partisi ke- i

Adapun perhitungan entropy total dari masing-masing kelas dengan persamaan (3.2) adalah sebagai berikut:

$$\begin{aligned} Entropy(S) &= \sum_{i=1}^n - p_i * \log_2 p_i \\ &= ((-726/827)*\text{Log}_2 (726/827))+(-101/827)*\text{Log}_2 (101/827)) \\ &= -0.87*-0.19+(-0.122)*-3.04 \\ &= 0.166+0.37 \\ &= 0.53 \end{aligned}$$

Jumlah Data = 827

Class 0 = 726

Class 1 = 101

Nilai Entropy = 0.53

Adapun hasil perhitungan entropy untuk masing-masing atribut dengan persamaan (3.2) adalah sebagai berikut :

1. Entropy dari atribut (Age):

$$\begin{aligned} Entropy(S) &= \sum_{i=1}^n - p_i * \log_2 p_i \\ X_1 < 20 &= ((-151/157)*\text{Log}_2 (151/157))+(-6/157)*\text{Log}_2 (6/157)) \\ &= -0.962*-0.056+(-0.038)*-4.717 \\ &= 0,054+0.1792 \\ &= 0.234 \\ 20 < X_1 \leq 30 &= ((-359/386)*\text{Log}_2 (359/386))+(-27/386)*\text{Log}_2 (27/386)) \\ &= -0.930*-0.1046+(-0.069)*-3.857 \\ &= 0.097+0.266 \\ &= 0.365 \end{aligned}$$

$$\begin{aligned}
30 < X_1 \leq 40 &= ((-204/219) * \text{Log}_2 (204/219) + (-15/219) * \text{Log}_2 (15/219)) \\
&= -0.931 * -0.1031 + (-0.068) * -3.878 \\
&= 0.0959 + 0.2637 \\
&= 0.36
\end{aligned}$$

$$\begin{aligned}
40 < X_1 < 50 &= ((-52/56) * \text{Log}_2 (52/56) + (-4/56) * \text{Log}_2 (4/56)) \\
&= -0.928 * -0.1078 + (-0.071) * -3.816 \\
&= 0.100 + 0.2709 \\
&= 0.371
\end{aligned}$$

$$\begin{aligned}
X_1 \geq 50 &= ((-7/9) * \text{Log}_2 (7/9) + (-2/9) * \text{Log}_2 (2/9)) \\
&= -0.777 * -0.364 + (-0.222) * -2.171 \\
&= 0.282 + 0.481 \\
&= 0.764
\end{aligned}$$

$$\begin{aligned}
\text{Entropy Total (Age)} &= 157/827 * (0.234) + 386/827 * (0.365) + 219/827 * (0.36) + \\
&\quad 56/827 * (0.371) + 9/827 * (0.764) \\
&= 0.343
\end{aligned}$$

2. Entropy dari Atribut (Number of Sexual Partners):

$$\begin{aligned}
X_2 \leq 7 &= ((-54/819) * \text{Log}_2 (54/819) + (-765/819) * \text{Log}_2 (765/819)) \\
&= -0.066 * -3.921 + (-0.934) * -0.098 \\
&= 0.258 + 0.091 \\
&= 0.350
\end{aligned}$$

$$\begin{aligned}
7 < X_2 \leq 14 &= ((-5/6) * \text{Log}_2 (5/6) + (-1/6) * \text{Log}_2 (1/6)) \\
&= -0.83 * -0.268 + (-0.17) * -2.56 \\
&= 0.222 + 0.44 \\
&= 0.65
\end{aligned}$$

$$\begin{aligned}
X_2 > 14 &= ((-1/2) * \text{Log}_2 (1/2) + (-1/2) * \text{Log}_2 (1/2)) \\
&= -0.5 * -1 + (-0.5) * -1 \\
&= 0.5 + 0.5 \\
&= 1
\end{aligned}$$

$$\begin{aligned}
\text{Entropy Total (Number of Sexual Partners)} &= 819/827 * (0.350) + 6/827 * (0.65) + 2/827 * (1) \\
&= 0.353
\end{aligned}$$

3. Entropy dari Atribut (First Sexual Intercourse):

$$\begin{aligned}
 X_3 \leq 17 &= ((-500/532) * \text{Log}_2 (500/532) + (-32/532) * \text{Log}_2 (32/532)) \\
 &= -0.94 * -0.09 + (-0.06) * -4.05 \\
 &= 0.08 + 0.243 \\
 &= 0.32 \\
 17 < X_3 \leq 24 &= ((-250/271) * \text{Log}_2 (250/271) + (-21/271) * \text{Log}_2 (21/271)) \\
 &= -0.922 * -0.117 + (-0.077) * -3.698 \\
 &= 0.109 + 0.284 \\
 &= 0.393 \\
 X_3 > 24 &= ((-23/24) * \text{Log}_2 (23/24) + (-1/24) * \text{Log}_2 (1/24)) \\
 &= -0.958 * -0.062 + (-0.041) * -4.608 \\
 &= 0.059 + 0.188 \\
 &= 0.249
 \end{aligned}$$

Entropy Total (First Sexual Intercourse)

$$\begin{aligned}
 &= 532/827 * (0.32) + 271/827 * (0.393) + 24/827 * (0.249) \\
 &= 0.342
 \end{aligned}$$

4. Entropy dari atribut (Num of Pregnancies):

$$\begin{aligned}
 X_4 \leq 4 &= ((-712/764) * \text{Log}_2 (712/764) + (-52/764) * \text{Log}_2 (52/764)) \\
 &= -0.932 * -0.102 + (-0.068) * -3.878 \\
 &= 0.095 + 0.263 \\
 &= 0.358 \\
 4 < X_4 \leq 8 &= ((-59/61) * \text{Log}_2 (59/61) + (-2/61) * \text{Log}_2 (2/61)) \\
 &= -0.967 * -0.048 + (-0.032) * -4.965 \\
 &= 0.046 + 0.158 \\
 &= 0.208 \\
 X_4 > 8 &= ((-1/2) * \text{Log}_2 (1/2) + (-1/2) * \text{Log}_2 (1/2)) \\
 &= -0.5 * -1 + (-0.5) * -1 \\
 &= 0.5 + 0.5 \\
 &= 1
 \end{aligned}$$

Entropy Total (Num of Pregnancies)

$$= 764/827*(0.358) + 61/827*(0.208) + 2/827*(1) \\ = 0.348$$

(2) nilai *Information Gain* pada setiap atribut:

$$Gain(S, A) = Entrop(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \quad (3.3)$$

Dimana:

S = Keseluruhan Dataset

A = Atribut Subset

n = Jumlah Partisi Atribut A

$|S_i|$ = Ukuran Subset dari Dataset yang dimiliki atribut A pada partisi ke-i

$|S|$ = Ukuran Jumlah Kasus dalam Dataset

Adapun menghitung *Information Gain* dari Atribut (**Age**) dengan persamaan (3.3) adalah sebagai berikut:

$$Gain(\text{Total}, \text{Age}) = 0.53 - 0.34 = 0.19$$

Adapun menghitung *Information Gain* dari Atribut (**Number of Sexual Partners**) dengan persamaan (3.2) adalah sebagai berikut:

$$Gain(\text{Total}, \text{Number of Sexual Partners}) = 0.53 - 0.35 = 0.18$$

Adapun menghitung *Information Gain* dari Atribut (**First Sexual Intercourse**) dengan persamaan (3.2) adalah sebagai berikut:

$$Gain(\text{Total}, \text{First Sexual Intercourse}) = 0.53 - 0.34 = 0.19$$

Adapun menghitung *Information Gain* dari Atribut (**Num of Pregnancies**) dengan persamaan (3.3) adalah sebagai berikut:

$$Gain(\text{Total}, \text{Num of Pregnancies}) = 0.53 - 0.34 = 0.18$$

(3) menghitung nilai *Split Information* untuk masing-masing atribut :

