TESIS

IMPLEMENTASI ALGORITMA C4.5 UNTUK KLASIFIKASI DATA REKAM MEDIS BERDASARKAN KODE PENYAKIT INTERNASIONAL

(Implementation of C4.5 Algorithm for Classification of Medical Records based on International Diseases Codes)

WENEFRIDA TULIT INA

PROGRAM PASCASARJANA TEKNIK ELEKTRO

KONSENTRASI TEKNIK INFORMATIKA

UNIVERSITAS HASANUDDIN

MAKASSAR

2013

IMPLEMENTASI ALGORITMA C4.5 UNTUK KLASIFIKASI DATA REKAM MEDIS BERDASARKAN KODE PENYAKIT INTERNASIONAL

Tesis

Sebagai Salah Satu Syarat untuk Mencapai Gelar Magister
Program Studi Teknik Elektro

Disusun dan diajukan oleh :

WENEFRIDA TULIT INA

Kepada :

PROGRAM PASCASARJANA
UNIVERSITAS HASANUDDIN
MAKASSAR
2013

PERNYATAAN KEASLIAN TESIS

Yang betanda-tangan di bawah ini :

Nama	: Wenefrida Tulit Ina
NIM	: P2700211431
Program Studi	: Teknik Elektro
Konsentrasi	: Teknik Informatika

Menyatakan dengan sebenarnya bahwa tesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil-alihan tulisan atau pemikiran orang lain. Adapun kutipan atau rujukan sebagai sumber informasi yang saya gunakan dari penulis lain, telah saya sebutkan namanya pada daftar pustaka tesis ini.

Apabila dikemudian hari ada terbukti bahwa tesis ini adalah hasil karya orang lain maka saya bersedia menerima sanksi apapun sesuai peraturan yang berlaku.

Makassar, Medio Juli 2013

Penulis

(Wenefrida Tulit Ina)

MOTTO

"Harus MAMPU, harus BISA, harus BERUSAHA..."

"SEMANGAT...!!!"

Persembahan untuk : (1). Yesus Tuhan Juru Selamatku. (2). Ayahanda Geroda Gerardus (alm), orang tuaku tercinta Bpk. Wilem Kopong & Mm. Maria Dai Rotok serta adik-adikku tersayang sekeluarga. (3). Mertuaku tercinta Bpk. Nikolaus Leki Kleden & Mm. Mariana Malay serta Kakak-adik ipar sekeluarga.

Special untuk :

Novrie Kleden (Suami)

Chinta Kleden (Anak)

Kasih Kleden (Anak)

Gerarld Malay Kleden (Anak)

PRAKATA

Puji syukur kehadirat Tuhan Yang Maha Kuasa yang telah melimpahkan rahmat dan berkat-Nya, sehingga penyusunan tesis ini dapat diselesaikan.

Penelitian ini berdasarkan ide dari ilmu data mining terutama pada pengimplementasian metode klasifikasi data menggunakan algoritma C4.5. Data yang dipakai sebagai objek penelitian berupa data rekam medis dari RSUD Malinau tahun 2010. Klasifikasi yang dihasilkan berdasarkan kode penyakit internasional (ICD-10)

Kelancaran dan keberhasilan penulis dalam menyelesaikan tesis ini adalah berkat bantuan dari berbagai pihak. Pada kesempatan ini penulis dengan tulus menyampaikan terima kasih kepada:

1. Prof.Dr.Ir.H.Salama Manjang, MT sebagai Ketua Komisi Penasehat sekaligus sebagai KPPS Teknik Elektro.
2. Dr.Armin Lawi, M.Eng sebagai Anggota Komisi Penasehat atas bantuan dan bimbingan yang telah diberikan mulai dari pembimbingan proposal penelitian, pelaksanaan penelitian sampai dengan penulisan tesis ini.
3. Muh.Niswar.,ST,MIT.PhD; Dr. Adnan.,ST,MT; Dr. Ir. Zulfajri B. Hasanuddin., M.Eng sebagai anggota penguji yang telah banyak memberikan kritik dan saran demi kesempurnaan tesis ini.
4. Suami dan anak-anakku tercinta, orang tua dan adik-adikku serta mertua dan ipar-iparku yang senantiasa mendoakan ku hingga terselesainya tesis ini.
5. Teman-teman seperjuangan PascaMelek'11 dan rekan-rekan angkatan 2010 (Suwoyo, Munawir, Abdur Rahman).
6. Keluarga besar Rumah Perintis yang selalu kompak dan saling mendukung selama meraih pendidikan di Makassar.

7. Semua pihak yang tidak sempat saya sebutkan namanya satu per satu disini, yang telah membantu dan mendampingi penulis selama pendidikan sampai pada ujian akhir magister.

Penulis merasa penelitian ini masih jauh dari kesempurnaan, oleh karena itu kritik dan saran sangat penulis harapkan guna menjadi bahan refleksi dan perbaikan pada penelitian-penellitian selanjutnya.

Makassar, Juli 2013

Wenefrida Tulit Ina

ABSTRAK

Wenefrida Tulit Ina. Implementasi Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan Kode Penyakit Internasional (dibimbing oleh **H.Salama Manjang** dan **Armin Lawi**)

Penelitian ini bertujuan untuk mengetahui model klasifikasi penyakit berdasarkan tumpukan data rekam medis dan untuk mengetahui kemampuan algoritma C4.5 dalam mengklasifikasikan data berkelas banyak.

