IMPLEMENTASI DATA MINING CLASSIFICATION DENGAN METODE DECISION TREE (MENGGUNAKAN ALGORITMA ID3 DAN C.45)

Amirhud Dalimunthe

Jurusan Pendidikan Teknik Elektro Unimed, Jl. Willem Iskandar Pasar V Medan Estate

e-mail : amir_unimed@yahoo.co.id

ABSTRAK

Data mining merupakan proses analisis data dengan menggunakan perangkat lunak untuk menemukan suatu pola dan aturan dalam himpunan data. Data mining mampu menganalisa data yang besar menjadi informasi berupa pola yang mempunyai arti bagi pendukung keputusan. Salah satu teknik yang ada pada data mining adalah classification yaitu proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Dalam penelitian ini dilakukan suatu analisis data yang mengimplementasikan salah satu metode dalam classification yaitu decision tree dengan algoritma ID3 dan C.45. Pengujian data dilakukan dengan menggunakan software data mining, yaitu Sipina_W© V.25, dimana output yang dihasilkan berupa data diskrit. Analisis yang dilakukan oleh software ini menggunakan metode decision tree dan didalam melakukan analisis software ini menggunakan algoritma ID3 dan algoritma C4.5. Hal ini sesuai dengan metode analisis dan algoritma yang ditentukan oleh penulis. Beberapa methode analisis yang tersedia didalam software ini antara lain adalah ID3, C.45, CART (Classification and Regression Tree), dan lain-lain.

Kata kunci : Data mining, Classification, Decision tree, ID3.

1. PENDAHULUAN

Seiring dengan makin majunya teknologi khususnya basis data maka timbul teknologi-teknologi untuk aplikasi baru , salah satunya adalah teknologi data mining. Data mining merupakan proses analisa data dengan menggunakan perangkat lunak untuk menemukan suatu pola dan aturan dalam himpunan data. Data mining mampu menganalisa data yang besar menjadi informasi berupa pola yang mempunyai arti bagi pendukung keputusan.

Hasil dari aplikasi data mining tersebut dievaluasi untuk menemukan suatu informasi/pengetahuan baru yang menarik dan bernilai bagi perusahaan, dan kemudian divisualisasikan agar mempermudah bagi user memilih informasi-informasi yang mempunyai arti bagi pendukung keputusan.

Salah satu teknik yang ada pada data mining adalah classification. Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapatmemperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika-maka", berupa decision tree, formula matematis atau neural network. Decision tree adalah salah satu metode classification yang paling populer karena mudah untuk diinterpretasi oleh manusia. Algoritma yang umum digunakan pada metode decision tree adalah algoritma ID3 dan algoritma C4.5. Metode classification yang lain adalah bayesian network, neural network, genetic algorithm, fuzzy, case-based reasoning, dan k-nearest neighbor.

Dari pembahasan diatas maka dilakukan penelitian dengan tujuan dari penelitian untuk 1.Memanfaatkan *data mining* sebagai cabang ilmu di bidang komputer dalam melakukan *ekstraksi* terhadap suatu *set* data dalam skala besar guna mendapatkan informasi yang tersembunyi dari *set* data tersebut. 2.Menerapkan teknik klasifikasi (*data mining classification*) sebagai salah satu teknik *data mining* dengan metode *decision tree* menggunakan algoritma ID3 dan C.45 terhadap suatu *set* data dalam skala besar. 3. Membangun pohon keputusan (*decision tree*) dari suatu sampel data.

Defenisi Data Mining

Terdapat beberapa pendapat yang mendefenisikan data mining, diantaranya adalah sebagai berikut :

1. Data mining adalah proses menemukan pola-pola didalam data, dimana proses penemuan tersebut dilakukan secara otomatis atau semi otomatis dan pola-pola yang ditemukan harus bermanfaat [Han, J., Kamber, M. ,2001].
2. Data mining adalah proses penemuan informasi yang berguna pada penyimpanan data yang besar secara otomatis [Tan P. N., Steinbach, M., Kumar, V. ,2006].
3. Data mining dapat juga didefinisikan sebagai "pemodelan dan penemuan pola-pola yang tersembunyi dengan memanfaatkan data dalam volume yang besar" [Sinclair, C., Pierce, L., Matzner, S.,2000].
4. Data mining adalah mencocokkan data dalam suatu model untuk menemukan informasi yang tersembunyi dalam basisdata [Dunham, H. Margareth ,2002].
5. Data mining atau Knowledge Discovery in Database (KDD) adalah pengambilan informasi yang tersembunyi, dimana informasi tersebut sebelumnya tidak dikenal dan berpotensi bermanfaat. Proses ini meliputi sejumlah pendekatan teknis yang berbeda, seperti clustering, data summarization, learning classification rules [Fayyad, U., Piatetsky-Shapiro, G. dan Smyth, P. ,1996].
6. Data mining, as we use the term, is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful paterns and rules [J.A.Berry dan Linoff, 1997]

Secara sederhana data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di basisdata yang besar. Data mining juga dikenal dengan nama KDD (Knowledge Discovery in Database).