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \quad (3.4)$$

Dimana:

D = Keseluruhan Dataset

A = Atribut Subset

v = Jumlah Partisi Atribut A

$|D_j|$ = Ukuran Subset dari Dataset yang dimiliki atribut A partisi ke-j

$|D|$ = Ukuran Jumlah Kasus dalam Dataset

Adapun penjumlahan *Split Information* dari Atribut (**Age**) dengan persamaan (3.4) adalah sebagai berikut:

$$\begin{aligned} &= ((-157/827 * \log_2(157/827)) + (-386/827 * \log_2(386/827)) + (-219/827 * \log_2(219/827)) + \\ &\quad (-56/827 * \log_2(56/827)) + (-9/827 * \log_2(9/827))) \\ &= 0.455623 + 0.513422 + 0.50844 + 0.26341 + 0.071064 \\ &= \mathbf{1.811959} \end{aligned}$$

Adapun penjumlahan *Split Information* dari Atribut (**Number of Sexual partners**) dengan persamaan (3.3) adalah sebagai berikut:

$$\begin{aligned} &= ((-819/827 * \log_2(819/827)) + (-6/827 * \log_2(6/827)) + (-2/827 * \log_2(2/827))) \\ &= 0.012874 + 0.051584 + 0.021016 \\ &= \mathbf{0.085474} \end{aligned}$$

Adapun penjumlahan *Split Information* dari Atribut (**First Sexual Intercourse**) dengan persamaan (3.3) adalah sebagai berikut:

$$\begin{aligned} &= ((-532/827 * \log_2(532/827)) + (-271/827 * \log_2(271/827)) + (-24/827 * \log_2(24/827))) \\ &= 0.4111705 + 0.360459 + 0.04666505 \\ &= \mathbf{0.8188295} \end{aligned}$$

Adapun penjumlahan *Split Information* dari Atribut (**Num of Pregnancies**) dengan persamaan (3.3) adalah sebagai berikut:

$$\begin{aligned} &= ((-764/827 * \log_2(764/827)) + (-61/827 * \log_2(61/827)) + (-2/827 * \log_2(2/827))) \\ &= 0.103996 + 0.2777285 + 0.021016 \\ &= \mathbf{0.402297} \end{aligned}$$

(4) Menghitung nilai *Average Gain* untuk masing – masing atribut :

$$\text{Average Gain}(A,T) = \frac{\text{Gain}(A,T)}{|A|} \quad (3.5)$$

Age $0.192/1.811959 = 0.105962663$

Number of Sexual partners $0.182/0.085474 = 2.129302478$

First Sexual Intercourse $0.193/0.8188295 = 0.235702304$

Num of Pregnances $0.187/0.402297 = 0.46483071$

- (5) Apabila atribut ada yang mempunyai *Average Gain* yang besar maka akan terpilih menjadi akar dan yang mempunyai nilai AG yang kecil dari akar maka akan menjadi cabang.
- (6) Melakukan perhitungan kembali nilai AG pada tiap atribut dengan tidak menambahkan atribut yang menjadi akar dari tahap sebelumnya.
- (7) Pada atribut yang memiliki AG yang besar akan menjadi cabang
- (8) Proses pengulangan pada langkah ke-4 dan ke-5 hingga menghasilkan nilai *Gain* = 0 pada atribut yang tersisa

3.3.3 Pengujian Akurasi menggunakan Confusion Matrix

Pengklasifikasi *decision tree* dari dataset UCI *Cervical Cancer* di uji, sehingga dilakukan perhitungan dalam pengujian kemudian dijelaskan pada tabel yang dinamakan *confusion matrix* (Witten, 20015). Semua tentang parameter baik dan buruk dalam pengklasifikasian dengan data uji pada kelas yang berbeda yaitu kelas negatif dan positif yaitu proses dari *confusion matrix*.

Tabel 3.7 *Confusion Matrix*

<i>Two-Class Prediction</i>		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

(Sumber: Witten, 2005)

Tabel 3.5 merupakan penjelasan dari parameter model dari klasifikasi 2 kelas yes dan no. jumlah klasifikasi yang berilai benar disebut dengan *True Positive* (TP) dan *True Negatives* (TN). Sedangkan prediksi menghasilkan nilai yang tak tepat atau mempunyai nilai positif pada saat melakukan proses prediksi diharapkan adalah negatif disebut dengan *False Positive* (FP). Apabila prediksi menghasilkan nilai yang tak tepat atau mempunyai nilai negatif pada saat melakukan proses prediksi yang diinginkan bernilai positif disebut dengan *False Negative* (FN). Kemudian hasil parameter *Confusion Matrix* disebut dengan akurasi, yang persamaan perhitungannya sebagai berikut: (Witten, 2005)

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (3.6)$$

TP (*True Positive*) merupakan banyaknya jumlah data didalam kelas positif dengan hasil prediksinya diklasifikasi dengan benar secara actual yang nilainya positif.

TN (*True Negative*) merupakan banyaknya jumlah data didalam kelas positif dengan hasil prediksinya diklasifikasi dengan benar secara actual yang nilainya negative.

FP (*False Positive*) merupakan banyaknya data yang salah satu dari kelas actual yang bernilai negative tapi hasil prediksi klasifikasinya pada kelas yang nilainya positif.

FN (*False Negative*) merupakan banyaknya data yang salah satu dari kelas actual yang bernilai positif tapi hasil prediksi klasifikasinya pada kelas yang nilainya negative.

3.4 Software dan Tools yang digunakan

Penelitian ini dibangun dengan dukungan perangkat lunak Rapid Miner® versi 5.3 dan menggunakan spesifikasi processor Intel Core i52.60GHz dengan kapasitas memory 4 GB. Untuk memudahkan perhitungan nilai-nilai dalam proses klasifikasi, pengujian dan evaluasi model klasifikasi *decision tree C4.5*, maka penelitian ini menggunakan dukungan perangkat lunak yaitu: Rapid Miner® versi 5.3.