Metode yang digunakan dalam penelitian ini adalah metode dalam data mining yaitu algoritma C4.5. Untuk mencapai tujuan penelitian yang dimaksud maka dipilih 4 atribut sesuai data rekam medis yang terdiri dari 1 atribut kelas / atribut tujuan yakni Diagnosa penyakit berdasarkan *International Classification of Diseases -10th* (ICD-10) dan 3 atribut kasus yakni Jenis Kelamin, Umur pasien, Bulan masuk pasien ke rumah sakit. Sebelum data-data ini diproses sesuai algoritma C4.5, data pada atribut Umur dan atribut Diagnosa dikelompokkan terlebih dahulu agar model yang dihasilkan lebih jelas dan lebih sederhana.

Hasil penelitian menunjukkan bahwa ada 5 jenis klasifikasi penyakit yaitu A00-B99, I00-I99, J00-J99, O00-O99 dan Z00-Z99. Penyakit A00-B99 umumnya diderita oleh laki-laki dengan kategori umur muda dan dewasa, perempuan dengan kategori umur tua, kategori bayi dan anak hanya terjadi pada bulan Januari, Maret, April, Mei. Penyakit I00-I99 umumnya diderita oleh laki-laki dengan kategori umur tua. Penyakit J00-J99 umumnya diderita oleh laki-laki dengan kategori umur bayi dan anak pada bulan Nopember. Penyakit O00-O99 umumnya diderita oleh perempuan dengan kategori umur muda dan dewasa. Penyakit Z00-Z99 umumnya diderita oleh bayi dan anak pada bulan Pebruari, Juni, Juli, Agustus, September, Oktober, Desember, sedangkan pada bulan Nopember diderita oleh bayi dan anak dengan jenis kelamin perempuan. Tingkat akurasi dalam klasifikasi data rekam medis menggunakan algoritma C4.5 sebesar 57,5 %, tapi akan berubah sesuai jumlah data training dan jumlah data pengujian sehingga algoritma C4.5 kurang mampu menghasilkan klasifikasi dengan jumlah kelas tujuan yang banyak.

Kata kunci : Algoritma C4.5, Rekam Medis, ICD-10

ABSTRACT

Wenefrida Tulit Ina. C4.5 Algorithm Implementation for Medical Record Data Classification is based on International Code of Diseases (led by **H.Salama Manjang** and **Armin Lawi**)

Medical records generated by a hospital every day stored away without further utilization. Pile data can be processed to produce information and knowledge that is useful for hospitals and improving health services.

This study aims to determine the classification of disease models based on the data stack of medical records, using one of the methods in the data mining algorithm C4.5. To achieve the research goal is then selected four attributes appropriate medical records consisting of 1 class attribute / attributes of the destination Diagnosis of disease based on the International Classification of Diseases-10th (ICD-10) and 3 cases the attribute Gender, patient age, Month patient admission to the hospital. Before the data is processed in accordance C4.5 algorithms, the data on the attributes Age and Diagnosis attributes grouped in advance so that the resulting model is simpler and clearer.

The results show that there are 5 types of disease classification are A00-B99, I00-I99, J00-J99, O00-O99 and Z00-Z99. A00-B99 disease generally affects men with young and adult age categories, women with older age category, the category of infants and children only occurred in January, March, April May. I0-I99 disease generally affects men with older age category. J00-J99 Diseases commonly suffered by men with age categories of infants and children in November. O00-O99 illnesses commonly suffered by women with young and adult age categories. Z00-Z99 disease usually affects infants and children in February, June, July, August, September, October, December, whereas in November suffered by infants and children with the female gender. Level of accuracy in the classification of medical record data using the C4.5 algorithm by 57.5%, but will change according to the amount of training data and amount of test data that the algorithm C4.5 classification are less able to produce the number of classes that a lot of goals.

Keywords: C4.5 Algorithm, Medical Records, ICD-10.

DAFTAR ISI

HALAMAN JUDUL i

LEMBAR PENGESAHAN ii

PERNYATAAN KEASLIAN TESIS..... iii

MOTTO..... iv

PRAKATA..... v

ABSTRAK..... vi

DAFTAR ISI viii

DAFTAR TABEL..... ix

DAFTAR GAMBAR..... x

BAB I. PENDAHULUAN

A. Latar Belakang 1

B. Rumusan Masalah 4

C. Tujuan Penelitian 5

D. Manfaat Penelitian..... 5

E. Batasan Masalah..... 5

BAB II. TINJAUAN PUSTAKA

A. Algoritma Klasifikasi 6

B. Rekam Medis 14

C. Kode Penyakit Internasional..... 15

D. Penelitian Sejenis 16

E. Kerangka Pemikiran..... 20

BAB III. METODODOLOGI PENELITIAN

A. Waktu dan Lokasi Penelitian……..……. 22

B. Metode Penelitian……..…….…….. 22

C. Alat dan Bahan…….….. 22

D. Desain Penelitian…….……. 23

E. Desain Proses…….... 25

F. Desain Aplikasi……. 27

BAB IV. HASIL PENELITIAN DAN PEMBAHASAN……..32

A. Gambaran Umum Sistem…..……..……. 32

B. Hasil Analisis dan Pembahasan…….…….. 33

C. Implementasi Sistem…… 61

BAB V. KESIMPULAN DAN SARAN…… 66

A. Kesimpulan…..……..……. 66

B. Saran…….…….. 66

DAFTAR PUSTAKA…… 67

LAMPIRAN 1. Kode Penyakit Berdasarkan ICD-10

LAMPIRAN 2. Data Sampel

LAMPIRAN 3. Hasil Perhitungan Nilai Entropy dan Gain Node 1 untuk data sampel.

LAMPIRAN 4. Pohon Keputusan hasil perhitungan Microsoft Excel.