Data mining menggunakan pendekatan discovery-based dimana pencocokan pola (pattern-matching) dan algoritma-algoritma yang lain digunakan untuk menentukan relasi-relasi kunci di dalam data yang diekplorasi. Data mining merupakan komponen baru pada arsitektur sistem pendukung keputusan (Decision Support System) di perusahaan-perusahaan.

Ruang Lingkup Data Mining

Data mining (penambangan data), sesuai dengan namanya, berkonotasi sebagai pencarian informasi bisnis yang berharga dari basis data yang sangat besar. Usaha pencarian yang dilakukan dapat dianalogikan dengan penambangan logam mulia dari lahan sumbernya. Dengan tersedianya basis data dalam kualitas dan ukuran yang memadai, teknologi data mining memiliki kemampuan-kemampuan sebagai berikut:

1. Mengotomatisasi prediksi tren dan sifat-sifat bisnis. Data mining mengotomatisasi proses pencarian informasi di dalam basis data yang besar. Pertanyaan-pertanyaan yang berkaitan dengan prediksi ini dapat cepat dijawab langsung dari data yang tersedia. Contoh dari masalah prediksi ini misalnya target pemasaran, peramalan kebangkrutan dan bentuk-bentuk kerugian lainnya.
2. Mengotomatisasi penemuan pola-pola yang tidak diketahui sebelumnya. Tools data mining "menyapu" basis data, kemudian mengidentifikasi pola-pola yang sebelumnya tersembunyi dalam satu sapuan. Contoh dari penemuan pola ini adalah analisis pada data penjulan ritel untuk mengidentifikasi produk-produk, yang kelihatannya tidak berkaitan, yang seringkali dibeli secara bersamaan oleh kustomer. Contoh lain adalah pendeteksian transaksi palsu dengan kartu kredit dan identifikasi adanya data anomali yang dapat diartikan sebagai data salah ketik (karena kesalahan operator).

Bahasan Teknis Data Mining

Penjelasan umum yang diberikan di atas memberikan pengertian bahwa seolah-olah teknologi data mining adalah teknologi utuh dan berdiri sendiri. Dibandingkan dengan Knowledge Discovery in Database (KDD), istilah data mining lebih dikenal para pelaku bisnis. Pada aplikasinya, sebenarnya data mining merupakan bagian dari proses Knowledge Discovery in Database (KDD). Sebagai komponen dalam Knowledge Discovery in Database (KDD), data mining terutama berkaitan dengan ekstraksi dan penghitungan pola-pola dari data yang ditelaah.

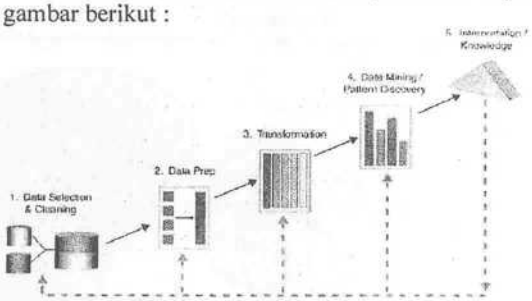
Proses Knowledge Discovery in Database (KDD)

Knowledge Discovery in Databases (KDD) adalah proses untuk mencari dan mengidentifikasi pola (pattern) dalam database. Pola yang ditemukan bersifat valid, potentially useful, dan ultimately understandable. Secara umum proses KDD terdiri dari langkah-langkah sebagai berikut [Han, J., Kamber, M. ,2001]:

1. Pembersihkan data (*data cleaning*), proses menghilangkan *noise* dan data yang tidak konsisten atau data tidak relevan;
2. Pengintegrasian data (*data integration*), penggabungan data dari berbagai basis data ke dalam satu basisdata baru;
3. Transformasi data, data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*;
4. Aplikasi *data mining*, suatu proses di mana metoda diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data;

5. Evaluasi pola *(pattern evaluation)*, untuk mengidentifikasi pola-pola menarik untuk di representasikan kedalam *knowledge based*;
6. Presentasi pengetahuan *(knowledge presentation)*, visualisasi dan penyajian mengenai teknik yang digunakan yang bermanfaat bagi pengguna (*user*).

