Penerapan Algoritma C4.5 untuk Klasifikasi Tingkat Keganasan Kanker Payudara

Dwi Ayu Nursela
A11.2010.05307
Teknik Informatika – S1, Fakultas Ilmu Komputer
Universitas Dian Nuswantoro
Jalan Nakula I No. 5-11, Semarang

Breast cancer is a disease characterized by abnormal cells grow out of control in the breast. This suggests that breast cancer is a highly malignant disease and requires the sufferer to do intensive examinations by detecting early the level of breast cancer malignancy. In this study used data mining techniques with the C4.5 algorithm. This study aims to classify the level of malignancy of breast cancer using the C4.5 algorithm. This study uses 9 attributes and attribute weighting performed to calculate the relevance of each attribute. The results of the weighting of attributes is of 9 attributes were used, only 6 attributes that affect the pattern of the decision tree and the accuracy obtained is equal to 98.57%. The results of the modeling is then performed using C4.5 algorithm and generates rules to be applied to the implementation of the classification system malignancies of breast cancer. The model is successfully applied to the system for classifying the level of malignancy of breast cancer.

Keyword : classification, C4.5 algorithm, breast cancer

I. PENDAHULUAN

Kanker payudara atau *breast cancer* adalah suatu penyakit yang ditandai dengan kelainan sel yang tumbuh tidak terkendali pada payudara [1] dan merupakan jenis kanker yang paling sering di derita pada wanita baik di negara maju maupun di negara berkembang [2]. Berdasarkan data Sistem Informasi Rumah Sakit (SIRS) tahun 2007, kanker payudara menempati urutan pertama pada pasien rawat inap di seluruh rumah sakit di Indonesia, yaitu sebesar 16,85% [3]. Menurut Departemen Kesehatan, kanker tertinggi yang diderita wanita Indonesia adalah kanker payudara dengan angka kejadian 26 per 100.000 perempuan [3]. Pada tahun 2011, WHO memperkirakan bahwa di seluruh dunia lebih dari 508.000 wanita meninggal karena kanker payudara [2].

Hal ini menunjukkan bahwa kanker payudara merupakan penyakit yang sangat ganas dan mengharuskan penderitanya untuk melakukan pemeriksaan yang intensif. Menurut [4], wanita yang positif terjangkit kanker payudara dan sudah melakukan tahap pengobatan, maka deteksi tingkat keganasan kanker payudara secara berkala sangat penting, karena digunakan untuk memilih terapi yang tepat atau menentukan tahap pengobatan selanjutnya. Tingkat keganasan kanker payudara terdiri dari dua jenis, yaitu *malignant* (ganas) dan *belign* (jinak).

Pada umumnya pendeteksian tingkat keganasan kanker payudara adalah dengan cara prognosis. Prognosis adalah "tebakan terbaik" tim medis dalam menentukan sembuh atau tidaknya pasien dari kanker payudara [4]. Selain dengan prognosis, cara lainnya adalah pemanfaatan *bioinformatic* dengan menggunakan teknik *data mining* [5] [6], karena telah terbukti dapat mendeteksi tingkat keganasan kanker payudara.

Data mining mempunyai beberapa algoritma, salah satunya yaitu algoritma C4.5 atau disebut juga algoritma *decision tree* [7]. Data mining adalah proses menemukan pola dengan memilah-milah sejumlah data yang besar menggunakan teknologi pengenalan pola [8]. Sedangkan algoritma C4.5 atau disebut juga algoritma *decision tree* merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal [7].

Berikut beberapa kelebihan dari *decision tree*, antara lain [9] :

a. Hasil analisa berupa diagram pohon yang mudah dimengerti.
b. Mudah untuk dibangun, serta membutuhkan data percobaan yang lebih sedikit dibandingkan algoritma klasifikasi lainnya.
c. Mampu mengolah data nominal dan kontinyu.
d. Model yang dihasilkan dapat dengan mudah dimengerti.
e. Menggunakan teknik statistik sehingga dapat divalidasikan.
f. Waktu komputasi relatif lebih cepat dibandingkan teknik klasifikasi lainnya.
g. Akurasi yang dihasilkan mampu menandingi teknik klasifikasi lainnya.