LAMPIRAN 5. Source code

LAMPIRAN 6. Data Rekam Medis RSUD Malinau Tahun 2010

LAMPIRAN 7. Surat Keterangan Penelitian

DAFTAR TABEL

Tabel 2.1 : Kode Penyakit berdasarkan ICD-10

Tabel 4.1 : Data Sampel

Tabel 4.2 : Hasil Perhitungan Entropi dan Gain Node 1

Tabel 4.3 : Hasil Perhitungan Entropi dan Gain Node 1.1

Tabel 4.4 : Hasil Perhitungan Entropi dan Gain Node 1.1.1

Tabel 4.5 : Hasil Perhitungan Entropi dan Gain Node 1.2

Tabel 4.6 : Hasil Perhitungan Entropi dan Gain Node 1.3

Tabel 4.7 : Hasil Perhitungan Entropi dan Gain Node 1.4

Tabel 4.8 : Hasil Perhitungan Entropi dan Gain Node 1.4.1

Tabel 4.9 : Hasil Perhitungan Entropi dan Gain Node 1.5.

Tabel 4.10 : Hasil Perhitungan Entropi dan Gain Node 1.6

Tabel 4.11 : Hasil Perhitungan Entropi dan Gain Node 1.7

Tabel 4.12 : Hasil Perhitungan Entropi dan Gain Node 1.7.1

Tabel 4.13 : Hasil Perhitungan Entropi dan Gain Node 1.8

Tabel 4.14 : Hasil Perhitungan Entropi dan Gain Node 1.9

Tabel 4.15 : Hasil Perhitungan Entropi dan Gain Node 1.9.1

Tabel 4.16 : Hasil Perhitungan Entropi dan Gain Node 1.10

Tabel 4.17 : Hasil Perhitungan Entropi dan Gain Node 1.10.1

Tabel 4.18 : Hasil Perhitungan Entropi dan Gain Node 1.11

Tabel 4.19 : Hasil Perhitungan Entropi dan Gain Node 1.12

Tabel 4.21 : Contoh Data Rekam Medis dari RSUD Malinau

Tabel 4.22 : Contoh Data Rekam Medis Hasil Pengelompokkan

Tabel 4.24 : Jumlah Kasus tiap Diagnosa

Tabel 4.25 : Jumlah Kasus sesuai Data Sampel

DAFTAR GAMBAR

Gambar 2.2 : Arsitektur Data Mining

Gambar 2.3 : Kerangka Pikir

Gambar 3.1 : Flowchart Algoritma C4.5

Gambar 3.2 : Diagram Usecase

Gambar 3.3 : Diagram Aktifitas

Gambar 4.1 : Gambaran Umum Sistem

Gambar 4.2 : Hasil Pohon Keputusan dari perhitungan manual

Gambar 4.3 : Hasil Pohon Keputusan dari WEKA

Gambar 4.4 : User interface Penginputan data pasien

Gambar 4.5 : Hasil perhitungan Node 1

Gambar 4.6 : Hasil perhitungan Node 1.1

Gambar 4.7 : Hasil perhitungan Node 1.1.1

BAB I

PENDAHULUAN

A. Latar Belakang

Rumah Sakit menghasilkan banyak data rekam medis pasien setiap harinya dan terus terakumulasi seiring dengan berjalannya aktifitas di rumah sakit. Kumpulan data rekam medis pada umumnya dibiarkan begitu saja tanpa diberdayagunakan, terkecuali pasien dengan penyakit langkah sehingga rekam medisnya dipakai untuk penelitian dan rekm medis juga digunakan sebagai bukti hukum bila ada kasus hukum terhadap pasien.

Pemanfaatan secara umum hanya pada penggunaan data-data tersebut dalam memberikan grafik secara statistic tentang jumlah pasien yang berobat dengan penyakit yang di deritanya beserta laporan kepulangan pasien, juga untuk laporan biaya rumah sakit. Laporan data inilah yang saat ini digunakan oleh rumah sakit dan juga Dinas Kesehatan untuk menghasilkan kebijakan-kebijakan yang berhubungan dengan penyuluhan kesehatan kepada masyarakat.

Mengenai pola kecenderungan penyakit yang diderita oleh masyarakat yang terpetakan secara rinci terhadap kelompok umur tiap bulan dalam setahun belum digali secara maksimal untuk dijadikan acuan yang lebih khusus dalam upaya merancang program-program penyuluhan

kesehatan kepada masyarakat yang mana pada era sekarang semakin intensif dilaksanakan sehingga bisa tepat sasaran.

Hadirnya ilmu Data Mining dengan berbagai metodenya menjadi sarana baru untuk menggali dan mengekstrak tumpukan data berskala besar untuk menghasilkan informasi dan pengetahuan yang berdayaguna, sehingga pada saat ini fungsi Data Mining banyak diterapkan dalam berbagai penelitian di segala bidang kehidupan yang memiliki data banyak dan akan terus bertambah setiap waktunya. (Shu, Pei, 2012)

Fungsi/peran Data Mining terdiri dari estimasi, prediksi, klasterisasi, klasifikasi, dan aturan asosiasi. Fungsi-fungsi ini bisa diterapkan secara tunggal dan juga bisa secara jamak untuk pencapaian solusi sesuai dengan karakteristik data dan tujuan yang ingin dicapai dalam suatu penelitian. (Sivanandam, 2006)

Metode klasifikasi digunakan untuk mengelompokan data ke dalam tingkatan tertentu berdasarkan atribut-atributnya. Pengelompokan yang dihasilkan menjadi model yang pada umumnya digunakan untuk memprediksi kelompok data baru. Algoritma yang dipakai yaitu Pohon Keputusan *(Decision Tree)*, C4.5, CART, k-NN *(k-Nearest Neighbor)*, JST (Jaringan Syaraf Tiruan), *Naïve Bayes,* SVM *(Support Vector Mechine)*. Aalgoritma C4.5 menempati peringkat 1 dalam urutan 10 algoritma terbaik dalam data mining (ICDM’06 Panel). Algoritma C4.5 lebih unggul dalam tingkat akurasi prediksi serta waktu komputasinya lebih cepat dibandingkan dengan algoritma k-NN (Kursini, 2009). Sehingga dalam penelitian ini, memakai algoritma C4.5 untuk menghasilkan model

klasifikasi data rekam medis berdasarkan Kode ICD-10 (*International Classification Of Diseases – 10th*) yang diambil dari data rekam medis RSUD Malinau – Kab. Malinau selama tahun 2010.

