# PENGGUNAAN ALGORITMA C4.5 DALAM MENENTUKAN BIOPSY PADA PENDERITA KANKER PAYUDARA

Adi Candra P[1], Bakhtiyar Hadi P[2]
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
adicandra012@gmail.com, bahtiar.hp@gmail.com

**ABSTRAK**

Breast cancer is the most common cancer diagnosis in women in the world. The incidence of breast cancer has increased by more than 20% since 2008 . Early detection of breast cancer in women who have no complaints can reduce mortality. Women who come with complaints usually have enlarged lesions and spread to other organs . Decision Tree C4.5 classification algorithm is a data mining method that is often used to classify diseases, the handling of doctors to patients or more generally the classification method using data mining has been widely used in the medical world. From the results of this study, it was found that the calculation value uses confussion matrix of the test data amounts to 197 records indicate that the precision accuracy of 100%, 93.4% and 97.41% recall.

Keywords: Breast cancer , Decison tree , C4.5 confussion matrix.

Kanker payudara merupakan diagnosis kanker yang paling sering terjadi pada wanita di dunia. Angka kejadian kanker payudara meningkat lebih dari 20% sejak tahun 2008. Deteksi dini kanker payudara pada wanita-wanita yang tidak memiliki keluhan dapat menurunkan angka kematian. Wanita yang datang dengan keluhan biasanya lesi sudah membesar dan menyebar ke organ lainnya. Algoritma klasifikasi Decision Tree C4.5 merupakan metode data mining yang kerap digunakan untuk mengklasifikasikan penyakit, penanganan dokter terhadap pasien atau lebih umumnya metode klasifikasi menggunakan data mining sudah banyak digunakan dalam dunia medis. Dari hasil penelitian ini, didapatkan bahwa hasil hitung menggunakan confussion matrix terhadap data uji berjumlah 197 record menunjukkan bahwa akurasi 93,4% presisi 100% dan recall 97,41%.

Kata kunci :Kanker payudara, Decison tree, C4.5 confussion matrix.

## BAB I PENDAHULUAN

### 1.1. LATAR BELAKANG

Kanker payudara merupakan diagnosis kanker yang paling sering terjadi pada wanita di dunia. Angka kejadian kanker payudara meningkat lebih dari 20% sejak tahun 2008. Menurut data WHO, pada tahun 2012 terdapat 1,7 wanita dengan diagnosis kanker payudara. Kanker ini juga merupakan penyebab umum kematian pada wanita. Deteksi dini kanker payudara pada wanita-wanita yang tidak memiliki keluhan dapat menurunkan angka kematian. Wanita yang datang dengan keluhan biasanya lesi sudah membesar dan menyebar ke organ lainnya. Dengan deteksi dini dapat menemukan kanker dengan ukuran kecil dan masih terbatas pada payudara.

## BAB II KAJIAN PUSTAKA

### 2.1. KANKER PAYUDARA

Kanker atau keganasan adalah suatu penyakit yang ditandai dengan pertumbuhan dan penyebaran jaringan secara abnormal (Tanjung, 2015). Kanker adalah pertumbuhan sel yang tidak normal/terus menerus dan tidak terkendali, dapat merusak jaringan sekitarnya serta dapat menjalar ketempat yang jauh dari asalnya yang di sebut metastasis (Anggriyani, 2015). Kanker payudara adalah suatu penyakit dimana terjadi pertumbuhan sel, akibat adanya onkogen sel normal menjadi sel kanker pada jaringan payudara (Palu, 2014).

### 2.2. KLASIFIKASI

Klasifikasi adalah proses penemuan model (atau fungsi) yang membedakan kelas data atau konsep yang bertujuan agar dapat digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Model ditemukan berdasarkan analisis data training (objek data yang kelasnya diketahui) (Han, et al, 2006: 24). Algoritma-algoritma yang sering digunakan untuk proses klasifikasi sangat banyak, yaitu k-nearest neighbor, rough set, algoritma genetika, metode rule based, C4.5, naive bayes, analisis statistik, memory based reasoning, dan support vector machines (SVM). Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah learning (fase training), dimana algoritma klasifikasi dibuat untuk menganalisa data training lalu direpresentasikan dalam bentuk aturan klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari aturan klasifikasi (Han, et al, 2006: 286).

