PENERAPAN ALGORITMA KLASIFIKASI C4.5 UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA

Laily Hermawanti
Program Studi Teknik Informatika Fakultas Teknik Universitas Sultan Fatah (UNISFAT)
Jl. Diponegoro No. 1B Jogoloyo Demak Telp (0291) 686227

Abstrak : Penyakit kanker payudara merupakan . Penelitian ini menggunakan algoritma C4.5 untuk mendiagnosis penyakit kanker payudara. Penelitian ini menghasilkan nilai akurasi untuk algoritma klasifikasi C4.5 senilai 94.56% dan nilai *Area Under Curve* (AUC) untuk algoritma *Naive Bayes* senilai 0.941, sehingga penelitian ini dalam mendiagnosis penyakit kanker payudara menghasilkan hasil yang akurat.

Kata Kunci: Penyakit kanker payudara, algoritma C4.5

PENDAHULUAN

Kanker payudara adalah kanker yang paling umum pada wanita dan penyebab utama kematian kanker di seluruh dunia (E. Technical and P. Series, 2006). Meskipun etiologi kanker payudara tidak diketahui, faktor risiko berbagai kemungkinan mempengaruhi perkembangan penyakit ini termasuk faktor genetik, hormonal, lingkungan, sosiobiologis dan fisiologis. Selama beberapa dekade terakhir, risiko kanker payudara meningkat di negara-negara industri dan berkembang sebesar 1% - 2% per tahun, tingkat kematian akibat kanker payudara menurun sedikit (E. Technical and P. Series, 2006). Maka dari itu, penyakit kanker payudara perlu didiagnosis.

Data mining dapat diaplikasikan di bidang kesehatan misalnya mendiagnosis penyakit kanker payudara, penyakit jantung, penyakit diabetes dan lain-lain (Larose, 2005).

Tujuan Penelitian

Menerapkan algoritma klasifikasi C4.5 untuk peningkatan akurasi dalam mendiagnosis penyakit kanker payudara.

KAJIAN PUSTAKA

Penyakit Kanker Payudara

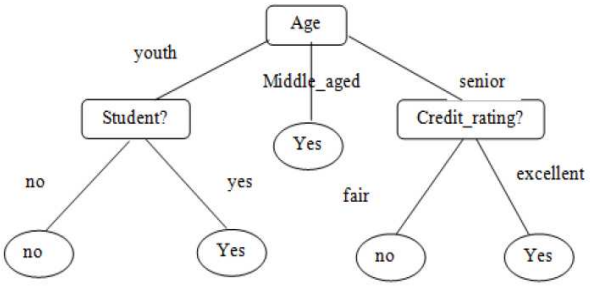
Kanker payudara adalah kanker yang paling umum pada wanita dan penyebab utama kematian kanker di seluruh dunia (E. Technical and P. Series, 2006). Meskipun etiologi kanker payudara tidak diketahui, faktor risiko berbagai kemungkinan mempengaruhi perkembangan penyakit ini termasuk faktor genetik, hormonal, lingkungan, sosiobiologis dan fisiologis. Selama

beberapa dekade terakhir, risiko kanker payudara meningkat di negara-negara industri dan berkembang sebesar 1% - 2% per tahun, tingkat kematian akibat kanker payudara menurun sedikit.

Gejala-gejala kanker payudara adalah payudara massa (*breast mass*), nyeri payudara (*breast pain*), keluarnya puting *(nipple discharge*), penarikan kembali puting atau kulit, pembengkakan lengan (*arm swelling*) dan lain-lain (E. Technical and P. Series, 2006).

Algoritma C4.5

Pohon keputusan mirip sebuah sebuah struktur pohon dimana terdapat node internal (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Gambar 1 menggambarkan pohon keputusan untuk memprediksi apakah seseorang membeli komputer. Node internal disimbolkan dengan persegi, cabang disimbolkan dengan garis, dan daun disimbolkan dengan oval. Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhan algoritma C4.5. Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (Iterative Dichotomiser), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B. Hunt, J. Marin, dan P.T. Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning*.