Kabupaten Malinau masuk kategori wilayah KaE dengan nilai IPKM (Indeks Pembangunan Kesehatan Masyarakat) 0,51 (KemenKes 2010). RSUD Malinau merupakan salah satu rumah sakit yang berada di perbatasan Indonesia-Malaysia dengan wilayah pelayanannya mencakup masyarakat Kabupaten Malinau dan masyarakat dari kabupaten sekitar seperti Kabupaten Tanah Tidung dan Kabupaten Nunukan. Kabupaten Malinau juga merupakan salah satu kabupaten dalam wilayah termuda dalam wilayah Republik Indonesia yaitu propinsi Kalimantan Utara yang baru disahkan tahun 2012. Tentunya pemerintah sangat memfokuskan pengembangan pembangunan di segala bidang pada wilayah ini termasuk bidang kesehatan dan untuk menunjang pembangunan perlu adanya penelitian-penelitian di wilayah ini, agar hasil penelitian-penelitian tersebut dapat bermanfaat untuk pengembangan pembangunannya.

Dengan demikian diharapkan hasil penelitian ini bisa memberikan informasi dan pengetahuan baru yang bermanfaat bagi RSUD Malinau dalam meningkatkan pelayanan kesehatan kepada masyarakat di Kabupaten Malinau dan sekitarnya.

B. Rumusan Masalah

Dari uraian latar belakang tersebut, maka dirumuskan permasalahan untuk diteliti sebagai berikut :

1. Bagaimana pola klasifikasi penyakit berdasarkan kode penyakit internasional sesuai jenis kelamin, kelompok umur, dan bulan masuk pasien yang diperoleh dari data rekam medis pasien di rumah sakit menggunakan algoritma C4.5 ?
2. Bagaimana performansi algoritma C4.5 untuk klasifikasi data dengan jumlah kelas klasifikasi yang banyak.

C. Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Menghasilkan pola klasifikasi penyakit berdasarkan kode penyakit internasional sesuai jenis kelamin, kelompok umur, dan bulan masuk pasien yang diperoleh dari data rekam medis pasien di rumah sakit menggunakan algoritma C4.5 ?
2. Mengetahui performansi algoritma C4.5 untuk klasifikasi data dengan jumlah kelas klasifikasi yang banyak.

3. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan beberapa manfaat sebagai berikut :

1. Bagi peneliti, yakni : meningkatkan pengetahuan dan pengembangan di bidang informatika dan aplikasinya pada kebutuhan masyarakat.
2. Bagi RSUD Malinau, yakni : sebagai salah satu sumber informasi baru dari tumpukan data rekam medis dalam upaya peningkatan pelayanan kesehatan kepada masyarakat.

4. Ruang Lingkup / Batasan Penelitian

Ruang lingkup materi yang dibahas dalam penelitian ini adalah:

1. Algoritma yang digunakan adalah algoritma C4.5.
2. Data yang dipakai berupa data rekam medis pasien di RSUD Malinau Kab. Malinau - Propinsi Kalimantan Timur tahun 2010.
3. Untuk tujuan kode etik, identitas/nama pasien dihilangkan.
4. Informasi penyakit yang dihasilkan dalam bentuk pola klasifikasi penyakit berdasarkan jenis kelamin, kelompok umur serta bulan masuknya pasien di rumah sakit.
5. Klasifikasi penyakit berdasarkan kode ICD-10.
6. Menggunakan WEKA 3.7.9 *(tools Data Mining)* sebagai pembanding.
7. Menggunakan bahasa pemrograman Visual Basic untuk membangun sistem aplikasi.
8. Menggunakan pengujian silang *(K-Fold Cross validation)* untuk mengetahui tingkat akurasi system yang dibuat.

BAB II

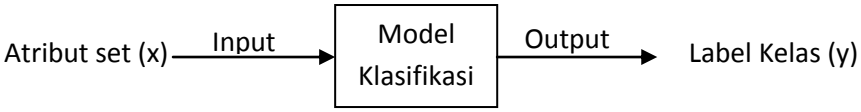
TINJAUAN PUSTAKA

A. Algoritma Klasifikasi

Klasifikasi diartikan sebagai tugas pembelajaran sebuah fungsi target f yang memetakan setiap himpunan atribut x ke salah satu label kelas y yang telah didefinisikan sebelumnya. Data input untuk klasifikasi adalah koleksi dari *Record*. Setiap *Record* dikenal sebagai *Instance* atau contoh, yang ditentukan oleh sebuah tuple (x,y) dimana x adalah himpunan atribut dan y adalah atribut tertentu yang dinyatakan sebagai label kelas (juga dikenal sebagai kategori atau atribut target). Fungsi target juga dikenal secara informal sebagai model klasifikasi.

Model klasifikasi berguna untuk keperluan sebagai berikut :

a. Pemodelan Deskriptif. Model klasifikasi dapat bertindak sebagai alat penjelas untuk membedakan objek-objek dari kela-kelas yang berbeda.
b. Pemodelan Prediktif. Model klasifikasi juga dapat digunakan untuk memprediksi label kelas dari *Record* yang tidak diketahui. Seperti pada gambar 2.1 tampak sebuah model klasifikasi dapat dipandang sebagai kotak hitan yang secara otomatis memberikan sebuah label ketika dipresentasikan dengan himpunan atribut dari *Record* yang tidak diketahui.