### 2.3. METODE C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti Structured Query Language untuk mencari record pada kategori tertentu.

$$Entropy(S) = \sum_{i=1}^{n} - pi.\log 2pi$$

$$gain(S.A) = Entropy(S) - \sum_{i}^{n} \frac{|si|}{|s|} * Entropy(Si)$$

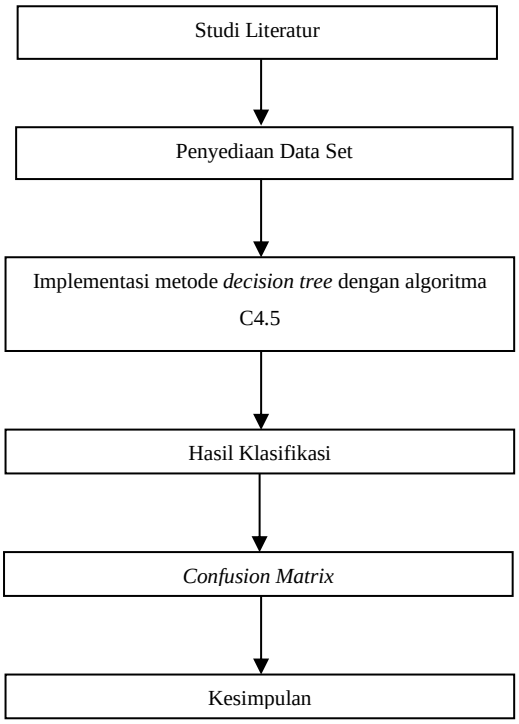
### 2.4. CONFUSSION MATRIX

Confusion matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi pada bidang Data mining. Confusion matrix ini nantinya akan melakukan perhitungan yang melakukan 4 keluaran, yaitu recall (proporsi kasus positif yang diidentifikasi dengan benar),

precision (proporsi kasus dengan hasil positif yang benar dan accuracy (perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus).

## BAB III METODOLOGI PENELITIAN

### 3.1. TAHAPAN PENELITIAN




## BAB IV IMPLEMENTASI DAN PENGUJIAN

### 4.1. IMPLEMENTASI

**1. PENGUMPULAN DATA**

Penelitian ini menggunakan dataset yang diunduh dari situs internet https://vincentarelbundock.github.io/Rdatasets/datasets.html. Dataset tersebut diambil dari hasil penelitian pada University of Wisconsin Hospitals oleh Dr. William H. Wolberg. He dengan jumlah data 761 dari tahun 1992. Pada dataset tersebut memiliki 9 atribut dalam pengukuran terhadap pasien penderita kanker payudara diantaranya; clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin,

**2. PREPROCESSING DATA**

Dari dataset yang telah disediakan selanjutnya akan melalui tahap preprocessing, ini bertujuan untuk memudahkan pembaca dan pengguna dalam mengklasifikasi output nantinya. Disini parameter yang sebelumnya berupa nominal akan diubah kedalam bentuk kategori sesuai teknik atau tata cara penghitungan dalam decision tree C4.5.

**Tabel 4.2 *Preprocessing* atribut**

| No. | Atribut awal | Atribut baru |
|---|---|---|
| 1 | clump thickness | Ketebalan Rumpun |
| 2 | uniformity of cell size | keseragaman ukuran sel. |
| 3 | uniformity of cell shape | keseragaman bentuk sel. |
| 4 | marginal adhesion | adhesi marjinal. |
| 5 | single epithelial cell size | ukuran sel epitel tunggal. |
| 6 | bare nuclei | bare nuclei |
| 7 | bland chromatin | bland chromatin |
| 8 | normal nucleoli | nukleoli normal. |
| 9 | mitoses | mitosis. |

**Tabel 4.3 *Preprocessing* parameter**

| No. | Parameter awal | Parameter baru |
|---|---|---|
| 1 | 1 | rendah |
| 2 | 2 | rendah |
| 3 | 3 | rendah |
| 4 | 4 | rendah |
| 5 | 5 | rendah |
| 6 | 6 | tinggi |
| 7 | 7 | tinggi |
| 8 | 8 | tinggi |
| 9 | 9 | tinggi |
| 10 | 10 | tinggi |

**3. IMPLEMENTASI**

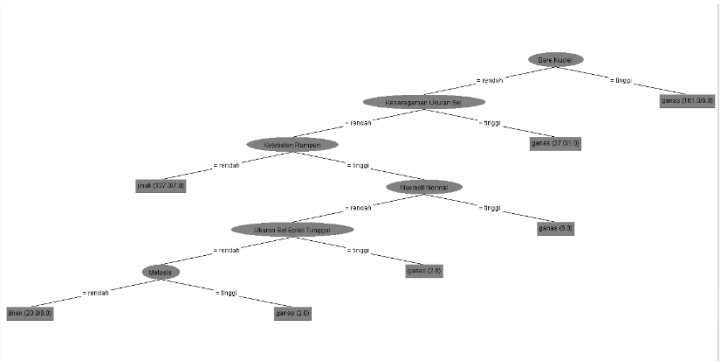
**Tabel 4.4 Hasil hitung *gain* dan entropi awal**