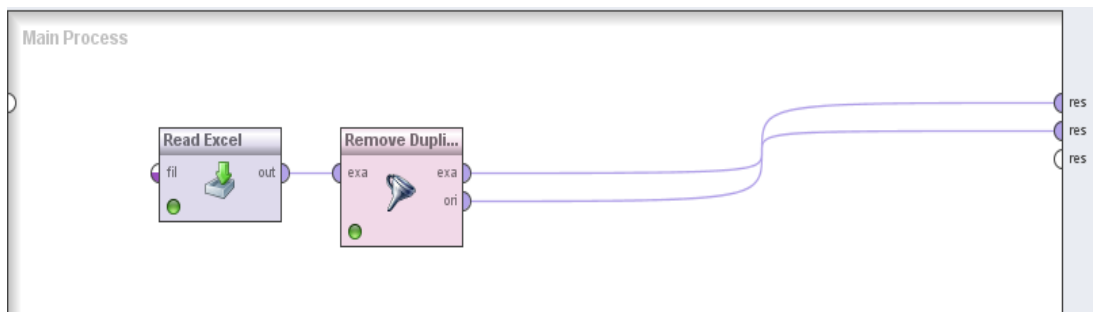
BAB IV

HASIL DAN KESIMPULAN

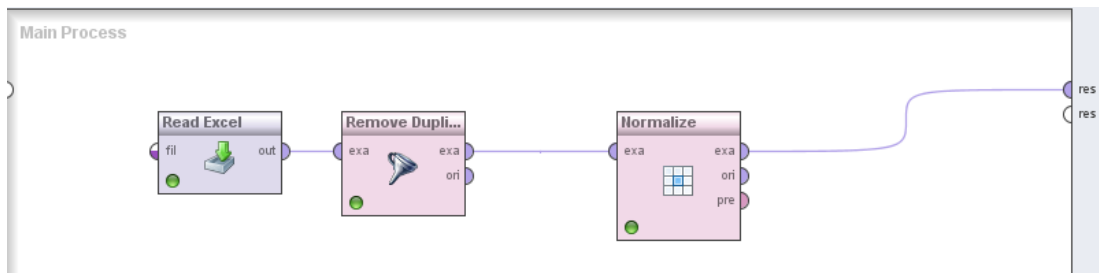
4.1. Hasil Pengujian

4.1.1. Persiapan Data Awal (Data Preprocessing)

Hasil *preprocessing* direpresentasikan kedalam bentuk label ($x_1, x_2, x_3 \dots x_{32}$) dengan perwakilan pengurutan data yang sama dengan data yang diuji.



Gambar 4.1 Proses *Remove Data Duplicate* (Rapidminer)



Gambar 4.2 Proses *Normalize Data* (Rapidminer)

Adapun hasil dari pre-processing adalah sebagai berikut:

Tabel 4.1 Hasil Data *Preprocessing* (*Cervical Cancer*)

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
1	18	4	15	1	0	0	0	0	0	0	...	0
2	15	1	14	1	0	0	0	0	0	0	...	0
3	34	1	17	1	0	0	0	0	0	0	...	0
4	52	5	16	4	1	37	37	1	3	0	...	0
5	46	3	21	4	0	0	0	1	15	0	...	0
6	42	3	23	2	0	0	0	0	0	0	...	0
7	51	3	17	6	1	34	3	0	0	1	...	1
8	26	1	26	3	0	0	0	1	2	1	...	0
9	45	1	20	5	0	0	0	0	0	0	...	0
10	44	3	15	2	1	1	3	0	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
827	29	2	20	1	0	0	0	1	0	0	...	0

Tabel 4.2 Hasil Data *Preprocessing* (*Kanker Rahim*)

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
1	30	0	0	1	0	0	1
2	29	0	0	0	0	0	0
3	29	0	0	0	0	0	0
4	29	0	0	0	0	0	0
5	24	0	1	0	0	0	1
6	25	0	0	0	0	1	1
7	26	0	0	0	0	0	0
8	26	0	0	0	0	0	0
9	26	0	0	0	0	0	0
10	27	0	1	0	0	0	1
11	28	0	0	0	0	0	0
12	29	0	0	0	0	0	0
13	29	0	0	0	0	0	0
14	24	0	0	0	0	0	0
15	24	0	0	0	0	0	0
16	24	0	1	0	0	0	1
17	25	0	0	0	0	0	0
18	26	0	0	0	1	0	1
19	26	0	0	0	0	0	0
20	24	0	0	0	0	0	0
21	24	0	0	0	0	1	1
22	27	0	1	0	0	0	1
23	26	0	0	0	0	0	0
24	26	0	0	1	0	0	1
25	28	0	0	1	0	0	1
26	28	0	0	0	0	0	0
27	28	0	0	1	0	0	1
28	28	0	0	0	0	0	0
...
716	57	0	0	1	0	0	1

Untuk hasil data *preprocessing* selengkapnya dapat dilihat pada lampiran 1

Dalam proses pembentukan model klasifikasi decision tree C4.5, hasil data *preprocessing* dari dataset UCI *Cervical Cancer* yang diperoleh sebanyak 827 data observasi kemudian dibagi menjadi 90% data sebagai data latih dan 10% data sebagai data uji. Tabel 4.2 adalah data uji yang digunakan pada pengujian model klasifikasi C4.5:

Tabel 4.3 Data Uji Dataset (*Cervical Cancer*)

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
1	45	1	20	5	0	0	0	0	0	0	...	1
2	43	2	18	4	0	0	0	1	15	0	...	1
3	40	3	15	3	0	0	0	1	3	0	...	1
4	41	4	16	5	0	0	0	1	15	0	...	1
5	33	3	21	6	1	7	2	1	0	0	...	1
6	35	2	18	6	0	0	0	1	1	0	...	1
7	35	1	21	1	0	0	0	1	5	0	...	1
8	35	3	16	5	0	0	0	1	4	1	...	1
9	31	1	20	5	0	0	0	1	2	0	...	1
10	33	5	19	1	1	4	0	1	2	0		1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
83	26	8	15	1	1	9	1	1	5	1	...	0

Untuk hasil selengkapnya dari data uji dataset Cervical Cancer dapat dilihat pada Lampiran 2

4.1.2. Pemberian Kategori

Sebelum dilakukan proses perhitungan entropy, maka terlebih dahulu dibentuk kedalam kategori untuk data yang bertipe numerik, Adapun hasil pembentukan kategori seperti berikut:

Tabel 4.4 Hasil Pemberian Kategori (*Cervical Cancer*)

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃
1	<20	≤7	≤17	≤4	0	≤13	≤13	0	≤13	0	...	1
2	<20	≤7	≤17	≤4	0	≤13	≤13	0	≤13	0	...	1
3	<40	≤7	≤17	≤4	0	≤13	≤13	0	≤13	0	...	1
4	≥50	≤7	≤17	≤4	1	>26	>26	1	≤13	0	...	1
5	<50	≤7	≤24	≤4	0	≤13	≤13	1	≤26	0	...	1
6	<50	≤7	≤24	≤4	0	≤13	≤13	0	≤13	0	...	1
7	≥50	≤7	≤17	≤8	1	>26	≤13	0	≤13	1	...	1
8	<30	≤7	>24	≤4	0	≤13	≤13	1	≤13	1	...	1
9	<50	≤7	≤24	≤8	0	≤13	≤13	0	≤13	0	...	1
10	<50	≤7	≤17	≤4	1	≤13	≤13	0	≤13	0		1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
827	<30	≤7	≤24	≤4	0	≤13	≤13	1	≤13	0	...	0

Tabel 4.5 Hasil Pemberian Kategori (Kanker Rahim)

No.	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
1	< 30	0	0	1	0	0	1
2	< 30	0	0	0	0	0	0
3	< 30	0	0	0	0	0	0
4	< 30	0	0	0	0	0	0
5	< 30	0	1	0	0	0	1
6	< 30	0	0	0	0	1	1
7	< 30	0	0	0	0	0	0
8	< 30	0	0	0	0	0	0
9	< 30	0	0	0	0	0	0
10	< 30	0	1	0	0	0	1
11	< 30	0	0	0	0	0	0
12	< 30	0	0	0	0	0	0
13	< 30	0	0	0	0	0	0
14	< 30	0	0	0	0	0	0
15	< 30	0	0	0	0	0	0
16	< 30	0	1	0	0	0	1
17	< 30	0	0	0	0	0	0
18	< 30	0	0	0	1	0	1
19	< 30	0	0	0	0	0	0
20	< 30	0	0	0	0	0	0
21	< 30	0	0	0	0	1	1
22	< 30	0	1	0	0	0	1
23	< 30	0	0	0	0	0	0
24	< 30	0	0	1	0	0	1
25	< 30	0	0	1	0	0	1
26	< 30	0	0	0	0	0	0
27	< 30	0	0	1	0	0	1
28	< 30	0	0	0	0	0	0
...
716	> 50	0	0	1	0	0	1

