

KOMPARASI ALGORITMA NAÏVE BAYES DAN C4.5 UNTUK MENDETEKSI KANKER PAYUDARA

TESIS

Diajukan untuk memenuhi salah satu persyaratan
Memperoleh gelar Magister Ilmu Komputer (MKOM)





**PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCA SARJANA
UNIVERSITAS BUDI LUHUR**

LEMBAR PERNYATAAN

Saya yang bertanda tangan dibawah ini:

Nama Mahasiswa : NI WAYAN PARWATI SEPTIANI
Nomor Induk Mahasiswa : 1111600555
Program Studi : REKAYASA KOMPUTASI TERAPAN
Konsentrasi : REKAYASA KOMPUTASI TERAPAN
Fakultas/Program : MAGISTER ILMU KOMPUTER

Menyatakan bahwa tesis saya yang berjudul:

KOMPARASI ALGORITMA NAIVE BAYES & C4.5
UNTUK MENDETEKSI KANKER PAYUDARA

1. merupakan hasil karya tulis ilmiah sendiri dan bukan merupakan karya yang pernah diajukan untuk memperoleh gelar akademik oleh pihak lain,
2. saya ijin untuk dikelola oleh Universitas Budi Luhur sesuai dengan norma hukum dan etika yang berlaku.

Pernyataan ini saya buat dengan penuh tanggung jawab dan saya bersedia menerima konsekuensi apapun sesuai aturan yang berlaku apabila di kemudian hari pernyataan ini tidak benar.

Jakarta, 23 SEPTEMBER 2013

Septiani



(NI WAYAN PARWATI S)



**PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCA SARJANA
UNIVERSITAS BUDI LUHUR**

PENGESAHAN

Nama Mahasiswa : Ni Wayan Parwati Septiani
Nomor Induk Mahasiswa : 1111600555
Konsentrasi : Rekayasa Komputasi Terapan
Judul : Komparasi Algoritma Naïve Bayes dan C4.5 untuk Mendeteksi Kanker Payudara

Jakarta, 23 Agustus 2013

Tim Penguji:

Ketua,

(Dr. Moedjiono, M.Sc)

Anggota,

(Samidi, M.M, M.Kom)

Pembimbing Utama,

(Dr., Ir. Nazori Az, M.T)

Tanda Tangan:

Ketua Program Studi,

(Dr., Ir. Nazori Az, M.T)

ABSTRAK

Kanker payudara merupakan penyakit yang banyak diderita oleh wanita di dunia dan merupakan penyebab kematian nomor dua yang diakibatkan oleh kanker. Kanker payudara dapat menyerang siapa saja, wanita bahkan pria. Beberapa penelitian telah dilakukan untuk mendiagnosa pasien yang terkena penyakit kanker payudara. Naïve bayes dan C4.5 merupakan algoritma yang sering digunakan untuk mendeteksi suatu penyakit. Dalam penelitian ini dilakukan komparasi metode naïve bayes dengan algoritma C4.5 dalam mendeteksi kanker payudara. Data yang digunakan dalam penelitian ini adalah *wiscosin original dataset* dari UCI *machine learning repository*. Tujuan dari penelitian ini adalah untuk mencari model dengan tingkat akurasi dan persisi yang paling tinggi dan menerapkannya dalam sistem pendeteksi kanker payudara. Dari hasil evaluasi dan validasi diketahui algoritma C4.5 memiliki tingkat akurasi dan persisi yang lebih tinggi dibandingkan dengan algoritma naïve bayes. Sehingga algoritma C4.5 dapat memprediksi kanker payudara dengan baik. Hal ini dibuktikan dengan akurasi dan persisi yang cukup tinggi yaitu 98,51% dan 100%.

Kata kunci: *Naïve bayes*, C4.5, klasifikasi, *data mining*, kanker payudara

ABSTRACT

Breast cancer is the great suffered disease in the world and becomes the second gratest death cancer. Breast cancer could attack everyone, both woman and man. Number of research has done to diagnose breast cancer patients. Naïve bayes and C4.5 is the common used to detect a disease. This research, the comparison method of naïve bayes with C4.5 in detecting breast cancer is conducted. The data used is Wisconsin original data set from UCI machine learning repository. The aim is at finding the highest accurate and percission model and applying it on breast cancer detection system. It is found that C4.5 Algorithm has the highest in accuracy and percission compared from naïve bayes algorithm. Therefore C4.5 algorithm can predict breast cancer well. It has been proved with 98,51% accuracy and 100% percission.

Keywords: Naïve Bayes, C4.5, classification, data mining, breast cancer

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Tuhan yang Maha Esa, yang telah melimpahkan rahmat dan karunia-Nya, sehingga pada akhirnya penulis dapat menyelesaikan tesis ini tepat pada waktunya. Judul tesis ini adalah "Komparasi naive bayes dan algoritma C4.5 untuk mendeteksi kanker payudara"

Tujuan penulisan tesis ini dibuat sebagai salah satu syarat untuk mendapatkan gelar Magister Ilmu Komputer (M.Kom) pada Program Pascasarjana Magister Ilmu Komputer Universitas Budi Luhur Jakarta.