Proses KDD tersebut diatas dapat dilihat seperti gambar berikut :



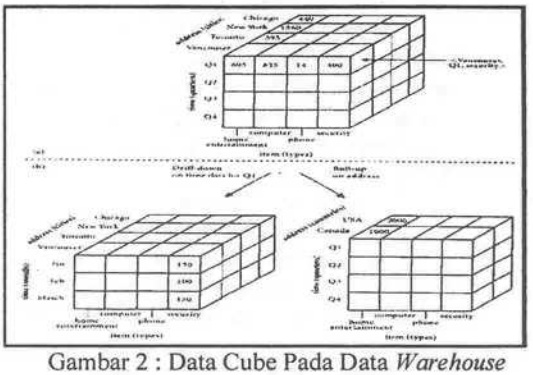
Gambar 1: Tahapan dalam Proses Knowledge Discovery in Database (KDD)

Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base.

Data Warehouse

Biasanya perusahaan memakai database dalam operasi sehari-harinya seperti pencatatan transaksi jual-beli, administrasi pengiriman barang, inventori, penggajian dan sebagainya yang lazim disebut dengan OLTP (*Online Transaction Processing*). Dengan makin besarnya kebutuhan akan analisa data untuk mempertahankan keunggulan dalam kompetisi, banyak perusahaan yang membangun database sendiri yang khusus digunakan untuk menunjang proses pengambilan keputusan (*decision making*) atau disebut juga dengan OLAP (*Online Analytical Processing*).

Berbeda dengan OLTP yang hanya memakai operasi query yang sederhana dan berulang-ulang, query untuk OLAP biasanya lebih rumit , bersifat adhoc, dan tidak melibatkan operasi data update. OLAP juga tidak memakai data operasi sehari-hari begitu saja, tetapi memakai data yang sudah terangkum dengan model data yang disebut data cube. *Data cube* adalah presentasi data multidimensi seperti jenis barang, waktu, lokasi dan sebagainya. Ilustrasi dari data cube ditunjukkan di gambar 2 berikut :



Gambar 2 : Data Cube Pada Data *Warehouse*

Dimensi pada data cube dapat dibuat bertingkat, contohnya dimensi lokasi dapat dibagi menjadi kota, propinsi dan negara. Sedangkan dimensi waktu mencakup jam, hari, minggu, bulan, tahun dan sebagainya. Dengan ini pemakai dapat dengan mudah mendapat rangkuman informasi dari tingkatan dimensi yang lebih luas atau lebih umum seperti negara atau tahun dengan operasi yang disebut *roll-up* seperti ditunjukkan di Gambar 2 diatas. Sebaliknya dengan operasi *drill-down*, pemakai dapat menggali informasi dari tingkatan dimensi yang lebih detil seperti data harian atau data di lokasi yang spesifik.

Data cube yang tersedia pada data *warehouse* memungkinkan pemakai untuk menganalisa data operasi sehari-hari dengan berbagai sudut pandang, dan sangat berguna untuk mengevaluasi suatu asumsi bisnis. Akan tetapi untuk mendapatkan informasi yang tidak diketahui secara eksplisit diperlukan satu tahap lagi yaitu aplikasi teknik *Data Mining*. Disini data *warehouse* merupakan data mentah untuk *Data Mining*. Data *warehouse* sendiri secara periodik diisi data dari OLTP (*Online Transaction Processing*) setelah menjalani pembersihan dan integrasi data. Karena itu ada pula anggapan bahwa *Data Mining* adalah tahap lanjut dari OLAP (*Online Analytical Processing*).

Teknik-Teknik dalam *Data Mining*

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basisdata. Dengan definisi *Data Mining* yang sangat luas, ada banyak jenis teknik analisa yang dapat digolongkan dalam *Data Mining*. Namun disini penulis akan memberikan sedikit gambaran tentang tiga teknik Data Mining yang paling populer, yaitu :

a) Association Rule Mining

Association rule mining atau analisis afinitas (affinity analysis) adalah teknik *mining* untuk menemukan aturan asosiatif antara suatu kombinasi atribut. Ini bisa berupa studi transaksi di superemarket, misalnya seseorang yang membeli susu bayi juga membeli sabun mandi. Disini berarti susu bayi bersama dengan sabun mandi. Karena awalnya berasal dari studi tentang database transaksi pelanggan untuk menentukan kebiasaaan suatu produk dibeli bersama dengan produk apa, maka aturan asosiasi juga sering disebut *market basket* analisis.