Berdasarkan uraian tersebut pada penelitian ini digunakan teknik *data mining* dengan menggunakan algoritma C4.5 untuk mengklasifikasi tingkat keganasan kanker payudara. Pada penelitian sebelumnya yang membahas perbandingan beberapa algoritma *data mining*, seperti algoritma *decision tree*, *back-propagated neural network* dan n*aive bayes* menunjukkan bahwa algoritma *decision tree* sudah terbukti keakuratannya dalam mendeteksi tingkat keganasan kanker payudara [5]. Pada penelitian lain oleh M. Ture, dkk [10] membandingkan algoritma *decision tree*, yaitu C&RT, CHAID, QUEST, C4.5 dan ID3 menggunakan 500 data pasien kanker payudara dimana hasil dari penelitian tersebut menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi yang lebih baik dari metode *decision tree* lainnya.

Berdasarkan latar belakang diatas, maka penelitian ini akan menerapkan algoritma C4.5 untuk mengklasifikasi tingkat keganasan kanker payudara.

II. METODE YANG DIUSULKAN

Metode dalam penelitian ini menggunakan *System Development Life Cycle Model* (SDLC) atau disebut juga dengan model *Waterfall*. Berikut langkah-langkah dari model *Waterfall* :

A. Tahap Perencanaan

Pada tahap ini perencanaan dilakukan dengan menganalisa kebutuhan sistem dan mendefinisikan kebutuhan tersebut yang akan digunakan pada tahap berikutnya.

B. Tahap Analisis Sistem

Pada tahap ini melibatkan metode lain, yaitu metode CRISP-DM. Metode CRISP-DM merupakan standar proses yang digunakan sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian dalam *data mining*. Dalam CRISP-DM, sebuah proyek *data mining* memiliki siklus hidup yang terbagi dalam enam fase, yaitu :

1. Pemahaman Bisnis (Bussiness Understanding)
2. Pemahaman Data (Data Understanding)
3. Pengolahan Data (Data Preparation)
4. Pemodelan (Modeling)
5. Evaluasi (Evaluation)
6. Penyebaran (Deployment)

C. Tahap Desain Sistem

Tahap ini merupakan proses menerjemahkan kebutuhan ke dalam representasi perangkat lunak untuk melakukan perancangan sebelum dilakukannya pengkodean. Tahap ini dibagi menjadi tiga, yaitu :

1. Merancang alur sistem dengan diagram. Pada penelitian ini menggunakan *Use Case Diagram* dan *Activity Diagram*.
2. Perancangan *database* untuk menampung semua data.
3. Perancangan *interface* untuk masukan dan keluaran sistem.

D. Tahap Implementasi

Implementasi merupakan tahap menerjemahkan perancangan sistem berupa diagram ke dalam bahasa pemrograman. Pembuatan aplikasi menggunakan bahasa pemrograman Java dan SQLyog untuk *database*.

E. Tahap Pengujian

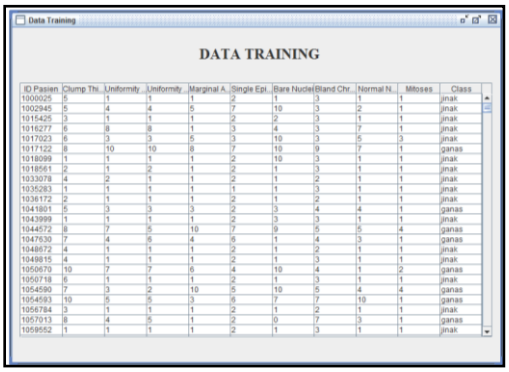
Pada tahap ini merupakan proses pengujian sistem yang akan dilakukan menggunakan metode *blackbox testing*. Pengujian ini dilakukan untuk mengetahui apakah fungsi-fungsi pada sistem sudah berjalan dengan baik.

III. IMPLEMENTASI

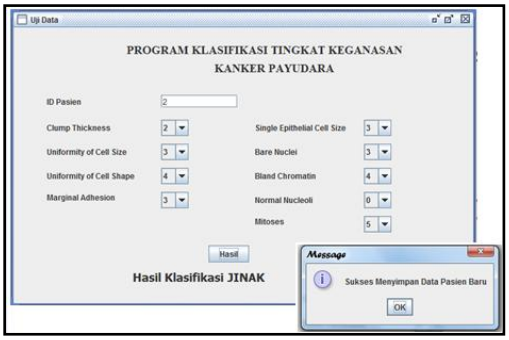
Berikut hasil implementasi sistem dari perancangan sistem yang telah dibuat :