Tesis ini menggunakan naive bayes dan algoritma C4.5 untuk mendeteksi kanker payudara. Data pasien yang digunakan adalah data yang terdapat pada *University of California Irvine machine learning data repository*. Penulis juga mencari dan menganalisa berbagai macam sumber referensi, baik dalam bentuk jurnal ilmiah, buku-buku literatur, internet, dll yang terkait dengan pembahasan pada tesis ini.

Penulis menyadari bahwa tanpa bimbingan dan dukungan dari semua pihak dalam pembuatan tesis ini, maka penulis tidak dapat menyelesaikan tesis ini tepat pada waktunya. Untuk itu izinkanlah penulis pada kesempatan ini untuk mengucapkan terima kasih yang sebesar-besarnya kepada:

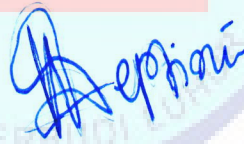
1. Bapak Dr. Ir Nazori AZ, MT selaku ketua Program Studi Magister Ilmu Komputer Universitas Budi Luhur dan pembimbing tesis yang telah menyediakan waktu, pikiran dan tenaganya dalam membimbing penulis untuk menyelesaikan tesis ini.
2. Bapak Dr. Moedjiono, M.Sc selaku Direktur Program Pascasarjana Magister Universitas Budi Luhur.
3. Orang tua tercinta serta seluruh keluarga yang telah memberikan dukungan material dan moral kepada penulis.

4. Seluruh staf pengajar (dosen) Program Pascasarjana Magister Ilmu Komputer Universitas Budi Luhur yang telah memberikan pelajaran yang berarti bagi penulis selama menempuh studi.
5. Seluruh staf dan karyawan Program Pascasarjana Magister Ilmu Komputer Universitas Budi Luhur yang telah melayani penulis dengan baik selama kuliah.

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan tesis ini. Penulis menyadari bahwa penulisan tesis ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah yang penulis hasilkan untuk yang akan datang.

Akhir kata semoga tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, Agustus 2013



Ni Wayan Parwati Septiani

Penulis

DAFTAR ISI

LEMBAR PENGESAHAN TESIS.....	i
ABSTRAK.....	ii
ABSTRACT.....	iii
KATA PENGANTAR.....	iv
DAFTAR ISI.....	vi
DAFTAR GAMBAR.....	viii
DAFTAR TABEL.....	ix
DAFTAR LAMPIRAN.....	x
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang Masalah.....	1
1.2 Masalah Penelitian.....	2
1.2.1 Identifikasi Masalah.....	2
1.2.2 Batasan Masalah.....	3
1.2.3 Rumusan Masalah.....	3
1.3 Tujuan dan Manfaat Penelitian.....	3
1.3.1 Tujuan Penelitian.....	3
1.3.2 Manfaat Penelitian.....	4
1.4 Tata Urut Penulisan.....	4
BAB II LANDASAN TEORI DAN KERANGKA KONSEP.....	5
2.1 Tinjauan Pustaka.....	5
2.1.1 Kanker Payudara.....	5
2.1.2 Data Mining.....	6
2.1.2.1 Data, informasi dan pengetahuan (<i>knowledge</i>).....	7
2.1.2.2 Pekerjaan dalam data mining.....	8
2.1.3 Klasifikasi.....	10
2.1.4 Model.....	11
2.1.5 Klasifikasi naïve bayes.....	11
2.1.6 Algoritma C4.5.....	18
2.2 Tinjauan Studi.....	27
2.3 Tinjauan Objek Penelitian.....	29
2.3.1 Objek penelitian.....	29
2.3.2 Perangkat Keras.....	29
2.3.3 Perangkat Lunak.....	29
2.4 Kerangka konsep.....	30
2.5 Hipotesis.....	31
BAB III METODOLOGI DAN RANCANGAN PENELITIAN.....	32
3.1 Metode Penelitian.....	32
3.2 Metode Pemilihan Sampel.....	32
3.2.1 Pengumpulan Data.....	32
3.2.2 Pengolaha data awal.....	32
3.2.3 Metode yang diusulkan.....	33

3.2.4 Eksperimen dan pengujian metode.....	33
3.2.5 Evaluasi dan validasi hasil.....	33
3.3 Metode Pengumpulan Data.....	33
3.4 Teknik analisis, perancangan dan pengujian data.....	34
3.5 Langkah-langkah penelitian.....	35
3.5.1 <i>Business Understanding</i>	36
3.5.2 <i>Data Understanding</i>	36
3.5.3 <i>Data Preparation</i>	36
3.5.4 <i>Modelling</i>	38
3.5.5 <i>Evaluation</i>	38
3.5.6 <i>Deployment</i>	38
3.6 Jadwal Penelitian.....	39
BAB IV PEMBAHASAN HASIL PENELITIAN.....	40
4.1 Hasil Penelitian.....	40
4.1.1 Algoritma Naïve Bayes.....	40
4.1.2 Algoritma C4.5.....	45
4.1.3 Komparasi Akurasi dan Persisi.....	53
4.2 Perancangan system.....	57
4.3 Implikasi Penelitian.....	59
4.3.1 Aspek Sistem.....	59
4.3.2 Aspek Manajerial.....	59
4.3.3 Aspek Penelitian Lanjutan.....	60
4.4 Rencana Implementasi.....	61
BAB V PENUTUP.....	63
5.1 Kesimpulan.....	63
5.2 Saran.....	63
DAFTAR PUSTAKA.....	65
LAMPIRAN-LAMPIRAN.....	67