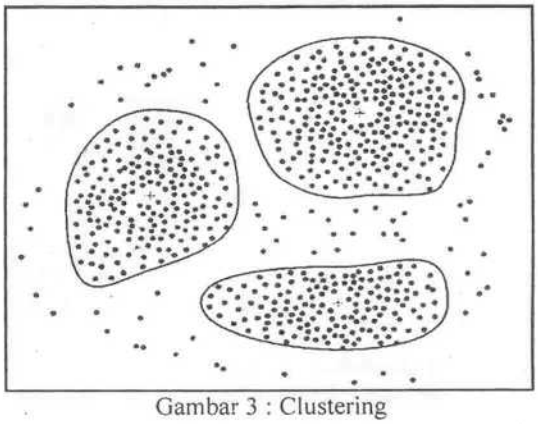
Penting tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, *support* yaitu persentasi kombinasi atribut tersebut dalam basisdata dan *confidence* yaitu kuatnya hubungan antar atribut dalam aturan asosiatif. Algoritma yang paling populer dalam teknik ini dikenal sebagai Apriori dengan paradigma generate and test, yaitu pembuatan kandidat kombinasi item yang mungkin berdasar aturan tertentu lalu diuji apakah kombinasi item tersebut memenuhi syarat support minimum. Kombinasi item yang memenuhi syarat tersebut dinamakan frequent itemset, yang nantinya dipakai untuk membuat aturan-aturan yang memenuhi syarat confidence minimum. Algoritma baru yang lebih efisien bernama FP-Tree [Han, J., Kamber, M. ,2001].

b) Clustering

Berbeda dengan *association rule mining* dan *klasifikasi* dimana kelas data telah ditentukan sebelumnya, *clustering* melakukan pengelompokan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*.

Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar *cluster*. *Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi. Ilustrasi dari *clustering* dapat dilihat di gambar 3 berikut, dimana lokasi dinyatakan dengan bidang dua dimensi, dari pelanggan suatu toko dapat dikelompokkan menjadi beberapa *cluster* dengan pusat *cluster* ditunjukkan oleh tanda positif (+).

Banyak algoritma *clustering* memerlukan fungsi jarak untuk mengukur kemiripan antar data, diperlukan juga metoda attacks untuk normalisasi bermacam atribut yang dimiliki data. Beberapa kategori algoritma clustering yang banyak dikenal adalah metode partisi dimana pemakai harus menentukan jumlah k partisi yang diinginkan lalu setiap data dites untuk dimasukkan pada salah satu partisi, metode lain yang telah lama dikenal adalah metode hierarki yang terbagi dua lagi : *bottom-up* yang menggabungkan cluster kecil menjadi cluster lebih besar dan *top-down* yang memecah cluster besar menjadi cluster yang lebih kecil. Kelemahan metode ini adalah bila salah satu penggabungan/ pemecahan dilakukan pada tempat yang salah, tidak dapat didapatkan cluster yang optimal. Pendekatan yang banyak diambil adalah menggabungkan metode hierarki dengan metode clustering lainnya seperti yang dilakukan oleh Chameleon [G. Karypis, E.-H. Han and V. Kumar,1997]



Gambar 3 : Clustering

c) Classification

Classification adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Hal ini digambarkan sebagai berikut:

Gambar 4: Klasifikasi memetakan atribut x ke dalam label kelas y

Masukan data untuk klasifikasi adalah kumpulan *record*. Setiap *record* dikenal sebagai *instance* atau contoh yang ditandai oleh *tuple* (x,y) dimana x adalah atribut dan y adalah atribut khusus yang menunjukkan label kelas (disebut juga kategori atau atribut target).

Ada beberapa metode yang terdapat pada *Data mining classification*, antara lain adalah decision tree, neural network, k-Nearest Neighbor Classifiers, Case-Based Reasoning dan algoritma genetika.

Pohon Keputusan (*Decision Tree*)

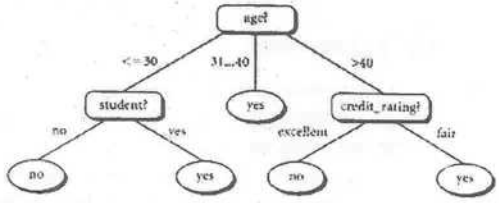
Salah satu metode classification yang paling populer dan mudah untuk diinterpretasi adalah Pohon Keputusan (*Decision Tree*). Metode ini merupakan salah satu fungsional dari *Data mining* yang menggunakan *representasi tree* untuk menentukan

aturan-aturan klasifikasi. Ada dua tipe *decision tree*, yaitu :

- *Classification* tree
  Memberi label dan memasukan record-record ke dalam kelas-kelas yang telah disediakan
- Regression tree
  Membuat estimasi nilai dari sebuah variabel target yang berdasar pada nilai numerik.