Gambar 1 Contoh konsep pohon keputusan untuk menentukan pembelian komputer berdasarkan atribut *age*, *student* dan *credit rating*.

Gambar 1 menggambarkan pohon keputusan untuk memprediksi apakah seseorang membeli komputer. Node internal disimbolkan dengan persegi, cabang disimbolkan dengan garis, dan daun disimbolkan dengan oval. Algoritma C4.5 dan pohon keputusan

merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhan algoritma C4.5. Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peneliti di bidang mesin pembelajaran mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (Iterative Dichotomiser), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B. Hunt, J. Marin, dan P.T. Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning*. Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 (Kusrini & Luthfi, 2009), yaitu :

1. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih,dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, hitung dahulu nilai entropy dapat dilihat pada persamaan 1 :

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

3. Kemudian hitung nilai Gain dengan metode *information gain* dapat dilihat pada persamaan 2:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad . (2)$$

   Keterangan :

   S : himpunan kasus

   A : atribut

   n : jumlah partisi atribut A

   |Si| : jumlah kasus pada partisi ke-i

   |S| : jumlah kasus dalam S
4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat :
   a. Semua tupel dalam node N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
   c. Tidak ada tupel di dalam cabang yang kosong.

Evaluasi dan Validasi pada Algoritma Klasifikasi Data Mining

Evaluasi adalah kunci untuk membuat program nyata dalam data mining (J. Han and M. Kamber, 2006). Untuk

menentukan mana yang akan digunakan pada suatu masalah tertentu, perlu cara-cara sistematis untuk mengevaluasi bagaimana metode-metode yang berbeda bekerja dan membandingkan satu dengan yang lain (J. Han and M. Kamber, 2006). Evaluasi dan validasi pada algoritma klasifikasi *data mining* adalah *Confusion Matrix* dan ROC (*Receiver Operating Characteristic) Curve.*

Confusion matrix

Confusion matrix adalah alat yang berguna untuk menganalisis bagaimana pengklasifikasi (*classifier*) dapat mengenali *tuple-tuple* pada kelas-kelas yang berbeda (J. Han and M. Kamber, 2006). Dalam kasus dengan dua klasifikasi data keluaran, seperti contoh "C_1" dan "C_2", atau contoh lainnya, tiap kelas yang diprediksi memiliki empat kemungkinan keluaran yang berbeda, yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP) dan *false negatif* (FN) menunjukkan ketepatan klasifikasi. *Confusion Matrix* dari dua kelas prediksi dapat dilihat pada tabel 1 (J. Han and M. Kamber, 2006).

Tabel 1 *Confusion Matrix* dari Dua Kelas Prediksi

		Predicted Class	
		C_1	**C_2**
Actual Class	**C_1**	*True Positive* – TP	*False Negative* – FN
	C_2	*False Positive* – FP	*True Negative* – TN

Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung *accuracy.* Rumus *accuracy* dapat dilihat pada persamaan 3 (F. Gorunesco, 2011) :

$$Accuracy = \frac{TN+TP}{TP+FP+FN+TN} \quad \text{..... (3)}$$

dimana :

TP = *true positive*

TN = *true negative*

FP = *false positive*

FN = *false negative*

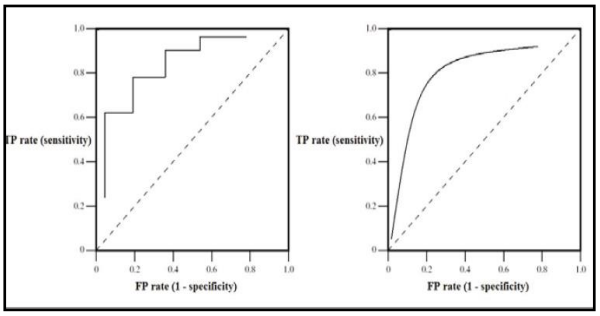
Receiver Operating Characteristic (ROC) Curve