4.1.3. Hasil Nilai *Entropy*

Berikut merupakan hasil nilai entropy yang telah dibahas pada persamaan (3.1). adapun hasil dari nilai tersebut adalah sebagai berikut:

Tabel 4.6 Nilai Entropy Per Atribut (*Cervical Cancer*)

		Jumlah Data	ENTROPY	TOTAL ENTROPY
TOTAL		827	0.535447454	
Age (X ₁)				0.343997022
	X ₁ <20	157	0.234054566	
	20 ≤ X ₁ <30	386	0.365730263	
	30 ≤ X ₁ <40	219	0.360275056	
	40 ≤ X ₁ <50	56	0.371232327	
	X ₁ ≥ 50	9	0.764204507	

Tabel 4.6 Nilai Entropy Per Atribut (*Cervical Cancer*) lanjutan

Number of Sexual partners (X_2)				0.354307018
	$X_2 < 7$	819	0.350563821	
	$7 \leq X_2 < 14$	6	0.650022422	
	≥ 14	2	1	
First Sexual Intercourse (X_3)				0.347149869
	$X_3 < 17$	532	0.328041566	
	$17 \leq X_3 < 24$	271	0.393275473	
	$X_3 \geq 24$	24	0.249882293	
Num of Pregnances (X_4)				0.349104709
	$X_4 < 4$	764	0.358652561	
	$4 \leq X_4 < 8$	61	0.208180946	
	≥ 8	2	1	
Smokes (X_5)				0.347620299
	0	704	0.337290067	
	1	123	0.406746183	
smokes(years) (X_6)				0.345174759
	$X_6 < 10$	789	0.335740726	
	$10 \leq X_6 < 20$	30	0.468995594	
	≥ 30	8	0.811278125	
smokes(packs/years) (X_7)				0.354868558
	$X_7 < 10$	819	0.350563821	
	$10 \leq X_7 < 20$	5	0.721928095	
	≥ 30	3	0.918295834	
Hormonal Contraceptives (X_8)				0.347689479
	0	249	0.312273212	
	1	578	0.36294666	
Hormonal Contraceptives(years) (X_9)				0.35002907
	$X_9 < 10$	800	0.351770921	
	$10 \leq X_9 < 20$	25	0.242292189	
	≥ 30	2	1	
IUD (X_{10})				0.106969507
	0	744	0.014759253	
	1	83	0.933528902	
IUD(years) (X_{11})				0.017583904
	$5 < X_{11} < 10$	20	0.286396957	
	$X_{11} \geq 10$	5	0.721928095	
	$X_{11} \geq 5$	745	0.00698565	
STDs (X_{12})				0.345612616
	0	748	0.364125023	
	1	79	0.170330576	
STDs(Number) (X_{13})				0.348119766
	$X_{13} < 4$	827	0.348119766	

Tabel 4.6 Nilai Entropy Per Atribut (*Cervical Cancer*) lanjutan

STDs: Condylomatosis (X_{14})				0.347830659
	0	783	0.352386073	
	1	44	0.266764988	
STDs:Cervical Condylomatosis (X_{15})				0.348119766
	0	827	0.348119766	
STDs:vaginal condylomatosis (X_{16})				0.348119766
	0	827	0.348119766	
STDs:vulvo-perineal condylomatosis (X_{17})				0.352379769
	0	784	0.356910537	
	1	42	0.276195428	
STDs:syphilis (X_{18})				0.352712508
	0	809	0.353673006	
	1	18	0.309543429	
STDs:pelvic inflammatory disease (X_{19})				0.348119766
	0	827	0.348119766	
STDs:genital herpes (X_{20})				0.348119766
	0	827	0.348119766	
STDs:molluscum contagiosum (X_{21})				0.348119766
	0	827	0.348119766	
STDs:AIDS (X_{22})				0.348119766
	0	827	0.348119766	
STDs:HIV (X_{23})				0.348093594
	0	809	0.348951323	
	1	18	0.309543429	
STDs:Hepatitis B (X_{24})				0.348119766
	0	827	0.348119766	
Dx:HPV (X_{25})				0.348093594
	0	809	0.348951323	
	1	18	0.309543429	
STDs: Number of diagnosis (X_{26})				0.348653453
	0	756	0.361365804	
	1	71	0.18512476	
	2	2	1	
STDs:Time since first diagnosis (X_{27})				0.361263369
	$5 < X_{27} < 10$	13	0.391243564	
	$10 \leq X_{27} < 15$	12	0.41381685	
	$X_{27} \leq 5$	803	0.350825451	
	$X_{27} \geq 15$	7	0.591672779	
STDs:Time since last diagnosis (X_{28})				0.360054179
	$5 < X_{27} < 10$	13	0.391243564	
	$10 \leq X_{27} < 15$	12	0.41381685	

Tabel 4.6 Nilai Entropy Per Atribut (*Cervical Cancer*) lanjutan

	$X_{27} \leq 5$	803	0.350825451	
	$X_{27} \geq 15$	6	0.650022422	
Dx:Cancer (X_{29})				0.352712508
	0	809	0.353673006	
	1	18	0.309543429	
Dx:CIN (X_{30})				0.347890812
	0	818	0.34618139	
	1	9	0.503258335	
STDs:HPV (X_{31})				0.348119766
	0	827	0.348119766	
Dx (X_{32})				0.347896025
	0	803	0.350825451	
	1	24	0.249882293	

Tabel 4.7 Nilai Entropy Per Atribut (Kanker Rahim)

		Jumlah Data	Class (0)	Class (1)	ENTROPY	TOTAL ENTROPY
TOTAL		716	568	148	0.735134964	
Umur						0.7320976
	<30 Tahun	88	68	20	0.773226674	
	30 - 39 Tahun	303	237	66	0.756170149	
	40 - 50 Tahun	258	213	45	0.667713541	
	> 50 Tahun	67	50	17	0.817138776	
Negatif						0.0152591
	No Data	716	715	1	0.015259082	
Positif						0.7333299
	0	679	542	137	0.725448654	
	1	37	26	11	0.877962001	
Curiga Kanker						0.4015112
	0	704	660	44	0.337290067	
	1	123	113	10	0.406746183	
Kelainan Rahim						0.5855842
	0	670	568	102	0.615414675	
	1	46	1	45	0.151096971	
Kanker Leher Rahim						0.5855842
	0	670	568	102	0.615414675	
	1	46	1	45	0.151096971	

4.1.4. Hasil Nilai *Information Gain*

Berikut merupakan hasil nilai *Information Gain* yang telah dibahas pada persamaan (3.2). adapun hasil dari nilai tersebut adalah sebagai berikut:

Tabel 4.8 Hasil Perolehan *Information Gain* (*Cervical Cancer*)

		Jumlah Data	TOTAL ENTROPY	INFORMATION GAIN
TOTAL		827		
Age (X_1)			0.343997022	0.191450433
	$X_1 < 20$	157		
	$20 \leq X_1 < 30$	386		
	$30 \leq X_1 < 40$	219		
	$40 \leq X_1 < 50$	56		
	$X_1 \geq 50$	9		
Number of Sexual partners (X_2)			0.354307018	0.181140437
	$X_2 < 7$	819		
	$7 \leq X_2 < 14$	6		
	≥ 14	2		
First Sexual Intercourse (X_3)			0.347149869	0.188297586
	$X_3 < 17$	532		
	$17 \leq X_3 < 24$	271		
	$X_3 \geq 24$	24		
Num of Pregnancies (X_4)			0.349104709	0.186342746
	$X_4 < 4$	764		
	$4 \leq X_4 < 8$	61		
	≥ 8	2		
Smokes (X_5)			0.347620299	0.187827155
	0	704		
	1	123		
smokes(years) (X_6)			0.345174759	0.190272696
	$X_6 < 10$	789		
	$10 \leq X_6 < 20$	30		
	≥ 30	8		
smokes(packs/years) (X_7)			0.354868558	0.645131442
	$X_7 < 10$	819		
	$10 \leq X_7 < 20$	5		
	≥ 30	3		
Hormonal Contraceptives (X_8)			0.347689479	0.187757975
	0	249		
	1	578		