Gambar 2.1. Pemetaan sebuah himpunan atribut (x) ke label kelas (y)

Dikenal ada beberapa teknik klasifikasi, diantaranya : *Decision Tree Classifier, Rule-Based Classifier, Neural Network, Support Vector Machine, dan naïve Bayes Classifier.* Setiap teknik klasifikasi menggunakan algoritma pembelajaran untuk mengidentifikasi model yang memberikan hubungan yang paling sesuai antara himpunan atribut dan label kelas dari data input.

Pendekatan umum yang digunakan dalam masalah klasifikasi adalah data training yang berisi *Record* yang mempunyai label kelas yang harus tersedia. Data training digunakan untuk membangun model klasifikasi yang kemudian diaplikasikan ke data test yang berisi *Record-Record* dengan label kelas yang tidak diketahui.

1. *Decision Tree*

Decision Tree merupakan salah satu fungsional dari data mining yang menggunakan representasi tree untuk menentukan aturan-aturan klasifikasi. Ada 2 tipe *Decision Tree* yaitu *Classification Tree* dan *Regression tree*.

Classification Tree memberi label dan memasukkan *Record-Record* ke ddalam kelas-kelas yang telah disediakan, sedangkan Regression tree membuat estimasi nilai dari sebuah variabel target

yang berdasar pada nilai numeric.Yang digunakan pada penelitian ini adalah *Classification Tree*.

Decision Tree terdiri dari node internal yang menggambarkan data yang diuji, cabang menggambarkan nilai keluaran dari data yang diuji, sedangkan leaf node menggambarkan distribusi kelas dari data yang digunakan. *Decision Tree* digunakan untuk mengklasifikasi suatu sampel data yang tidak dikenal.

Pembentukan *Decision Tree* terdiri dari 3 tahap, yaitu :

a. Pembentukan Pohon

   Pada tahap ini akan dibentuk suatu pohon yang terdiri dari akar yang merupakan node paling awal, daun sebagai distribusi kelas, dan batang yang menggambarkan hasil keluaran dari pengujian. Pada pembentukan pohon ini dilakukan pemilihan atribut untuk penentuan posisi-posisi dalam pohon.

b. Pemangkasan Pohon

   Pemangkasan pohon bertujuan untuk mengidentifikasi dan membuang cabang yang tidak diperlukan pada pohon yang telah terbentuk. Ada dua metode dalam melakukan pemangkasan pohon yaitu : *Prepruning*, dan *Post-pruning*. *Prepruning* adalah pemangkasan yang dilakukan sejak awal pembentukan pohon sedangkan *post-pruning* adalah pemangkasan yang dilakukan saat pohon telah terbentuk.

Namun kedua metode pemangkasan dapat dilakukan secara kombinasi untuk menghasilkan pohon yang lebih baik.

c. Pembentukan Aturan Keputusan

Aturan yang dihasilkan dari *Decision Tree* dapat ditampilkan dalam bentuk aturan *IF-THEN*. Aturan diibentuk dari tiap bagian pohon. Setiap node yang bukan leaf node berperan sebagai bagian IF sedangkan bagian *THEN* diambil dari leaf node yang merupakan konsekuen dari aturan.

Aturan *IF-THEN* lebih mudah dipahami oleh pengguna apalagi jika pohonnya dalam ukuran besar.

Decision Tree memiliki beberapa cara dalam menentukan ukuran data dalam membentuk pohon. Pada algoritma C4.5 dipakai information *Gain* sebagai penentuan akar pohon.

2. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Algoritma ini 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 dapat diekspresikan dalam bentuk bahasa basis data seperti SQL (*Structured Query Language*) yang berguna untuk mencari *Record* pada kategori tertentu. Pohon keputusan berguna untuk mengeksplorasi data, menemukan

hubungan tersembunyi antara sejumlah calon variabel input dengan variabel target.

Data dalam pohon keputusan biasanya dinyatakan dalam bentuk table dengan atribut dan *Record*. Atribut menyatakan suatu parameter yang dibuat sebagai criteria dalam pembentukan pohon. Salah satu atribut merupakan atribut yang menyatakan data solusi per item data yang disebut target atribut atau atribut kelas tujuan. Atribut memiliki nilai-nilai yang dinamakan *Instance*.

Ada 2 variabel yang dipakai dalam menentukan akar dari pohon keputusan yaitu nilai *Entropy* dan nilai *Gain*.

Nilai *Entropy* diperoleh dari rumus :

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

Keterangan :
S = Himpunan Kasus
n = Jumlah partisi S
p = Proporsi dari S_i terhadap S

Nilai *Gain* diperoleh dari rumus :

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

Keterangan :
S = Himpunan Kasus
A = Atribut
n = Jumlah partisi atribut A
$|S_i|$ = Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

Atribut dengan nilai *Gain* tertinggi akan dipilih menjadi akar dari pohon keputusan.

Secara umum dalam membangun pohon keputusan dengan algoritma C4.5 akan melalui proses sebagai barikut :

a. Pilih atribut dengan *Gain* tertinggi sebagai akar pohon
b. Buat cabang untuk tiap-tiap nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