ROC *Curve* adalah peralatan visual yang berguna untuk membandingkan dua model-model klasifikasi (J. Han and M. Kamber, 2006). Nilai ROC (*Receiver Operating Characteritics*) *Curve* sering

digunakan untuk menilai hasil dari prediksi berupa grafik. Berikut penjelasan singkat tentang ROC (F. Gorunesco, 2011) :

1. Kurva ROC pertama kali dikembangkan oleh teknik elektro dan radar dalam perang dunia II untuk mendetesi objek musuh dalam benteng pertahanan (contohnya: cerita pearl harbor yang diserang tahun 1941, atau masalah operator penerima radar di inggris)
2. ROC sudah lama digunakan dalam teori deteksi sinyal.
3. ROC sering digunakan dalam penelitian kesehatan.
4. ROC juga sering digunakan dalam penelitian *machine learning* dan *data mining* (salah satu pendahulu yang menggunakan ROC dalam *machine learning* adalah Spackman, seseorang yang mendemonstrasikan nilai dari kurva ROC dalam evaluasi dan perbandingan algoritma)
5. Dalam permasalahan klasifikasi, ROC merupakan dasar kinerja dari teknik untuk visualisasi, pengorganisasian dan pemilihan klasifikasi

Dalam permasalahan klasifikasi digunakan keputusan 2 kelas (klasifikasi biner), salah satu objek digambarkan satu elemen yang saling berpasangan yaitu positive atau negative. Model klasifikasi yang lain yaitu dengan penamaan label pada class. Kurva ROC juga dikenal sebagai grafik ROC yaitu grafik 2 dimensi yang mana TP rate adalah plot untuk Y-axis dan FP rate adalah plot untuk X-axis. Grafik ROC (*discrete* dan *continuous*) dapat dilihat pada Gambar 2 (F. Gorunesco, 2011).



Gambar 2 ROC (*Discrete* dan *Continuous*)

Pada gambar 2, ruang ROC dipisah oleh garis diagonal hasil ROC dikategorikan ke dalam *"good classification"* jika poin berada di atas garis diagonal begitu juga sebaliknya dikategaorikan *"poor classification"* jika poin berada di bawah garis diagonal. Dapat disimpulkan bahwa satu poin dalam ruang ROC

lebih baik daripada jika satu poin ke utara-selatan dari persegi (jika TP rate lebih tinggi dan FP rate lebih rendah atau kedua-duanya). Kurva ROC adalah alat dua dimensi yang digunakan untuk menilai kinerja klasifikasi. ROC sering digunakan untuk perbandingan model klasifikasi. Kategori klasifikasi untuk mencari akurasi dengan menggunakan AUC *(Area Under Curve)* dapat dilihat pada Tabel 2 (F. Gorunesco, 2011).

Tabel 2 Kategori Klasifikasi dengan menggunakan hasil AUC

Nilai AUC	**Kategori**
0.90 - 1.00	Klasifikasi sangat baik (*excellent classification*)
0.80 - 0.90	Klasifikasi baik (*good classification*)
0.70 - 0.80	Klasifikasi sama (*fair classification*)
0.60 - 0.70	Klasifikasi rendah (*poor classification*)
0.50 - 0.60	Klasifikasi gagal (*failure classification*)

METODE PENELITIAN

Metode penelitian pada penelitian ini sebagai berikut:

1. Pengumpulan Data (*Data Gathering*)

   Tahap ini dilakukan sebagai langkah awal dari suatu penelitian.Untuk memperoleh data yang benar-benar akurat, maka penentuan jenis dan sumber data sangatlah penting. *Dataset* penyakit kanker payudara yang digunakan adalah *Wisconsin Breast Cancer* (WBC) dari UCI *Dataset Repository.*
2. Pengolahan Awal Data (Data *Pre-processing*)

   Data yang didapat diolah untuk mendapatkan atribut yang relevan dan sesuai.
3. Metode Yang Diusulkan (*Proposed Model/Method*)