| atribut | jumlah kasus | jinak | ganas | *entropy* | *gain* |
|---|---|---|---|---|---|
| | 564 | 349 | 215 | 0.958889 | |
| Ketebalan Rumpun | | | | | 0.660185 |
| rendah | 394 | 333 | 61 | 0.621778 | |
| tinggi | 170 | 16 | 154 | 0.450067 | |
| keseragaman ukuran sel. | | | | | 0.423317 |
| rendah | 437 | 346 | 91 | 0.738101 | |
| tinggi | 127 | 3 | 124 | 0.161321 | |
| keseragaman bentuk sel. | | | | | 0.454707 |
| rendah | 432 | 344 | 88 | 0.729274 | |
| tinggi | 132 | 5 | 127 | 0.232481 | |
| adhesi marjinal. | | | | | 0.340577 |
| rendah | 460 | 345 | 115 | 0.811278 | |
| tinggi | 104 | 4 | 100 | 0.235193 | |
| ukuran sel epitel tunggal. | | | | | 0.320519 |
| rendah | 467 | 344 | 123 | 0.831809 | |
| tinggi | 97 | 5 | 92 | 0.29293 | |

| | | | | | |
|---|---|---|---|---|---|
| bare nuclei | | | | | 0.590683 |
| rendah | 403 | 343 | 60 | 0.607039 | |
| tinggi | 161 | 6 | 155 | 0.229618 | |
| bland chromatin | | | | | 0.403029 |
| rendah | 446 | 343 | 103 | 0.779649 | |
| tinggi | 118 | 6 | 112 | 0.289986 | |
| nukleoli normal. | | | | | 0.42281 |
| rendah | 443 | 341 | 102 | 0.778444 | |
| tinggi | 121 | 8 | 113 | 0.351258 | |
| mitosis. | | | | | 0.087154 |
| rendah | 534 | 348 | 186 | 0.932554 | |
| tinggi | 30 | 1 | 29 | 0.210842 | |

Dari penghitungan diatas didapatkan pohon keputusan seperti dibawah ini.

```
if ($n_bare_nuclei=='tinggi')
{$keputusan='ganas';}
elseif ($n_bare_nuclei == 'rendah' &&
$n_keseragaman_ukuran_sel=='tinggi')
{$keputusan='ganas';}
elseif ($n_bare_nuclei == 'rendah' &&
$n_keseragaman_ukuran_sel=='rendah' &&
$n_ketebalan_rumpun=='rendah')
{$keputusan='jinak';}
elseif ($n_bare_nuclei == 'rendah' &&
$n_keseragaman_ukuran_sel=='rendah' &&
$n_ketebalan_rumpun=='tinggi' &&
$n_nukleoli_normal=='tinggi')
{$keputusan='ganas';}
```



Gambar 4.1 Pohon keputusan yang dihasilkan

## 4.2. CONFUSSION MATRIX

Dari hasil klasifikasi diatas selanjutnya dibandingkan dengan hasil sebenarnya untuk mendapatkan kategori yang diinginkan oleh *confussion matrix*. Hasil dari penghitungan *confussion matrix* sebagai berikut.

Jumlah data yang diprediksi: 197
Jumlah data yang diprediksi tepat: 184
Jumlah data yang diprediksi tidak tepat: 13

Dari jumlah diatas dijabarkan lagi yaitu:

Data asli jinak | klasifikasi jinak = TP

Data asli jinak | klasifikasi ganas = FN

Data asli ganas | klasifikasi ganas = TN

Data asli ganas | klasifikasi jinak = FP

TP: 113 | TN: 71 | FP: 10 | FN: 3

Mengacu pada rumus pada 2.3, 2.4 dan 2.5 dalam menghitung *confussion matrix* yaitu

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad \textbf{(2.3)}$$

$$Precison = \frac{TP}{TP+FP} \quad \textbf{(2.4)}$$

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(2.5)}$$

Hail *confussion matrix* yang diperoleh sebagai berikut:

Akurasi = 93,4 %
Presisi = 100 %
*Recall* = 97,41 %

## 4.3. UJI COBA

Dari pohon keputusan yang dihasilkan selanjutnya mengaplikasikannya pada data uji. Data uji disini berjumlah 361 data.

**Tabel 4.5 Potongan data uji**

| | **Ketebalan Rumpun** | **Keseragaman Ukuran Sel** | **Keseragaman Bentuk Sel** | **Adhesi Marjinal** | **Ukuran Sel Epitel Tunggal** | **Bare Nuclei** | **Brand Chromatin** | **Nucleoli Normal** | **Metosis** | **Class *Output*** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |
| 2 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |
| 3 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |
| 4 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |
| 5 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |
| 6 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** |

Dari data uji diatas selanjutnya diujikan menggunakan pohon keputusan yang sudah diperoleh.