Gambar 1. Halaman depan program

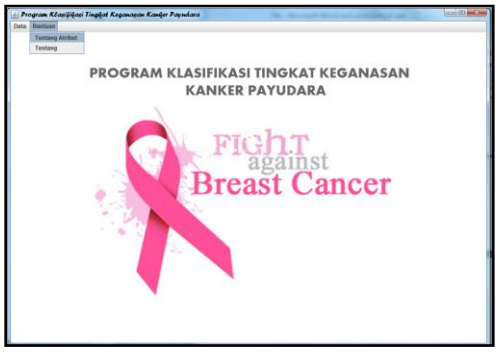
Gambar 2. Submenu pada menu data



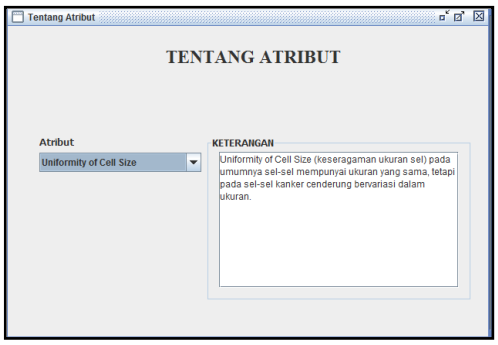
Gambar 3. Tampilan submenu data training



Gambar 4. Tampilan submenu uji data



Gambar 5. Submenu pada menu bantuan



Gambar 6. Tampilan submenu tentang atribut pada menu bantuan

IV. HASIL & PEMBAHASAN

A. *Hasil Model Pembentukan Rules*

Hasil penelitian dari pemodelan algoritma C4.5 dengan pembobotan atribut adalah pembentukan *rules.* Berikut penjelasan dari r*ules* yang terbentuk dari pemodelan algoritma C4.5 :

1. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness > 6,5 maka class = ganas
2. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei > 8,5 maka class = ganas
3. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin > 4,5 dan clump thickness > 5,5 dan uniformity of cell shape > 6,5 maka class = ganas
4. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin > 4,5 dan clump thickness > 5,5 dan uniformity of cell shape ≤ 6,5 maka class = jinak
5. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin > 4,5 dan clump thickness ≤ 5,5 maka class = ganas
6. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin ≤ 4,5 dan clump thickness > 5,5 maka class= jinak
7. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin ≤ 4,5 dan clump thickness ≤ 5,5 dan uniformity of cell size > 8,5 maka class = ganas
8. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin ≤ 4,5 dan clump thickness ≤ 5,5 dan uniformity of cell size ≤ 8,5 dan bare nuclei > 3,5 maka class = jinak
9. Jika uniformity of cell size > 2,5 dan uniformity of cell size > 4,5 dan clump thickness ≤ 6,5 dan bare nuclei ≤ 8,5 dan bland chromatin ≤ 4,5 dan clump thickness ≤ 5,5 dan uniformity of cell size ≤ 8,5 dan bare nuclei ≤ 3,5 maka class = ganas
10. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness > 6,5 maka class = ganas
11. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness ≤ 6,5 dan bland chromatin > 3,5 dan bare nuclei > 6 maka class = ganas
12. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness ≤ 6,5 dan bland chromatin > 3,5 dan bare nuclei ≤ 6 dan clump thickness > 4,5 maka class = ganas
13. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness ≤ 6,5 dan bland chromatin > 3,5 dan bare nuclei ≤ 6 dan clump thickness ≤ 4,5 maka class = jinak
14. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness ≤ 6,5 dan bland chromatin ≤ 3,5 dan uniformity of cell size > 3,5 maka class = jinak
15. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness ≤ 6,5 dan bland chromatin ≤ 3,5 dan uniformity of cell size ≤ 3,5 dan uniformity of cell shape > 2,5 maka class = ganas
16. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei > 2,5 dan clump thickness ≤ 6,5 dan bland chromatin ≤ 3,5 dan uniformity of cell size ≤ 3,5 dan uniformity of cell shape ≤ 2,5 maka class = jinak
17. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei ≤ 2,5 dan uniformity of cell size > 3,5 maka class = ganas