Decision tree adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas[Andrew W. Moore].

Contoh dari decision tree dapat dilihat di gambar 6 berikut:



Gambar 6 : Decision Tree

Disini setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan kelas data. Contoh di gambar 6 diatas adalah identifikasi pembeli komputer. Dari decision tree tersebut diketahui bahwa salah satu kelompok yang potensial membeli komputer adalah orang yang berusia di bawah 30 tahun dan juga pelajar.

Proses classification biasanya dibagi menjadi dua fase yaitu learning dan test. Pada fase learning, sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan. Kemudian pada fase test model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut.

Algoritma ID3 dan Algoritma C4.5

Sebelum membahas algoritma C4.5 perlu dijelaskan terlebih dahulu algoritma ID3 karena C4.5 adalah ekstensi dari algoritma decision-tree ID3. Algoritma ID3 atau C4.5 ini secara rekursif membuat sebuah *decision tree* berdasarkan *training* data yang telah disiapkan. Algoritma ini mempunyai inputan berupa *training samples* dan *samples*. *Training samples* berupa data contoh yang akan digunakan untuk membangun sebuah *tree* (pohon). Sedangkan *samples* merupakan field-field data yang nantinya akan kita gunakan sebagai parameter dalam melakukan klasifikasi data. Berikut adalah algoritma dasar dari ID3 dan C4.5

1. **Algoritma ID3** [Han,Jiawei and Khamber, Micheline ,2001]

   Input : *Training* samples, samples
   Output : *Decision tree*

   Method :
   (1) Create node N;
   (2) **If** samples are all of the same class, C **then**
   (3) Return N as a leaf node labeled with the class C;
   (4) **if** atribute-list is empty then
   (5) Return N as a leaf node labeled with the most common class in samples; // majority voting
   (6) select test-atribute, atribute among atribute-list with the highest information gain;
   (7) label node N with test-atribute;
   (8) for each known value ai of test-atribute // partition the samples
   (9) grow a branch from node N for the condition test-atribute = ai;
   (10) let s_i be the set of samples in samples for which test-atribute = a_i; // a partition
   (11) **if** s_i is empty then
   (12) attach a leaf labeled with the most common class in samples;
   (13) **else** attach the node returned by Generate_decision_tree(s_i, attriute-list-test-atribute);

2. **Algoritma C4.5** [Solorio, Thamar I. and Fuentes, Olac]

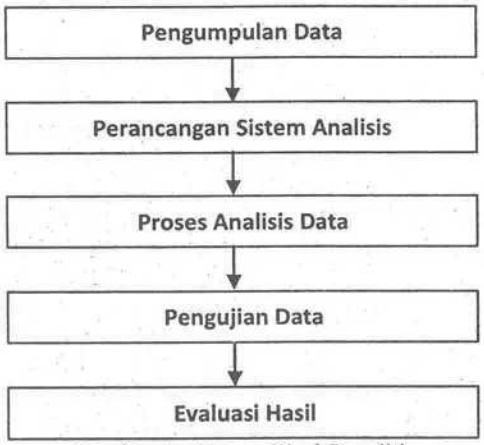
   (1) Build the *decision tree* form the *training* set (conventional ID3).
   (2) Convert the resulting tree into an equivalent set of rules. The number of rules is equivalent to the number of possible paths from the root to a leaf node.
   (3) Prune (generalize) each rule by removing preconditions that increase *classification* accuracy.
   (4) Sort pruned rules by their accuracy, and use them in this order when classifying future test examples.

METODE PENELITIAN

Kerangka Kerja (*Framework*) Penelitian

Metode penelitian adalah gambaran langkah-langkah yang akan dilakukan dalam melakukan penelitian. Hal ini perlu diterapkan agar penelitian dapat dilakukan dengan terstruktur. Langkah-langkah yang dilakukan harus mencakup persoalan mulai dari mempelajari masalah sampai dengan adanya suatu

analisis yang dapat dihasilkan sehingga masalah dapat teratasi. Maka disini ditetapkan beberapa tahapan yang dapat dilakukan dan dapat dilihat pada gambar 3.1 berikut ini :



Gambar 7 : *Frame Work* Penelitian

Metodologi Pengumpulan Data

Metode pengumpulan data adalah cara-cara yang dapat digunakan peneliti untuk mengumpulkan data. Beberapa metodologi pengumpulan data yang digunakan dalam penelitian ini adalah :