Tabel 4.8 Hasil Perolehan *Information Gain* (Cervical Cancer) lanjutan

Hormonal Contraceptives(years) (X_9)			0.35002907	0.185418384
	$X_9 < 10$	800		
	$10 \leq X_9 < 20$	25		
	≥ 30	2		
IUD (X_{10})			0.106969507	0.428477947
	0	744		
	1	83		
IUD(years) (X_{11})			0.017583904	0.517863551
	$5 < X_{11} < 10$	20		
	$X_{11} \geq 10$	5		
	$X_{11} \geq 5$	745		
STDs (X_{12})			0.345612616	0.189834839
	0	748		
	1	79		
STDs(Number) (X_{13})			0.348119766	0.187327689
	$X_{13} < 4$	827		
STDs: Condylomatosis (X_{14})			0.347830659	0.187616796
	0	783		
	1	44		
STDs:Cervical Condylomatosis (X_{15})			0.348119766	0.187327689
	0	827		
STDs:vaginal condylomatosis (X_{16})			0.348119766	0.187327689
	0	827		
STDs:vulvo-perineal condylomatosis (X_{17})			0.352379769	0.183067685
	0	784		
	1	42		
STDs:syphilis (X_{18})			0.352712508	0.182734947
	0	809		
	1	18		
STDs:pelvic inflammatory disease (X_{19})			0.348119766	0.187327689
	0	827		
STDs:genital herpes (X_{20})			0.348119766	0.187327689
	0	827		
STDs:molluscum contagiosum (X_{21})			0.348119766	0.187327689
	0	827		
STDs:AIDS (X_{22})			0.348119766	0.187327689
	0	827		
STDs:HIV (X_{23})			0.348093594	0.187353861
	0	809		
	1	18		

Tabel 4.8 Hasil Perolehan *Information Gain* (Cervical Cancer) lanjutan

STDs:Hepatitis B (X_{24})			0.348119766	0.187327689
	0	827		
Dx:HPV (X_{25})			0.348093594	0.187353861
	0	809		
	1	18		
STDs: Number of diagnosis (X_{26})			0.348653453	0.186794001
	0	756		
	1	71		
	2	2		
STDs:Time since first diagnosis (X_{27})			0.361263369	0.177640302
	$5 < X_{27} < 10$	13		
	$10 \leq X_{27} < 15$	12		
	$X_{27} \leq 5$	803		
	$X_{27} \geq 15$	7		
STDs:Time since last diagnosis (X_{28})			0.360054179	0.177932412
	$5 < X_{27} < 10$	13		
	$10 \leq X_{27} < 15$	12		
	$X_{27} \leq 5$	803		
	$X_{27} \geq 15$	6		
Dx:Cancer (X_{29})			0.352712508	0.182734947
	0	809		
	1	18		
Dx:CIN (X_{30})			0.347890812	0.187556642
	0	818		
	1	9		
STDs:HPV (X_{31})			0.348119766	0.187327689
	0	827		
Dx (X_{32})			0.347896025	0.18755143
	0	803		
	1	24		

Tabel 4.9 Hasil Perolehan *Information Gain* (Kanker Rahim)

		Jumlah Data	ENTROPY	TOTAL ENTROPY	INFORMATI ON GAIN
TOTAL		716	0.735134964		
Umur				0.7320976	0.00303735
	<30 Tahun	88	0.773226674		

Tabel 4.8 Hasil Perolehan *Information Gain* (*Cervical Cancer*) lanjutan

	30 - 39 Tahun	303	0.756170149		
	40 - 50 Tahun	258	0.667713541		
	> 50 Tahun	67	0.817138776		
Negatif				0.0152591	0.71987588
	No Data	716	0.015259082		
Positif				0.7333299	0.00180503
	0	679	0.725448654		
	1	37	0.877962001		
Curiga Kanker				0.4015112	0.33362381
	0	704	0.337290067		
	1	123	0.406746183		
Kelainan Rahim				0.5855842	0.14955076
	0	670	0.615414675		
	1	46	0.151096971		
Kanker Leher Rahim				0.5855842	0.14955076
	0	670	0.615414675		
	1	46	0.151096971		

4.1.5. Hasil Nilai *Split Info*

Berikut merupakan hasil nilai *Split Info* yang telah dibahas pada persamaan (3.3). adapun hasil dari nilai tersebut adalah sebagai berikut:

Tabel 4.10 Hasil Nilai *Split Info*(*Cervical Cancer*)

No	TOTAL	SPLIT INFO	No	TOTAL	SPLIT INFO
1	Age	1.8098	17	STDs:vulvo-perineal condylomatosis	0.29138
2	Number of Sexual partners	0.08647	18	STDs:syphilis	0.15124
3	First Sexual Intercourse	1.08508	19	STDs:pelvic inflammatory disease	0
4	Num of Pregnances	0.40404	20	STDs:genital herpes	0
5	Smokes	0.60665	21	STDs:molluscum contagiosum	0

Tabel 4.8 Hasil Perolehan *Information Gain* (Cervical Cancer) lanjutan

6	smokes(years)	0.30305	22	STDs:AIDS	0
7	smokes(packs/years)	0.08785	23	STDs:HIV	0.15124
8	Hormonal Contraceptives	0.88262	24	STDs:Hepatitis B	0
9	Hormonal Contraceptives(years)	0.21994	25	Dx:HPV	0.15124
10	IUD	0.47014	26	STDs: Number of diagnosis	0.44349
11	IUD(years)	0.31013	27	STDs:Time since first diagnosis	0.22405
12	STDs	0.45465	28	STDs:Time since last diagnosis	0.22405
13	STDs(Number)	0	29	Dx:Cancer	0.15124
14	STDs: Condylomatosis	0.29986	30	Dx:CIN	0.08659
15	STDs:Cervical Condylomatosis	0	31	STDs:HPV	0
16	STDs:vaginal condylomatosis	0	32	Dx	0.18946

Tabel 4.11 Hasil Nilai *Split Info*(Kanker Rahim)

No.	Total	Split Info
1	Umur	1.74717
2	Negatif	0
3	Positif	0.29347
4	Curiga Kanker	0.46054
5	Kelainan Rahim	0.34407
6	Kanker Leher Rahim	0.34407

Tabel 4.12 Rules Model Klasifikasi *Decision Tree C4.5* (Average Gain)

Rule No.	Rules	Class Pengujian
R1	if $Dx:cancer=0$ and Hormonal Contraceptives (years) ≤ 13 and $Dx=0$ and STDs:HIV = 0 and STDs (number) = 0 and IUD (years) ≤ 14 and Hormonal Contraceptives = 0 then	0

Tabel 4.12 Rules Model Klasifikasi *Decision Tree* C4.5 (*Average Gain*) lanjutan

R2	if STDs:genital herpes = 0 and Hormonal Contraceptives (years) <=13 and Dx = 0 and STDs:HIV = 0 and STDs (number) = 0 and IUD (years)<=14 and Hormonal Contraceptives = 1 then	1
⋮	⋮	⋮
R74	if STDs:genital herpes = 1 then	1