   Tahap ini akan membahas metode yang akan digunakan untuk penelitian ini. Dalam penelitian, setelah dilakukan studi literatur dari buku dan jurnal, ditemukan bahwa salah satu cara yang dapat membantu mengklasifikasikan diagnosis penyakit kanker payudara dari UCI Dataset Repository dengan menggunakan algoritma C4.5.
4. Eksperimen dan Pengujian Model/Metode (*Method Test and Experimen*)

   Metode eksperimen dan pengujian ini dengan algoritma C4.5.
5. Evaluasi dan Validasi Hasil (*Result Evaluation*)

Tahap ini akan membahas tentang hasil evaluasi dari eksperimen yang telah dilakukan. Pengujian hasil implementasi dengan menggunakan model *ROC CURVE (AUC)*. Hasil pengujian yang di dapat dari metode *Confusion Matrix* adalah akurasi.

EVALUASI DAN VALIDASI PADA ALGORITMA C4.5

Penelitian ini evaluasi dan validasi hasil menggunakan *confusion matrix* (accuracy) dan ROC *Curve*.

Confusion Matrix

Tabel *Confusion Matrix* algoritma C4.5 menggunakan *software* RapidMiner dapat dilihat pada Tabel 3.

accuracy: 94.56% +/- 2.99% (mikro: 94.56%)

	true benign	true malignant	class precision
pred. benign	437	17	96.26%
pred. malignant	21	224	91.43%
class recall	95.41%	92.95%	

Tabel 3. *Confusion Matrix* algoritma *Naive Bayes* menggunakan RapidMiner

Dari tabel 3, dapat dihitung nilai *accuracy* sebagai berikut :

TP = 437 FN = 17

FP = 21 TN = 224

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$= \frac{437+224}{437+224+21+17}$$

$$= 661 : 699$$

$$= 94.56\%$$

Berdasarkan hasil perhitungan, tingkat akurasi menggunakan algoritma C4.5 sebesar 94.56%.

Kurva ROC (*Receiver Operating Characteristic*)

Grafik ROC dengan nilai AUC (*Area Under Curve*) dengan algoritma C4.5 sebesar 0.941 dapat dilihat pada gambar 3. Nilai AUC termasuk kategori "klasifikasi sangat baik" karena nilainya 0.941.

Gambar 2. Nilai AUC dalam grafik ROC dengan Algoritma *Naive Bayes*

Analisa dan Validasi Model

Penggunaan algoritma C4.5 akurasinya tinggi dalam mendiagnosis kanker payudara, dapat dilihat pada tabel 4.

Tabel 4. Analisa hasil evaluasi dan validasi

	Algoritma C4.5
Accuracy	94.56%
AUC	0.941

Tabel 4 menunjukkan analisa evaluasi hasil C4.5 nilai akurasi sebesar 94.56% dan nilai AUC sebesar 0.941.

KESIMPULAN DAN SARAN

Kesimpulan

Setelah dilakukan evaluasi dengan algoritma C4.5, hasil akurasinya tinggi. Nilai akurasi untuk algoritma klasifikasi C4.5 senilai 94.56% dan nilai AUC untuk algoritma C4.5 senilai 0.941.

Saran

Untuk menambah akurasi algoritma, akan lebih baik apabila dioptimasi dengan algoritma seperti *Particle Swarm optimization* (PSO), *Genetic Algorithm* (GA) ataupun algoritma-algoritma optimasi yang lain pada algoritma C4.5 supaya didapatkan hasil akurasi yang lebih tinggi.

DAFTAR PUSTAKA

E. Technical and P. Series, *Guidelines for management of breast cancer*. World Health Organization, 2006.

Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Publishing.

D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. United States of America: John Wiley & Sons, Inc, 2005.

J. Han and M. Kamber, *Data Mining Concept dan Techniques*, 2nd ed. United States of America: Diane Cerra, 2006.

F. Gorunesco, *Data Mining Concept Model Technique*. Romania: Springer, 2011.