1. Study Literatur untuk memahami materi yang berhubungan dengan penelitian. Metode yang dilakukan adalah dengan cara mempelajari teori-teori/ literatur dan buku-buku ilmiah serta referensi-referensi yang berhubungan dengan objek tesis sebagai dasar dalam penelitian ini. Literatur yang dipelajari disini adalah literatur yang berhubungan *Data Mining* dan aplikasinya terhadap bidang-bidang teknologi, pemasaran dan bisnis.
2. Pengumpulan data lapangan, berupa data dari pemain tennis (petennis) yang mendaftar untuk bermain tennis lapangan di suatu lapangan tennis.
3. Pengamatan Langsung (*Observasi*), teknik ini dilakukan untuk mengetahui secara langsung keadaan lapangan tennis tersebut.

Perancangan Sistem Analisis

Beberapa hal yang harus dipersiapkan dalam melakukan proses analisis data dalam penulisan ini adalah sebagai berikut:

1. Identifikasi Masalah;
   Pada tahap ini dimulai dengan menetapkan masalah yang akan dianalisis dalam penelitian ini dan mempelajari metodologi yang tepat dalam melakukan proses analisis data guna mendapatkan informasi yang diharapkan dari hasil analisis yang dilakukan.
2. Menetapkan variabel-variabel yang diperlukan.
   Untuk dapat menghasilkan suatu *output* berupa informasi yang sesuai dengan yang diharapkan dalam melakukan proses analisis data, maka harus ditetapkan terlebih dahulu variabel-variabel dan kebutuhan yang diperlukan untuk persyaratan sistem yang berpengaruh terhadap proses penganalisaan data.
3. Menetapkan batasan-batasan dalam sistem analisis data.

Proses Analisis Data

Pada tahap ini dilakukan proses pengolahan dan analisis terhadap data yang sudah disiapkan pada tahap pengumpulan data berdasarkan literatur-literatur yang ada, observasi lapangan, data disusun dalam bentuk tabel sederhana.

Pengujian Data

Penelitian dan pengujian yang dilakukan adalah untuk mendapatkan hasil dari analisis yang bertujuan untuk memperoleh informasi yang tersembunyi dari set data yang tersedia. Spesifikasi peralatan atau alat bantu penelitian yang digunakan adalah sebagai berikut:

Hardware :

a. Notebook Dell Inspiron 6000
b. Processor Intel P4 Mobile 1.73 GHz
c. Space Harddisk 80 GB.
d. Memori 512 MB

Software :

a. Microsoft Windows XP Profesional
b. *Software Data Mining* : Sipina_W© V.25

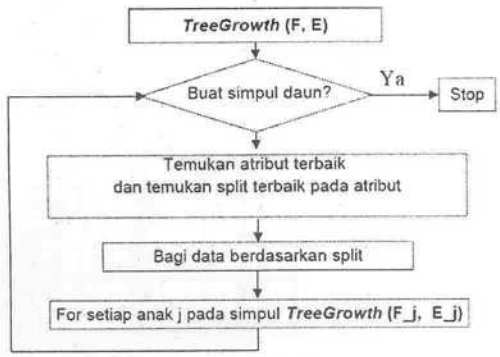
Evaluasi Hasil

Evaluasi terhadap hasil analisis dari pengujian data yang dilakukan perangkat lunak Sipina_W© V.25 ini dilakukan untuk menjelaskan *output* yang dihasilkan. *Output* yang dihasilkan berupa *decission* tree yang menggambarkan informasi dari set data yang sudah dianalisis menggunakan perangkat lunak ini.

HASIL PEMBAHASAN

Konstruksi Pohon Keputusan

Setelah langkah di atas dilakukan, dilakukan penerapan dari algorithma ID3 untuk membangun pohon keputusan (*decision tree*). Rangka dari algoritma ID3 disebut *TreeGrowth* seperti yang terlihat pada gambar di bawah ini. Masukan terdiri dari *record* pelatihan (*training record)* E dan atribut F. Cara kerja algoritma ini yaitu dengan memilih atribut yang terbaik untuk memisahkan data secara rekursif dan mengembangkan simpul daun pada tree sampai ditemui kriteria untuk berhenti.



Gambar 8 : Algoritma *Induksi of Decision"3"(ID3)*

Berdasarkan algoritma diatas, untuk mendapatkan *decision tree* yang terbaik (minimal), maka dilakukan perhitungan *information gain* dari setiap atribut untuk mendapatkan atribut yang akan menyediakan prediksi terbaik untuk target atribut kelulusan.