Rule R1 dan R2 memiliki class pengujian masing-masing bernilai 0 dan 1 artinya jika bernilai 0 maka diklasifikasikan sebagai class *unsuspected* (tanpa pengujian medis lanjutan) sebaliknya jika bernilai 1 maka diklasifikasikan sebagai *suspected* (pengujian medis lanjutan) bagi pasien yang dicurigai terinfeksi kerservik

4.2. Pengujian Akurasi menggunakan *Confusion Matrix*

Setelah diperoleh model klasifikasi C4.5 dalam bentuk *rules*, maka selanjutnya model klasifikasi tersebut diuji menggunakan data uji dataset *Cervical Cancer*. Tabel 4.8 adalah hasil pengujian model klasifikasi C4.5 menggunakan data uji dari dataset *Cervical Cancer*:

Tabel 4.13 Hasil Pengujian Model Klasifikasi C4.5 (*Cervical Cancer*)

X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	...	X ₃₃	Prediksi Class
45	1	20	5	0	0	0	0	0	0	...	1	0
43	2	18	4	0	0	0	1	15	0	...	1	0
40	3	15	3	0	0	0	1	3	0	...	1	0
41	4	16	5	0	0	0	1	15	0	...	1	0
33	3	21	6	1	7	2	1	0	0	...	1	0
35	2	18	6	0	0	0	1	1	0	...	1	0
35	1	21	1	0	0	0	1	5	0	...	1	0
35	3	16	5	0	0	0	1	4	1	...	1	0
31	1	20	5	0	0	0	1	2	0	...	1	0

Tabel 4.13 Hasil Pengujian Model Klasifikasi C4.5 (*Cervical Cancer*) lanjutan

33	5	19	1	1	4	0	1	2	0	...	1	0
31	3	17	2	0	0	0	0	0	0	...	0	0
30	3	16	3	0	1	0	1	2	0	...	0	0
33	1	16	4	0	0	0	0	0	0	...	0	0
31	3	14	4	0	0	0	1	4	0	...	0	0
29	5	20	2	0	0	0	1	2	0	...	0	1
30	2	22	2	0	0	0	1	0	0	...	0	0
30	3	14	3	0	0	0	1	12	1	...	0	0
29	4	10	5	0	0	0	0	0	0	...	0	0
26	10	16	1	1	9	0	1	2	0	...	0	0
27	1	26	1	0	0	0	0	0	0	...	0	0
28	3	15	6	1	14	2	1	7	1	...	0	1
25	2	17	2	0	0	0	1	3	0	...	0	0
29	4	16	5	0	0	0	1	6	0	...	0	0
25	6	18	1	0	0	0	1	2	0	...	0	0
22	3	17	1	0	0	0	0	0	0	...	0	0
21	2	16	1	0	0	0	0	0	0	...	0	0
21	2	19	1	0	0	0	1	1	0	...	0	0
26	1	18	2	0	0	0	0	0	0	...	0	0
22	2	18	2	0	0	0	0	0	0	...	0	0
21	4	15	2	0	0	0	0	0	0	...	0	0
23	1	18	2	0	0	0	1	1	0	...	0	0
24	1	16	3	0	0	0	1	0	0	...	0	0
21	1	17	2	0	0	0	1	3	0	...	0	0
22	2	19	1	0	0	0	1	2	0	...	0	0
21	1	18	1	0	0	0	1	0	0	...	0	0
27	1	15	4	0	0	0	1	2	1	...	0	0
31	1	17	5	0	0	0	1	1	0	...	0	0
20	5	14	3	1	4	3	1	2	0	...	0	0
19	3	18	1	0	0	0	1	1	0	...	0	0
20	1	15	1	1	2	0	1	2	0	...	0	0
18	1	17	1	0	0	0	0	0	0	...	0	0
19	2	15	2	0	0	0	1	0	0	...	0	0
18	1	16	2	0	0	0	0	0	0	...	0	0
19	1	16	1	0	0	0	0	0	0	...	0	0
21	1	15	3	0	0	0	1	0	0	...	0	0
17	3	15	1	0	0	0	0	0	0	...	0	0
18	1	14	2	0	0	0	0	0	0	...	0	0
18	2	17	1	0	0	0	1	0	0	...	0	0
14	2	14	1	0	0	0	1	2	0	...	0	0
17	5	15	1	0	0	0	0	0	0	...	0	0
15	2	15	1	0	0	0	0	0	0	...	0	0

Tabel 4.13 Hasil Pengujian Model Klasifikasi C4.5 (*Cervical Cancer*) lanjutan

31	2	18	2	0	0	0	0	0	0	...	0	0
25	2	16	1	0	0	0	1	2	0	...	0	0
19	1	17	1	0	0	0	1	1	0	...	0	0
20	1	18	1	0	0	0	1	0	0	...	0	0
18	2	17	1	0	0	0	0	0	0	...	0	0
42	2	18	2	0	0	0	1	0	1	...	0	0
18	3	16	1	0	1	0	1	2	0	...	0	0
40	2	21	4	0	0	0	1	0	0	...	0	1
21	1	15	2	0	0	0	1	0	0	...	0	0
36	1	28	1	1	16	2	0	0	0	...	0	0
44	2	25	1	0	0	0	0	0	0	...	0	0
23	3	16	1	0	0	0	0	0	0	...	0	0
19	2	15	2	0	0	0	1	1	0	...	0	0
19	2	17	1	0	0	0	1	0	0	...	0	0
23	1	15	3	0	0	0	0	0	0	...	0	0
21	1	14	1	0	0	0	1	5	0	...	0	0
38	3	22	2	0	1	0	1	3	1	...	0	1
52	2	19	5	0	0	0	0	0	1	...	0	0
24	1	21	2	0	0	0	1	1	0	...	0	0
19	2	16	2	1	1	0	1	2	0	...	0	0
21	5	17	1	0	0	0	0	0	0	...	0	0
21	1	17	1	0	0	0	1	5	0	...	0	0
40	3	17	2	0	0	0	1	20	1	...	0	0
40	1	20	7	0	0	0	0	0	0	...	0	0
24	2	13	5	0	0	0	1	2	1	...	0	0
18	3	13	1	0	0	0	0	0	0	...	0	0
19	2	17	1	0	0	0	1	2	0	...	0	0
26	2	18	1	0	0	0	0	0	0	...	0	0
33	3	18	2	0	1	0	0	0	0	...	0	0
23	2	16	1	0	0	0	1	6	0	...	0	0
35	3	18	3	0	0	0	1	5	0	...	0	0
26	8	15	1	1	9	1	1	5	1	...	0	0

Tabel 4.14 Hasil Pengujian Model Klasifikasi C4.5 (Kanker Rahim)

X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	Prediksi Class
30	0	0	1	0	0	1
29	0	0	0	0	0	0
29	0	0	0	0	0	0
29	0	0	0	0	0	0
24	0	1	0	0	0	1

Tabel 4.14 Hasil Pengujian Model Klasifikasi C4.5 (Kanker Rahim) lanjutan

25	0	0	0	0	1	1
26	0	0	0	0	0	0
26	0	0	0	0	0	0
26	0	0	0	0	0	0
27	0	1	0	0	0	1
28	0	0	0	0	0	0
29	0	0	0	0	0	0
29	0	0	0	0	0	0
24	0	0	0	0	0	0
24	0	0	0	0	0	0
24	0	1	0	0	0	1
25	0	0	0	0	0	0
26	0	0	0	1	0	1
26	0	0	0	0	0	0
24	0	0	0	0	0	0
24	0	0	0	0	1	1
27	0	1	0	0	0	1
26	0	0	0	0	0	0
26	0	0	1	0	0	1
28	0	0	1	0	0	1
28	0	0	0	0	0	0
28	0	0	1	0	0	1
28	0	0	0	0	0	0
...
57	0	0	1	0	0	1