**Tabel 4.6 hasil klasifikasi menggunakan pohon keputusan yang diperoleh**

| No | Ketebalan Rumpun | Keseragaman Ukuran Sel | Keseragaman Bentuk Sel | Adhesi Marjinal | Ukuran Sel Epitel Tunggal | Bare Nuclei | Brand Chromatin | Nucleoli Normal | Metosis | Class *Output* | Class Prediksi | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 38 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |
| 39 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |
| 40 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |
| 41 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |
| 42 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |
| 43 | tinggi | tinggi | tinggi | rendah | rendah | tinggi | rendah | rendah | rendah | **jinak** | **ganas** | **Salah** |
| 44 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |
| 45 | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | rendah | **jinak** | **jinak** | Benar |

## BAB V PENUTUP

### 5.1. KESIMPULAN

1. Hasil klasifikasi menggunakan algoritma terhadap penderita kanker payudara untuk mengetahui kanker jinak atau kanker ganas menunjukkan bahwa algoritma yang digunakan dan sistem yang dibangun sudah cukup baik dan efisien bila dibandingkan dengan penghitungan klasifikasi manual atau menggunakan excel. Serta hasil menunjukkan pada pohon keputusan yaitu, jika Bare Nuclei rendah, Keseragaman Ukuran Sel rendah dan Ketebalan Rumpun rendah maka data terklasifikasi jinak. Jika Ketebalan Rumpun tinggi, Nucleoli Normal rendah, Ukuran Sel Epitel Tunggal rendah, dan Metosis rendah maka data terklasifikasi jinak tapi bila Metosis tinggi, data terklasifikasi ganas. Jika Ukuran Sel Epitel Tunggal tinggi maka data terklasifikasi ganas. Jika Nucleoli Normal tinggi maka data terklasifikasi ganas. Jika Keseragaman Ukuran Sel tinggi data terklasifikasi ganas. Dan jika Bare Nuclei tinggi data terklasifikasi ganas.
2. Hasil hitung menggunakan confussion matrix terhadap data uji berjumlah 197 record menunjukkan bahwa akurasi 93,4% presisi 100% dan recall 97,41%.

### 5.2. SARAN

1. Pengembang dapat menggunakan algoritma klasifikasi yang lain untuk membandingkan dalam mencari algoritma terbaik.
2. Pengembang juga dapat mengambil referensi untuk atribut terbaru sesuai perkembangan dunia kesehatan.
3. Pengembang dapat menambahkan database penyimpanan daftar pasien untuk dijadikan referensi selanjutnya.

## DAFTAR PUSTAKA


Bernita Laurensia Maria Nindia. (2017). *Klasifikasi Persalinan Normal Atau Caesar Menggunakan Algoritma C4.5.* Universitas Sanata Dharma Yogyakarta.

Reem Alyami. (2017). *Investigating the effect of Correlation based Feature Selection on breast cancer diagnosis using Artificial Neural Network and Support Vector Machines*, Computer Science and Information Technology University of Dammam Saudi Arabia.

Achmad Ramadhan Safutra. (2014). *Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Desktop*. Universitas Darwan Ali Sampit.

Baraa M. Abed. (2016). *A Hybrid Classification Algorithm Approach for Breast Cancer Diagnosis.* Computer Science and Information Technology, University of Anbar Iraq.

Shuo Liu. (2018). *Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach*. Faculty of Mathematics and Informatics, Fujian Normal University China.

Supriyadi. (2013). *Penerapan Neural Network Untuk Prediksi Penyakit Kanker Payudara*. Teknik Informatika-S2 Universitas Dian Nuswantoro Semarang.

Yuniarti. (2012). *Hubungan Antara Pengetahuan Dan Sikap Remaja Tentang Kanker Payudara Dengan Praktik Pemeriksaan Payudara Sendiri (Sadari) Pada Siswi Smk Ibu Kartini Semarang 2012.* Kesehatan Masyarakat-S1 Universitas Dian Nuswantoro .

Haris Taqwa. (2010). *Pengaruh Edukasi Tentang Kanker Payudara Terhadap Skor Kesadaran Bahaya Penyakit Kanker (Breast Cancer Awareness) Di Desa Glagah Bantul Dan Desa Kerso Jepara*. Universitas Muhammadiyah Yogyakarta, Yogyakarta.

Young Gyo Jung, Kyung Tae Kim, Byungjun Lee, Hee Yong Youn. (2016). *Enhanced Naive Bayes Classifier for Real-time Sentiment Analysis with SparkR.* Sungkyunkwan University Korea.