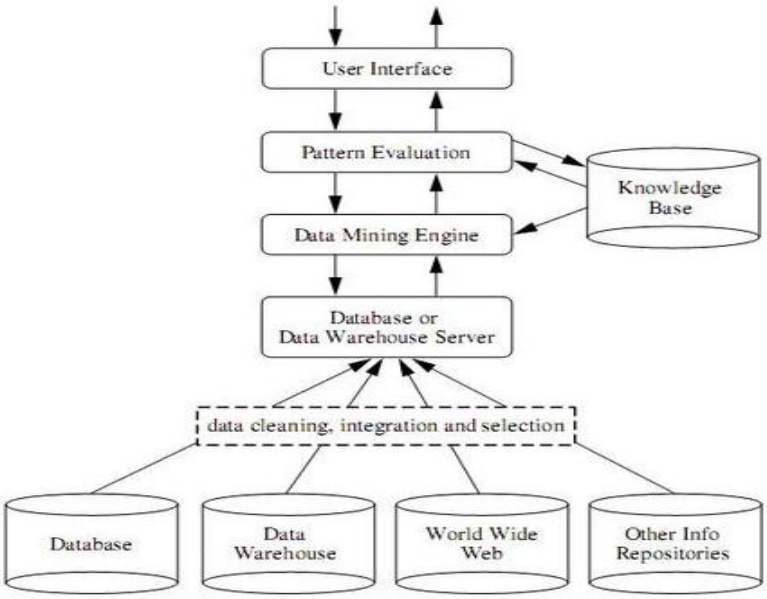
3. Data.

Data warehouse merupakan salah satu cara untuk mengekstak informasi penting dari data yang beranekaragam sehingga dapat dianalisa menggunakan teknik dan metode tertentu. Data warehouse bersifat orientasi subjek, terintegrasi, waktunya bervariasi dan hanya bisa dibaca (bersifat *read only*) tidak bisa diubah tetapi bisa ditambah dan digunakan untuk mendukung proses pengambilan keputusan. (Sivanandam,2007).

Data warehouse berisikan data-data historikal yang terintegrasi dan mudah diakses. Data-data ini direpresentasikan dalam bentuk multi-dimensi (*data cube)*. Berbeda dengan database operasional yang dapat melakukan *update*, *insert* dan *delete* terhadap data yang mengubah isi dari database sedangkan pada data warehouse hanya ada dua kegiatan memanipulasi data yaitu pengambilan data (*data loading*) dan akses data (mengakses data warehouse seperti

melakukan query atau menampilan laporan yang dibutuhkan, tidak ada kegiatan *updating* data).

Dalam proses KDD (*Knowledge Discovery in Database*), sebelum melakukan data mining terlebih dahulu dilakukan proses seleksi data (*data selection*), pembersihan data (*Data Cleaning*), transformasi data (*Data transformation*). Proses-proses inilah yang akan menghasilkan data warehouse yang siap dipakai sebagai input untuk proses selanjutnya dalam KDD yaitu proses data mining. Arsitektur Data Warehouse seperti pada gambar 2.2.



Gambar 2.2. Arsitektur Data Mining

Proses yang dilakukan pada data warehouse adalah sebagai berikut :

1. Ekstraksi Data (*Extract*)

   Ekstraksi data adalah proses dimana data diekstrak dari berbagai sumber data.

2. Transformasi Data (*Transformation*)

   Trannsformasi data adalah proses dimana data mentah hasil ekstraksi disaring dan diubah sesuai dengan kaidah bisnis yang berlaku. Langkah-langkah dalam transformasi data adalah sebagai berikut :

   a. Memetakan data input dari skema data asli ke skema data warehouse.
   b. Melakukan konversi tipe data atau format data.
   c. Pembersihan serta pembuangan duplikasi dan kesalahan data.
   d. Perhitungan nilai-nilai derivatif (mula-mula)
   e. Perhitungan nilai-nilai agregat (rangkuman)
   f. Pemeriksaan integritas referensi data.
   g. Pengisisan nilai-nilai kosong denan nilai .
   h. Penggabungan data.

3. Pengambilan Data (*Loading*)

   Proses pengambilan data yang sudah ditransformasi untuk dimasukan ke dalam data warehouse.

B. Rekam Medis

Rekam medis adalah berkas yang berisikan catatan dan dokumen tentang identitas pasien, pemeriksaan, pengobatan, tindakan dan pelayanan lain yang diberikan kepada pasien (PermenkesRI, No. 269/MENKES/PER/III/2008).

Rekam medis harus dibuat secara tertulis, lengkap dan jelas atau secara elektonik. Penyelenggaraan rekam medis dengan menggunakan teknologi informasi elektronik diatur oleh peraturan tersendiri. Informasi dalam rekam medis dijaga kerahasiannya oleh dokter, tenaga kesehatan dan petugas pengelola serta pimpinan sarana pelayanan kesehatan.

Informasi dalam rekam medis dapat dibuka dalam hal :

1. Untuk kepentingan kesehatan pasien.
2. Memenuhi permintaan aparatur penegak hukum dalam rangka penegakan hukum atas permintaan pengadilan.
3. Permintaan/persetujuan pasien.
4. Permintaan institusi/lembaga berdasarkan ketentuan perundang-undangan.
5. Untuk kepentingan penelitian, pendidikan dan audit medis sepanjang tidak menyebutkan nama pasiennya.

Pelayanan rekam medis bukan pelayanan dalam bentuk pengobatan, tetapi merupakan bukti pelayanan, financial, aspek hokum dan ilmu pengetahuan. Peran rekam medis sangat dibutuhkan untuk

mengelola bahan bukti pelayanan kesehatan dengan aman, nyaman, efisien, efektif dan rahasia.

Pemanfaatan rekam medis dapat dipakai sebagai :

1. Pemeliharaan kesehatan dan pengobatan pasien.
2. Alat bukti dalam proses penegakkan hukum, disiplin kedokteran dan kedokteran gigi dan penegakkan etika kedokteran dan kedokteran gigi.
3. Keperluan pendidikan dan penelitian.
4. Dasar pembayaran biaya pelayanan kesehatan.
5. Data statistic kesehatan.

C. Kode Penyakit Internasional

Kode penyakit internasional adalah standar klasifikasi diagnostic penyakit dan masalah kesehatan lainnya yang ditetapkan menurut criteria tertentu oleh WHO (*World Health Organizations*) untuk dipakai negara-negara di dunia. Kode penyakit internasional lebih dikenal dengan nama ICD-10 (*International Statistical Classification of Diseases and Related Health Problem – 10th*).