Berdasarkan hasil selengkapnya dari tabel diatas, diperoleh jumlah data yang bernilai **True Positive** sebanyak 1 record karena hanya record tersebut yang memiliki jumlah aktual *class* dan prediksi *class* yang sama. Jumlah data yang bernilai **False Positive** diperoleh sebanyak 3 record. Jumlah data yang bernilai **False Negative** diperoleh sebanyak 9 record. Jumlah data yang bernilai **True Negative** diperoleh sebanyak 70 record. Sehingga dapat diperoleh hasil evaluasi model klasifikasi C4.5 dan C4.5 average gain dari dataset *Cervical Cancer* dengan menggunakan *confusion matrix* seperti pada tabel berikut:

Tabel 4.15 *Confusion Matrix C4.5(Cervical Cancer)*

Kinerja Klasifikasi	Predicted Class	
Actual Class	Predicted. Class 1	Predicted. Class 0
Actual. Class 1	3 (True Positive)	44 (False Negative)
Actual. Class 0	26 (False Positive)	654 (True Negative)

Berdasarkan tabel diatas, maka dilanjutkan dengan menghitung nilai *Accuracy*, *Particularity* dan tingkat error pengklasifikasian (*Classification_error*) dari model klasifikasi C4.5 dari dataset *Cervical Cancer*. Berikut hasil perhitungannya:

$$a. \text{ Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3+654}{3+654+26+44} = \frac{657}{727} = 0.9037 * 100\% = 90.37\%$$

Tingkat kedekatan antara prediksi *class* dengan aktual *class* atau jumlah prediksi *class* yang benar dari model klasifikasi C4.5 adalah sebesar **90.37%**

$$b. \text{ Classification Error} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{26+44}{3+654+26+44} = \frac{70}{727} = 0.9628 * 100\% = 96.3\%$$

Maka dilanjutkan dengan menghitung nilai *Accuracy*, *Particularity* dan tingkat error pengklasifikasian (*Classification_error*) dari model klasifikasi C4.5 Average Gain dari dataset *Cervical Cancer*. Berikut hasil perhitungannya:

Tabel 4.16 *Confusion Matrix C4.5 Average Gain(Cervical Cancer)*

Kinerja Klasifikasi	Predicted Class	
Actual Class	Predicted. Class 1	Predicted. Class 0
Actual. Class 1	0 (True Positive)	5 (False Negative)
Actual. Class 0	0 (False Positive)	77 (True Negative)

Berdasarkan tabel diatas, maka dilanjutkan dengan menghitung nilai *Accuracy*, *Particularity* dan tingkat klasifikasi yang error (*Classification error*) dari pengujian pada klasifikasi C4.5 Average Gain dari dataset *Cervical Cancer*. Berikut hasil perhitungannya:

$$a. \text{ Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+77}{0+77+0+5} = \frac{77}{82} = 0.9390 * 100\% = 93.90\%$$

Tingkat kedekatan antara prediksi *class* dengan aktual *class* atau jumlah prediksi *class* yang benar dari model klasifikasi *C4.5 average Gain* adalah sebesar **93.90%**

$$\text{b. Classification Error} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{0+5}{0+77+0+5} = \frac{5}{82} = 0.0609 * 100\% = 6.10\%$$

Tabel 4.17 *Confusion Matrix C4.5 (Kanker Rahim)*

Kinerja Klasifikasi	Predicted Class	
	Predicted. Class 1	Predicted. Class 0
Actual. Class 1	102 (True Positive)	0 (False Negative)
Actual. Class 0	27 (False Positive)	487 (True Negative)

Berdasarkan tabel diatas, maka dilanjutkan dengan menghitung nilai *Accuracy*, *Particularity* dan tingkat error pengklasifikasian (*Classification_error*) dari model klasifikasi *C4.5* dari dataset *Kanker Rahim*. Berikut hasil perhitungannya:

a. Accuracy

=

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{102+487}{102+487+0+27} = \frac{589}{616} = 0.9561 * 100\% = 95.61\%$$

Tingkat kedekatan antara prediksi *class* dengan aktual *class* atau jumlah prediksi *class* yang benar dari model klasifikasi *C4.5* adalah sebesar **95.61%**

$$\text{b. Classification Error} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{0+27}{102+487+0+27} = \frac{27}{616} = 0.0438 * 100\% = 4.38\%$$

Maka dilanjutkan dengan menghitung nilai *Accuracy*, *Particularity* dan tingkat error pengklasifikasian (*Classification_error*) dari model klasifikasi *C4.5 Average Gain* dari dataset *Kanker Rahim*. Berikut hasil perhitungannya:

Tabel 4.18 *Confusion Matrix C4.5 Average Gain (Kanker Rahim)*

Kinerja Klasifikasi	Predicted Class	
	Predicted. Class 1	Predicted. Class 0
Actual. Class 1	16 (True Positive)	0 (False Negative)
Actual. Class 0	1 (False Positive)	55 (True Negative)

Berdasarkan tabel diatas, maka dilanjutkan dengan menghitung nilai *Accuracy*, *Particularity* dan tingkat klasifikasi yang error (*Classification error*) dari pengujian

pada klasifikasi C4.5 *Average Gain* dari dataset Kanker Rahim. Berikut hasil perhitungannya:

$$a. \text{ Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{16+55}{16+55+0+1} = \frac{71}{72} = 0.9861 * 100\% = 98.61\%$$

Tingkat kedekatan antara prediksi *class* dengan aktual *class* atau jumlah prediksi *class* yang benar dari model klasifikasi C4.5 *average Gain* adalah sebesar 98.61%

$$b. \text{ Classification Error} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{0+1}{16+55+0+1} = \frac{1}{72} = 0.0140 * 100\% = 1.4\%$$

4.3. Kesimpulan Pengujian

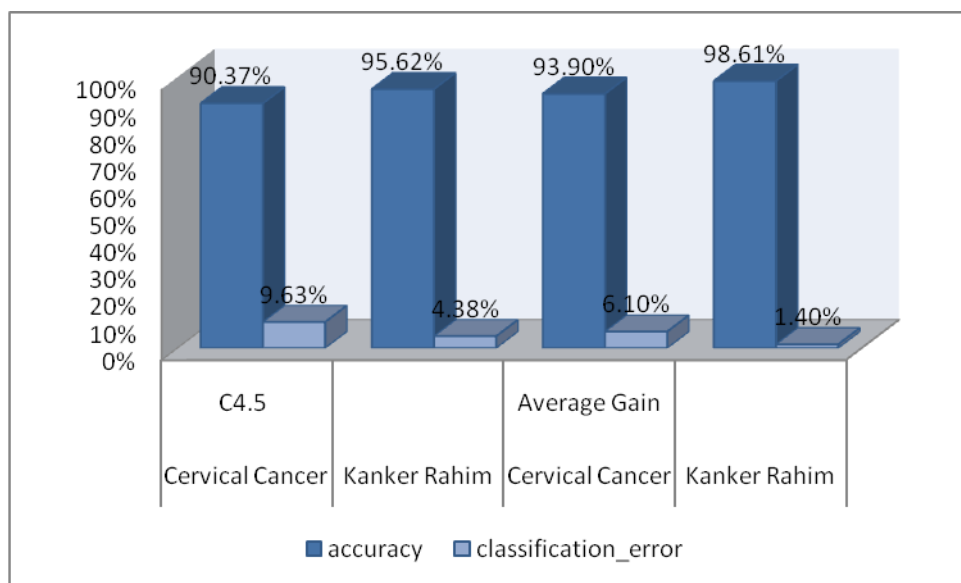
Kesimpulan pengujian yang telah dilakukan yang menggunakan dataser *Cervical Cancer* pada metode C4.5 yang memiliki tingkat akurasi sebesar 90.37% , dengan tingkat kesalahan pengklasifikasian dengan nilai 9.63%. Sedangkan klasifikasi model C4.5 *Average Gain* memiliki akurasi sebesar 93.90%, dengan tingkat kesalahan pengklasifikasian sebesar 6.10%. Pada dataser *Kanker Rahim* pada metode C4.5 yang memiliki tingkat akurasi sebesar 95.61% , dengan tingkat kesalahan pengklasifikasian dengan nilai 4.38%. Sedangkan klasifikasi model C4.5 *Average Gain* memiliki akurasi sebesar 98.61%, dengan tingkat kesalahan pengklasifikasian sebesar 1.4%. Perbedaan ini disebabkan oleh jumlah dari atribut yang berbeda, semakin banyak atribut yang diuji maka menghasilkan tingkat akurasi yang lebih rendah dari atribut yang sedikit, maka dataset Kanker Rahim memiliki akurasi yang lebih tinggi dibandingkan dengan dataset *Cervical Cancer*.