Penghitungan Information Gain

Information gain adalah salah satu a*tribute selection measure* yang digunakan untuk memilih test atribute tiap node pada tree. Atribut dengan information gain tertinggi dipilih sebagai test atribut dari suatu node [Han,Jiawei and Khamber, Micheline,2001]. Ada 2 kasus berbeda pada saat penghitungan Information Gain, pertama untuk kasus penghitungan atribut tanpa *missing value* dan kedua, penghitungan atribut dengan *missing value.*

1. Penghitungan Information Gain tanpa missing value

Misalkan S berisi *s* data samples. Anggap atribut untuk class memiliki *m* nilai yang berbeda, C_i (untuk i = 1, ...,I). anggap s_i menjadi jumlah samples S pada class C_i. Maka besar information-nya dapat dihitung dengan :

$$I(s_1, s_2,\ldots,s_m) = -\sum_{i=1}^{m} p_i * \log_2(p_i)$$

dimana $p_i = \frac{si}{s}$ adalah probabilitas dari sample yang mempunyai class C_i.

Misalkan atribut A mempunyai *v* nilai yang berbeda, $\{a_1, a_2,...,a_v\}$. Atribut A dapat digunakan untuk mempartisi S menjadi *v* subset, $\{S_1, S_2,...,S_v\}$, dimana S_j berisi samples pada S yang mempunyai nilai a_j dari A. Jika A terpilih menjadi test atribut (yaitu, best atribut untuk splitting), maka subset-subset akan berhubungan dengan pertumbuhan node-node cabang yang berisi S. Anggap s_{ij} sebagai jumlah samples class C*i* pada subset S_j. Entropy, atau nilai information dari subset A adalah:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_1, s_2,\ldots,s_m)$$

$\frac{s_{1j} + \ldots + s_{mj}}{s}$ adalah bobot dari subset *j*th dan jumlah samples pada subset (yang mempunyai nilai a_j dari A) dibagi dengan jumlah total samples pada S. Untuk subset S_j,

$$I(s_{1j}, s_{2j},\ldots,s_{mj}) = -\sum_{i=1}^{m} p_{ij} * \log_2(p_{ij})$$

dimana $p_{ij} = \frac{s_{ij}}{|s_j|}$ adalah probabilitas sample S_j yang mempunyai class C*i*.

Maka nilai information gain atribut A pada subset S adalah

$$Gain(A) = I(s_1, s_2,\ldots,s_m) - E(A)$$

Pada penelitian ini penulis menggunakan data play tennis berikut. Dari data tersebut ingin mencari apakah petennis akan masuk class Yes atau No berdasarkan data berikut:

Tabel 1. Data play tennis tanpa missing value

Outlook	Temp.	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	78	False	Yes
Rain	70	96	False	Yes
Rain	68	80	False	Yes
Rain	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rain	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes
Rain	71	80	True	No

Dari data pada tabel kita akan mencoba untuk membangun sebuah classifier yang berdasarkan atribut Outlook, Temperature, Humidity dan Windy. Disana ada dua kelas yaitu Yes dan No. Dan ada 14 examples, 5 examples menyatakan No dan 9 examples menyatakan Yes. Maka,

$$I(s_1,s_2) = I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0{,}940$$

Entropy untuk atribut Outlook adalah :

$$E(Outlook) = \frac{5}{14} * I(2,3) + \frac{4}{14} * I(4,0) + \frac{5}{14} * I(3,2)$$

$$= 0{,}694$$

Dengan nilai Gain(Outlook) yaitu:

$$Gain(Outlook) = I(s_1, s_2) - E(Outlook)$$

= 0,94 – 0,694
= 0,246

dengan menggunakan cara yang sama, Gain dari semua atribut dapat dicari.

Gain (Outlook) = 0,246
Gain (Humidity) = 0,151
Gain (Windy) = 0,048
Gain (Temperature) = 0,029

Setelah nilai information gain pada semua atribut dihitung, maka atribut yang mempunyai nilai information gain terbesar yang dipilih menjadi test atribut.

2. Penghitungan Information Gain dengan missing value

Untuk atribut dengan *missing value* penghitungan information gain-nya diselesaikan dengan Gain Ratio. Sebelum menghitung gain ratio terlebih dahulu dihitung I ($s_1, s_2,\ldots,s_m$) dan E(A).

$$I(s_1, s_2,\ldots,s_m) = -\sum_{i=1}^{m} p_i * \log_2(p_i)$$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s} I(s_1, s_2,\ldots,s_m)$$

Dimana penghitungan I ($s_1, s_2,\ldots,s_m$) dan E(A) hanya dilakukan pada atribut yang ada nilainya.