Fungsi ICD-10 sebagai berikut :

1. Sebagai standar istilah diagnosis dan prosedur medis.
2. Sebagai catatan medis yang sistematik.

3. Sebagai alat untuk analisis, menerjemmahkan dan membandingkan peristiwa penyakit dan kematian yang telah dikumpulkan di berbagai tempat dan Negara pada saat yang berlainan.
4. Sebagai alat standar penerjemah nama penyakit menjadi kode atau sandi dalam bentuk alfanumerik
5. Untuk mempermudah penyimpanan, pencarian dan analisis suatu penyakit.
6. Sebagai standar dalam pencatatan untuk keperluan epidemologi dan berbagai masalah dan upaya penyelesian kesehatan.
7. Sebagai dasar pengelompokkan pembayaran rumah sakit.

Keunggulan ICD-10 yaitu dapat menganalisis keadaan kesehatan suatu kelompok penduduk serta dapat memantau kasus baru (insiden) dan semua kasus (prevalensi) penyakit dan masalah kesehatan lain dalam hubungannya dengan beberapa variabel seperti ciri dan keadaan orang yang terkena penyakit. Klasifikasi penyakit sesuai ICD-10 ditunjukkan pada tabel 2.1.

D. Penelitian Sejenis

Ada beberapa penelitian terdahulu yang meneliti tentang klasifikasi rekam medis antara lain :

1. ***Patient Status Classification by using Rule Based Sentence Extraction and BM25-kNN Based Classifier*****,** oleh Eiji Aramaki, Ph.D, Tekeshi Imai, Ph.D, Kengo Miyo, Ph.D, Kazuhiko Ohe, Ph.D.,

M.D dari The University of Tokyo Hospital, Japan – 2007. Penelitian ini menitikberatkan pada Klasifikasi pasien perokok berdasarkan data rekam medisnya menggunakan algoritma *Rule Based Sentence Extraction* dan BM25-kNN.

2. ***System and Method for Large Scale Code Classification for Medical Patient Records*****,** oleh Jinbo Bi, Lucian Vlad Lita, Radu Stefan Niculescu, R.Bharat Rao, Shipeng Yu, dari Amerika – 2008 dimuat dalam *Patent Application Publication – US, May 13,2008*. Penelitian ini difokuskan pada pengkodean diagnosis pasien dalam skala besar sesuai standar pengkodean diagnosis penyakit pasien secara internasional yang dinamakan IDC-9 *(International Classification of Diseases)* menggunakan algoritma SVM, *Ridge Regretion* dan *Weighted Ridge Regretion*.
3. ***An Effective Retrieval of Medical Record using Data Mining Techniques.*** Oleh T.Sakthimurugan dan S.Poonkuzhali dari Department of Computer Science and Engineering – Rajalakshmi Engineering College, Chennai-India, dimuat dalam *International Journal of PharmaceuticalScience and Heath Care*, Vol 2 - April 2012. Penelitian ini difokuskan pada pencarian hubungan antara penyakit dan pengobatannya berdasarkan data rekam medis pasien menggunakan dua algoritma yaitu Algoritma *Key Word Searching* untuk pencarian kata kuncinya dan algoritma k-NN *(k-*

Nearest Neighbor) untuk mendapatkan klasifikasi hubungan antara penyakit dan pengobatannya.

Penelitian dalam negeri mengenai Rekam Medis masih didominasi oleh Pembuatan Sistem Informasi Rekam Medis dan Rumah Sakit serta tinjauan yuridis rekam medis, seperti dilakukan beberapa peneliti berikut :

1. **Analisis dan Perancangan Sistem Informasi Rekam Medis Puskesmas Induk Banguntapan II Bantul.** Oleh Uun Kurniasih, STIMIK AMIKOM Yogyakarta,2010.
2. **Rekam Medis dan Sistem Informasi Kesehatan di Sarana Pelayanan Kesehatan Primer (PUSKESMAS).** Oleh dr,Sharon Gondodiputro, MARS. Bagian Ilmu kesehatan Masyarakat, Fak. Kedokteran Universitas Padjadjaran Bandung,2007.
3. **Tinjauan Yuridis terhadap Rekam Medis.** Oleh Anny Retnowati, Justitia Et Pax, Vol.26 No.1, juni 2001, pp 1-12.
4. **Tinjauan Pelepasan Informasi Rekam Medis dalam Menjamin Aspek Hukum Kerahasiaan Rekam Medis di RSUD Dr. H.Moch. Ansari Saleh Banjarmasin.** Oleh Enggar Normanto, Program Perekam dan Informasi Kesehatan, STIKES HUSADA BORNEO Banjarbaru, 2011.
5. **Arsip Rekam Medis *(Medical Record)* serta Pemanfaatan Data Non Medis dalam Mendukung Pencapaian Visi Misi Instansi.**