Tabel berikut menunjukkan hasil Performance Matrix yang diperoleh dari masing – masing klasifikasi decision tree C4.5 dari kedua Dataset.

Tabel 4.19 Performance Matrix Klasifikasi Decision Tree C4.5

	Cervical Cancer	Kanker Rahim	Cervical Cancer	Kanker Rahim
	C4.5		Average Gain	
Accuracy	90.37%	95.62%	93.90%	98.61%
classification_error	9.63%	4.38%	6.10%	1.40%

Berikut merupakan Grafik perbandingan antara dua dataset yang diuji:

**Gambar 4.3** Grafik Perbandingan Akurasi dan Error

BAB V

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Pada penelitian yang telah dilakukan, maka penulis menghasilkan beberapa kesimpulan sebagai berikut:

1. Kesimpulan pengujian yang telah dilakukan yang menggunakan dataser *Cervical Cancer* pada metode C4.5 yang memiliki tingkat akurasi sebesar 90.37% , dengan tingkat kesalahan pengklasifikasian dengan nilai 9.63%. Sedangkan klasifikasi model C4.5 Average Gain memiliki akurasi sebesar 93.90%, dengan tingkat kesalahan pengklasifikasian sebesar 6.10%. Pada dataser *Kanker Rahim* pada metode C4.5 yang memiliki tingkat akurasi sebesar 95.61% , dengan tingkat kesalahan pengklasifikasian dengan nilai 4.38%. Sedangkan klasifikasi model C4.5 Average Gain memiliki akurasi sebesar 98.61%, dengan tingkat kesalahan pengklasifikasian sebesar 1.4%.
2. Perbedaan pada penelitian ini disebabkan oleh jumlah dari atribut yang berbeda, semakin banyak atribut yang diuji maka menghasilkan tingkat akurasi yang lebih rendah dari atribut yang sedikit, maka dataset Kanker Rahim memiliki akurasi yang lebih tinggi dibandingkan dengan dataset *Cervical Cancer*.
3. Keberhasilan pengimplementasian dengan baik yang menggunakan metode C4.5 dalam memprediksi *Cervical Cancer dataset* dan Dataset Kanker Rahim.

5.2. Saran

Pada penelitian selanjutnya yang diharapkan penulis adalah untuk mengembangkan metode analisa dalam memprediksi data yang memiliki banyak atribut, serta tidak hanya membahas tingkat kinerja antara decision tree C4.5 dengan induksi *gain ratio* dan decision tree C4.5 konvensional saja. Karena masih ada kekurangan dalam

penelitian kedepannya dapat memperoleh hasil lebih baik dari sebelumnya. Maka dari itu penulis mengharapkan penelitian ini dilanjutkan dengan melakukan pengujian terhadap algoritma lainnya yang dapat membuat algoritma C4.5 lebih baik lagi dan memperoleh hasil akhir yang sesuai dengan keinginan. Semoga mendapatkan keakuratan yang lebih besar serta menghasilkan konsep prediksi yang lebih baik.

DAFTAR PUSTAKA

- Dai Qin-yun,. Zang Chun-Ping., Wu Hao. 2016. *Research of Decision tree Classification Algorithm in Data Mining*. Dept. of Electric and Electronic Engineering, Shijiazhuang Vocational and Technology Institute. China
- Han, J., Kamber, M. & Pei, J. 2012.*Data Mining: Concepts and Techniques*. 3rd Edition. Morgan Kaufmann Publishers: San Francisco.
- Hou, S., Hou, R., Shi, X., Wang, J., & Yuan, C. 2014. Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics. *International Journal of Control and Automation*, 7(1), 181-190.
- Hussain, H., Quazilbash. N.Z., Bai. S. &Khoja, S. 2015. Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis.*International Conference on Computer-Based Medical Systems*, pp. 131-134.
- Kavitha, K. V., Tiwari, S., Purandare, V. B., Khedkar, S., Bhosale, S. S., Unnikrishnan, A. G. (2014). Choice of wound care in diabetic foot ulcer: A practical approach. *World J Diabetes*.5(4):546–56. doi: 10.4239/wjd.v5.i4.546.
- Kavitha, R., Kannan, E. 2016. An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining. *International Conference on Emerging Trends in Engineering, Technology and Science(ICETETS)*, pp. 1-5.
- Kotu, V. & Deshpande, B. 2015. *Predictive Analytics and Data Mining*. Morgan Kaufmann Publisher: San Francisco.
- Larose, D.T. 2005.*Discovering Knowledge in Data: An Introduction to Data Mining*, John Willey & Sons. Inc. pp. 129-240
- Maimon, O. dan Last, M. 2000. Knowledge Discovery and Data Mining, The Fuzzy network (IFN) Methodology. Dordrecht: Kluwer Akademik.

- Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. *Croatian Operational Research Review (CRORR)*, 367-388.
- Quinlan, J.R. 1992. C4.5 Programs for Maching Learning. San Mateo, CA: Morgan Kaufmann.
- Raviya. Kaushik H & Gajjar, Biren. 2013. *Performance Evaluation of Different Data Mining Classification Algoritma Using WEKA*. *Indian Journal of Research*. Volume. 2. Issue.1. ISSN: 2250-1991.
- Sahu, Mridu., Nagwani. N.K., Verma Shrish., Shirke. Saransh. 2015. *Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal*. *International Journal of Knowledge Engineering*, Volume.1, No.2.
- Seema., Rathi Monika., Mamta. 2012. *Decision Tree: Data Mining Techniques*. Department of Computer Science Engineering. India.
- Sharma, R., Purushottam, Saxena, K. 2016. Efficient Heart Disease Prediction System using Decision Tree. *International Conference on Computing, Communication and Automation (ICCCA)*, Noida, India, 15-16 May. 72-77. DOI: 10.1109/CCAA.2015.7148346
- Sivapriya, T. R., Nadira, B. K. 2013. Hybrid Feature Selection for Enhanced Classification of High Dimensional Medical Data. *International Conference on Computational Intelligence and Computing Research*, pp. 1-4.
- Steinbach, M., Karypis, G. & Kumar, V. 2000. A comparison of document clustering techniques. *KDD Workshop on Text Mining*, pp. 525.
- Zhang, S. (2012). Decision tree classifiers sensitive to heterogeneous costs. *Journal of Systems and Software*, 85(4), 771–779. doi:10.1016/j.jss.2011.10.007