DAFTAR GAMBAR

Gambar		Halaman
II-1	Posisi data mining dalam bisni cerdas (<i>business intelligence</i>).	7
II-2	Proses pekerjaan klasifikasi.....	11
II-3	Pohon keputusan hasil perhitungan node 1.....	22
II-4	Pohon keputusan hasil perhitungan node 1.1.....	24
II-5	Pohon keputusan hasil perhitungan node 1.1.2.....	26
II-6	Kerangka konsep penelitian.....	31
III-1	Alur metode naïve bayes.....	34
III-2	Tahapan penelitian.....	35
IV-1	Pohon Keputusan Deteksi Kanker Payudara.....	50
IV-2	Proses Pengujian Pada RapidMiner.....	54
IV-3	GUI Sistem Pendeteksi Kanker Payudara.....	58

DAFTAR TABEL

Tabel		Halaman
II-1	Data latih klasifikasi hewan.....	16
II-2	Probabilitas fitur dan kelas.....	17
II-3	Keputusan bermain tenis.....	19
II-4	Perhitungan node 1.....	21
II-5	Perhitungan node 1.1.....	23
II-6	Perhitungan node 1.1.2.....	25
II-7	State of the art penggunaan klasifikasi dalam pendeteksi kanker payudara.....	28
III-1	Informasi atribut data.....	37
III-2	Jadwal penelitian.....	39
IV-1	Probabilitas Prior.....	40
IV-2	Probabilitas Posterior.....	44
IV-3	Perhitungan Gain Node 1.....	45
IV-4	Nilai Akurasi dan Persisi.....	53
IV-5	Nilai Akurasi dan Persisi Data Latih Menggunakan Algoritma Naïve Bayes.....	55
IV-6	Nilai Akurasi dan Persisi Data Uji Menggunakan Algoritma Naïve Bayes.....	55
IV-7	Nilai Akurasi dan Perisisi Data Latih Menggunakan Algoritma C 4.5.....	56
IV-8	Nilai Akurasi dan Persisi Data Uji Menggunakan Algoritma C 4.5.....	56
IV-9	Komparasi Nilai Akurasi Data Latih dan Data Uji.....	56
IV-10	Komparasi Nilai Persisi Data Latih dan Data Uji.....	56
IV-11	Tabel Nilai Akurasi dan Persisi Data Baru Algoritma C4.5.....	57
IV-12	Spesifikasi Perangkat keras dan perangkat lunak.....	59
IV-13	Rencana Implementasi Sistem Pendeteksi Kanker Payudara.....	61

DAFTAR LAMPIRAN

Lampiran		Halaman
1	Data Pasien Penyakit Kanker Payudara.....	64
2	Source Code Sistem Pendeteksi Kanker Payudara.....	79
3	Pengujian Terhadap Deployment Data/Data Baru.....	88



DAFTAR PUSTAKA

- [Aruna 2011] Aruna S. et al. (2011). *Knowledge based analysis of various statistical tools in detecting breast cancer*.
- [Berka 2009] Berka, P., Rauch, J., & Zighed, D. A. (2009). *Data Mining and Medical Knowledge Management: Cases and Applications*. New York: Information Science Reference (an imprint of IGI Global).
- [Bhuvaneswari 2012] Bhuvaneswari, R., & Kalaiselvi, K. (2012). Naive Bayesian Classification Approach in Healthcare Applications. *International Journal of Computer Science and Telecommunications Volume 3, Issue 1, January*, 106-112.
- [Bramer 2007] Bramer, M. (2007). *Principles of Data Mining*. London: Springer.
- [Cohen 2005] Cohen, L., Manion, L., & Morrison, K. (2005). *Research Methods in Education 5th Edition*. London: The Taylor & Francis e-Library.
- [Gauda 2012] Gauda I. Salama, Abdelhalim, M. B., Zeid Abd-Elghany. (2012). *Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers*. *International Journal of Computer and Information Technology Vol.01, issue 01*.
- [Han 2006] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques Second Edition*. California: Morgan Kaufmann.
- [Kothari 2004] Kothari, C. R. (2004). *Research Methodology Methods and Techniques*. India: New Age International Limited.
- [Kumar 2011] Kumar, D. Senthil., Sathyadevi, G., & Sivanesh, S. (2011). Decision Support System for Medical Diagnosis Using Data Mining. *International Journal of Computer Science Issues Vol.8, Issue 3, No 1*, 147-153.
- [Kuarini 2009] Kusrini, & Luthfi, T. E. (2009). *Algoritma Data Mining*. Yogyakarta: Penerbit Andi.