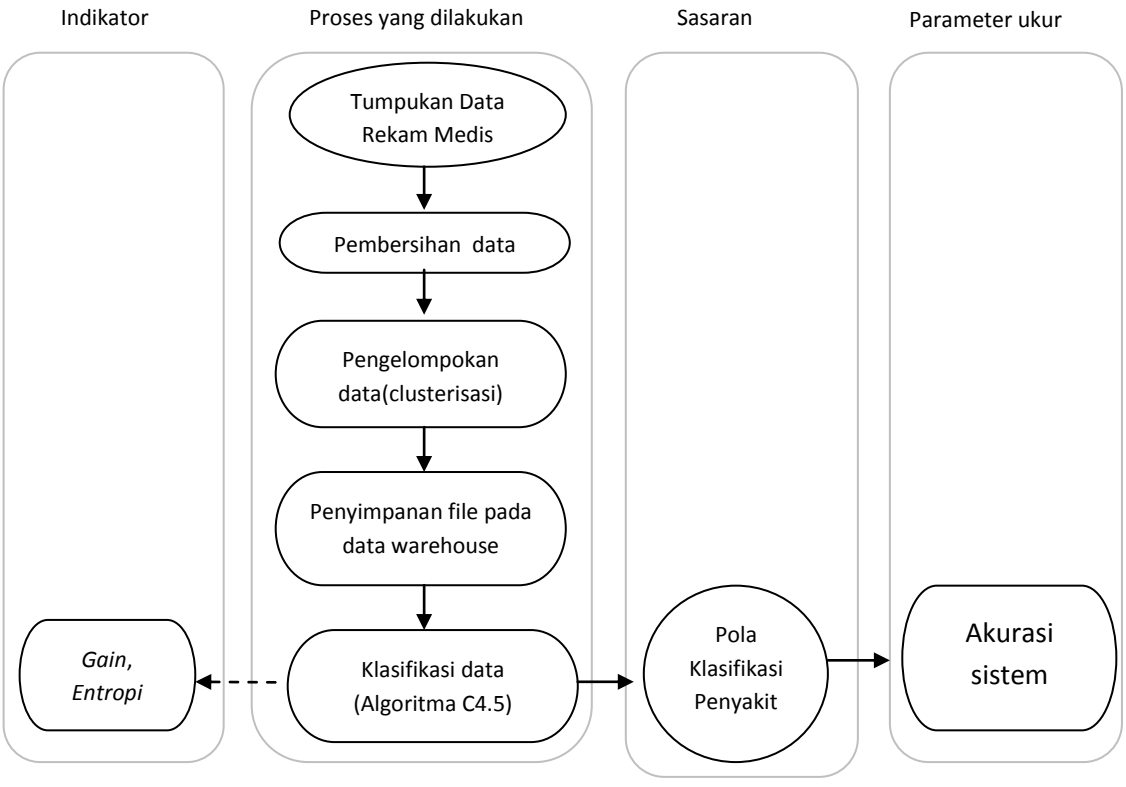
Oleh Emmi Haryanto, Badan Perpustakaan dan Arsip Daerah Propinsi DIY, 2009.

Beberapa penelitian menggunakan algoritma C4.5 yang dipublikasikan pada jurnal-jurnal terkemuka di Indonesia, diantaranya :

1. **Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit,** oleh Angga Ginanjar Mabrur dan Riani Lubis – Prodi Teknik Informatika- Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia, yang dimuat pada Jurnal Komputer dan Informatika (KOMPUTA) edisi I Vol 1, Maret 2012. Penelitian ini difokuskan pada perancangan aplikaksi data mining yang berfungsi untuk memprediksi kritteria nasabah kredit yang berpotensi melakukan kredit terhadap bank. Algoritma yang dipakai adalah C4.5 dengan data yang digunakan berupa data angsuran nasabah Bank XY pada bulan Juni 2009 dalam format Microsoft Excel.
2. **Klasifikasi Data Nasabah sebuah Asuransi menggunakan Algoritma C4.5,** oleh Sunjana – Universitas Widyatama, dipublikasikan pada Seminar Nasional Aplikasi Teknologi Informasi 2010 (SNATI 2010), Yogyakarta, 19 Juni 2010. Penelitian ini difokuskan pada pengelompokan nasabah ke kelas lancer dan kelas tidak lancer menggunakan algoritma C4.5. Hasilnya digunakan oleh asuransi untuk memprediksi nasabah baru yang mau bergabung.

3. **Pemanfaatan Data Mining untuk Prakiraan Cuaca,** oleh Subekti Mujiasih – Pusat Meteorologi Penerbangan dan Maritim BMKG Jakarta, dipublikasikan pada Jurnal Meteorologi dan Geofisika, Vol 12 No 2 – Sept'2011, pp 189-195. Penelitian ini difokuskan pada pembentukan model prediksi prakiraan cuaca menggunakan *Association Rule* dan *Classification (Classification Tree, C4.5, Random Forest)*. Hasil penelitian ini menyimpulkan bahwa Association Rule mempunyai tingkat akurasi prediksi 60,9% sedangkan C4.5 mempunyai tingkat akurasi prediksi 68,5% sehingga model yang dipilih adalah model prediksi dari C4.5.

E. Kerangka Pemikiran



Gambar 2.3. Kerangka Pikir

Penelitian ini berawal dari permasalahan sebagai berikut : adanya penumpukan data rekam medis pasien di rumah sakit yang terus terakumulasi setiap hari tanpa adanya pendayagunaan lebih lanjut secara maksimal dari tumpukan data tersebut, sehingga perlu adanya penerapan ilmu data mining khususnya metode klasifikasi untuk menghasilkan klasifikasi penyakit berdasarkan ICD-10.

Pada penelitian ini, data rekam medis yang digunakan adalah data rekam medis pasien tahun 2010 yang diambil di RSUD Malinau – Kab. Malinau – Propinsi Kalimantan Timur.

Metode yang diusulkan adalah algoritma C4.5 untuk mengklasifikasi penyakit berdasarkan data rekam medis pasien di rumah sakit. Sebelum melakukan klasifikasi data terlebih dahulu dilakukan klusterisasi terhadap data umur dan data diagnose penyakit.

Indikator yang diobservasi yaitu *Gain* dan *Entropi* sebagai parameter dalam menentukan atribut sebagai akar pohon keputusan. Sasaran penelitian ini pada performansi algoritma C4.5 untuk klasifikasi penyakit berdasarkan data rekam medis. Parameter ukurnya berupa tingkat akurasi dan waktu komputasi dari algoritma tersebut.