Kemudian untuk mencari gain dari atribut A dihitung dengan rumus sebagai berikut :

Gain (A) = Prob S yang diketahui * E(A)

dimana,

A = atribut dengan missing value yang sedang dicari nilai gain-nya,

S = jumlah samples pada subset A yang diketahui nilainya.

Sedangkan nilai split pada atribut A dinyatakan dengan :

$$Split(A) = -u * \log_2 u - \sum_{i=1}^{m} p_j * \log_2(p_j)$$

dimana,

u adalah probabilitas samples pada atribut A yang merupakan *missing values.*

$p_j = \frac{s_j}{|s|}$ adalah probabilitas sample S_j yang diketahui nilainya.

Nilai Gain Ratio pada atribut A :

Gain Ratio(A) = Gain(A) / Split(A)

Sebagai contoh : kita ingin mencari apakah petenis akan masuk class yes atau no berdasarkan data berikut:

Tabel 2. data play tennis dengan missing value

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	true	yes
Sunny	80	90	true	no
Sunny	85	85	false	no
Sunny	72	95	false	no
Sunny	69	70	false	yes
?	72	90	true	yes
Cloudy	83	78	false	yes
Cloudy	64	65	true	yes
Cloudy	81	75	false	yes
rain	71	80	true	no
rain	65	70	true	no
rain	75	80	false	yes
rain	68	80	false	yes
rain	70	96	false	yes

Pertama, kita menghitung frekuensi pada Outlook sebagai berikut :

	Yes	No	Total
Sunny	2	3	5
Cloudy	3	0	3
Rain	3	2	5
Total	8	5	13

Untuk data pada tabel 2.2 maka penghitungan information gainnya adalah sebagai berikut :

$$I(s_1,s_2) = I(8,5) = -\frac{8}{13}\log_2\frac{8}{13} - \frac{5}{13}\log_2\frac{5}{13}$$
$$= 0.961$$

$$I(outlook) = \frac{5}{13}*(-\frac{2}{5}*\log_2\frac{2}{5} - \frac{3}{5}*\log_2\frac{3}{5}) + \frac{3}{13}*(-\frac{3}{3}*\log_2\frac{3}{3} - \frac{0}{3}*\log_2\frac{0}{3}) + \frac{5}{13}*(-\frac{3}{5}*\log_2\frac{3}{5} - \frac{2}{5}*\log_2\frac{2}{5}) = 0.747$$

$$Gain\ (Outlook) = \frac{13}{14} * (0.961-0.747) = 0.199$$

$$Split = -\frac{5}{14}*\log_2\frac{5}{14} - \frac{3}{14}*\log_2\frac{3}{14} - \frac{5}{14}*\log_2\frac{5}{14} - \frac{1}{14}*\log_2\frac{1}{14}$$
$$= 1.809$$

$$Gain\ ratio = \frac{0,199}{1,809} = 0.110$$

Setelah semua Gain diketahui maka dapat ditentukan atribut mana yang layak menjadi root. Atribut yang layak adalah atribut yang mempunyai Gain terbesar.

KESIMPULAN

- *Data mining* dapat diimplementasikan dengan metode *decision tree* menggunakan algoritma ID3 dan C.45.
- Perangkat lunak Sipina_W©V.25 dapat digunakan sebagai perangkat lunak bantu untuk melakukan analisis terhadap suatu set data dan

mendapatkan informasi yang tersembunyi dari set data tersebut.

- Pemahaman yang baik tentang teori dan konsep pemakaian perangkat lunak Sipina_W© V.25 akan sangat berpengaruh terhadap output yang dihasilkan.

DAFTAR PUSTAKA

Han,Jiawei and Khamber, Micheline. "*Data Mining : Concepts and Techniques*", Morgan Kaufmann Publishers, San Francisco,USA, 2001.

Tan P. N., Steinbach, M., Kumar, V. (2006), *Introduction to Data Mining*, Addison Wesley.

Fayyad, U., Piatetsky-Shapiro, G. dan Smyth, P. (1996), *From Data Mining to Knowledge Discovery in Databases*, AAAI and The MIT Pres, 37-53.

Berry, Michael J.A. and Gordon Linoff, "*Data Mining Techniques: For Marketing, Sales, and Customer*", John Wiley & Sons, New York,1997.

Dunham, H. Margareth (2002), *Data Mining: Introductory and Advanced*,

Prentice Hall.

Andrew W. Moore, "*Decision Trees*", Carnegie Mellon University. http://www.cs.cmu.edu/~awm

Sinclair, C., Pierce, L., Matzner, S. (2000), *An Application of Machine Learning to Network Intrusion Detection*, Applied Research Laboratory, University of Texas at Austin.