18. Jika uniformity of cell size > 2,5 dan uniformity of cell size ≤ 4,5 dan bare nuclei ≤ 2,5 dan uniformity of cell size ≤ 3,5 maka class = jinak
19. Jika uniformity of cell size ≤ 2,5 dan bare nuclei > 4,5 dan clump thickness > 3,5 maka class = ganas
20. Jika uniformity of cell size ≤ 2,5 dan bare nuclei > 4,5 dan clump thickness ≤ 3,5 maka class = jinak
21. Jika uniformity of cell size ≤ 2,5 dan bare nuclei ≤ 4,5 dan clump thickness > 6,5 dan uniformity of cell shape > 2,5 maka class = ganas
22. Jika uniformity of cell size ≤ 2,5 dan bare nuclei ≤ 4,5 dan clump thickness > 6,5 dan uniformity of cell shape ≤ 2,5 maka class = jinak
23. Jika uniformity of cell size ≤ 2,5 dan bare nuclei ≤ 4,5 dan clump thickness ≤ 6,5 maka class = jinak

Dapat dilihat bahwa dalam pembentukan r*ules* terbentuk 23 *rules* yang nantinya akan diimplementasikan ke dalam program.

B. Hasil Akurasi Pemodelan Algoritma C4.5

Pada pemodelan algoritma C4.5 dengan menggunakan pembobotan atribut juga menghasilkan akurasi sebesar 98,57%. Akurasi tersebut diperoleh dari kesesuaian antara prediksi klasifikasi dan hasil klasifikasi.

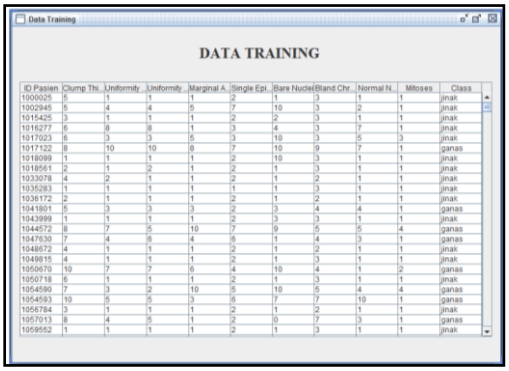
C. Hasil Implementasi Program

Gambar 7. Halaman depan program

Gambar 7 menjelaskan tentang tampilan halaman depan saat program dijalankan. Pada halaman depan terdapat menu data dan bantuan. Menu data dari beberapa submenu, yaitu data training dan uji data, seperti gambar 8 dibawah ini :

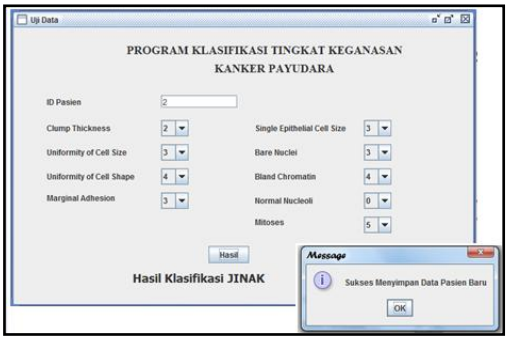
Gambar 8. Submenu pada menu data

Jika *user* memilih submenu data training, program akan menampilkan data training pada tabel, seperti gambar 9 dibawah ini :



Gambar 9. Tampilan submenu data training

Sedangkan untuk menginputkan data testing, *user* dapat memilih submenu uji data yang terdapat pada menu data dan Program akan menampilkan tampilan halaman uji data dan *user* tinggal menginputkan data testing yang akan di klasifikasi. Hasil klasifikasi akan muncul saat *user* mengklik tombol hasil dan hasil klasifikasi akan di simpan ke dalam *database*, seperti gambar 10 dibawah ini :

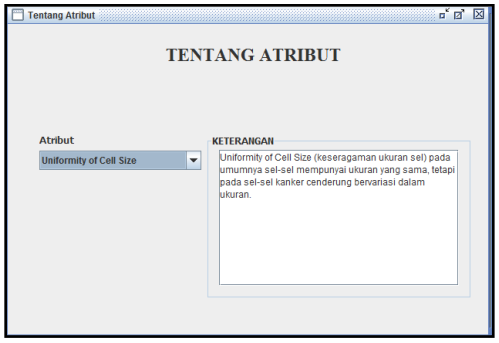


Gambar 10. Tampilan submenu uji data

Gambar 11. Submenu pada menu bantuan

Gambar 11 menjelaskan bahwa jika *user* ingin mengetahui tentang atribut yang dipakai dalam penentuan klasifikasi, *user* dapat mengklik menu bantuan dan memilih submenu tentang atribut. Pada tampilan dibawah ini, *user* dapat memilih atribut yang diinginkan dengan memilih salah satu atribut pada menu *dropdown*. Submenu tentang atribut berisi keterangan atribut-atribut yang digunakan

dalam penentuan klasifikasi. Berikut gambar 12 tampilan submenu tentang atribut :



Gambar 12. Tampilan submenu tentang atribut pada menu bantuan

V. PENUTUP

Dari permasalahan di atas dapat disimpulkan bahwa klasifikasi tingkat keganasan kanker payudara dapat diselesaikan menggunakan teknik *data mining*, yaitu algoritma C4.5, karena *rules* yang terbentuk sederhana.

Akurasi yang dihasilkan dari pemodelan algoritma C4.5 dengan pembobotan atribut, yaitu sebesar 98,57%. Akurasi tersebut diperoleh dari kesesuaian antara prediksi klasifikasi dan hasil klasifikasi.

Data yang digunakan dalam penelitian ini adalah data sampel dari *public dataset* UCI, karena data dari pasien kanker payudara merupakan data yang *private* dan sulit untuk didapatkan.

Beberapa ide yang dapat digunakan untuk penelitian selanjutnya adalah sebagai berikut :

1. Program inputan data ini akan lebih bermanfaat jika data yang digunakan adalah data valid yang bersumber dari pasien kanker payudara.
2. Menambahkan visualisasi pohon keputusan dari *rules* yang terbentuk.
3. Penelitian selanjutnya dapat dikembangkan dengan menggabungkan beberapa metode *data mining*, seperti algoritma C4.5 yang digabungkan dengan *Particle Swarm Optimization* (PSO).

REFERENCES

[1] [Online]. Available: http://www.breastcancer.org/symptoms/understand_bc. [Diakses 6 January 2014].

[2] "WHO | Breast cancer: prevention and control," [Online]. Available: http://www.who.int/cancer/detection/breastcancer/en/index1.html. [Diakses 30 December 2013].

[3] [Online]. Available: http://www.depkes.go.id/index.php?vw=2&id=1060. [Diakses 30 December 2013].

[4] F. Rachman dan S. W. Purnami, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," *JURNAL SAINS DAN SENI ITS Vol. 1, No. 1, (Sept. 2012) ISSN: 2301-928X*, pp. D-130, 2012.

[5] A. Bellaachia dan E. Guven, "Predict ing Breast Cancer Survivability Using Data Mining Techniques," dalam *SIAM Conference on Data Mining*, Washington DC, 2006.

[6] G. I. Salama, M. B. Abdelhalim dan M. A.-e. Zeid, "Experimental Comparison of Classifiers for Breast Cancer Diagnosis," dalam *International Conference Computer Engineering and Systems (ICCES)*, Cairo, 2012.

[7] Kusrini dan E. T. Luthfi, Algoritma Data Mining, Yogyakarta: ANDI, 2009.

[8] D. T. Larose, Discovering Knowledge In Data : An Introduction to Data Mining, Hoboken, New Jersey: John Wiley & Sons, Inc., 2005.

[9] F. Gorunescu, Data Mining : Concepts and Techniques, Verlag Berlin Heidelberg: Springer, 2011.

[10] M. Ture, F. Tokatli dan I. Kurt, "Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients," *Expert Systems with Applications*, pp. 2017-2026, 2009.

[11] D. Soria, J. M. Garibaldi, E. Biganzoli dan I. O. Ellis, "A Comparison of Three Different Methods for Classificat ion of Breast Cancer Data," *Seventh International Conference on Machine Learning and Applications*, pp. 619-624, 2008.

[12] Turban, Efraim dan dkk, Decision Support System and Intelligent System Edisi 7 Jilid 1, Yogyakarta: ANDI, 2005.

[13] I. H. Witten, E. Frank dan M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques, USA: Morgan Kaufmann Publishers, 2011.

[14] B. Santoso, Data Mining : Teknik Pemanfaatan Data untuk keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.

[15] M. Melanie, An Introduction to Genetic Algorithms, Cambridge, Massachusetts: A Bradford Book The MIT Press, 1996.

[16] "Confusion Matrix," [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/cconfusion_matrix/confusion_matrix.html. [Diakses 23 January 2014].

[17] A. E. Hassanien, "Classificat ion and Feature Selection of Breast Cancer Data Based on Decision Tree Algorithm," *Studies in Informatics and Control*, vol. XII No. 1, pp. 1-7, 2003.