BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Sistem *data mining* akan lebih efektif dan efisiensi dengan komputerisasi yang tepat. Sistem *data mining* mampu memberikan informasi yang tepat dan pengelohan data yang akurat sehingga bisa langsung digunakan dan dilaporkan.

Penelitian mengenai *data mining* sebelumnya sudah banyak dilakukan, tetapi tempat dan program aplikasi yang digunakan berbeda – beda. Adapun sistem *data mining* yang berkaitan yang pernah dibuat adalah sebagai berikut:

Menurut penelitian yang dilakukan oleh Swastina. L. (2013), yang berjudul "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa". Metode yang digunakan dalam penelitian ini adalah Algoritma C4.5, Algoritma C4.5 digunakan untuk menentukan jurusan yang akan diambil oleh mahasiswa sesuai dengan latar belakang, minat dan kemampuannya sendiri. Parameter pemilihan jurusan adalah Indeksi Prestasi Kumulatif Semester 1 dan 2. Hasil eksperimen dan evaluasi dari penelitian tersebut menunjukan bahwa Algoritma Decision Tree C4.5 akurat diterapkan untuk penentuan kesesuaian jurusan mahasiswa dengan tingkat akurasi 93,31% dan akurasi rekomendasi jurusan sebesar 82,64%.

Andriani. A. (2013), melakukan penelitian tentang "Sistem Pendukung Keputusan Berbasis *Decision Tree* dalam Pemberian Beasiswa (Studi Kasus: AMIK "BSI Yogyakarta"). Tujuan penelitian ini adalah membuat klasifikasi mahasiswa penerima beasiswa dengan *Decision Tree* yang menggunakan Algoritma C4.5. Hasil klasifikasi dievaluasi dan divalidasi dengan *Confusion*

Matrix dan *Kurva ROC* untuk mengetahui tingkat akurasi *Decision Tree* dalam membuat klasifikasi beasiswa. Hasil klasifikasi digunakan untuk membuat sistem pendukung keputusan dalam pemberian beasiswa. Sistem yang digunakan dibuat dengan *Microsoft Visual Basic* 6.0. Dengan adanya sistem pendukung keputusan ini dapat mempermudah dan mempercepat pengambilan keputusan untuk pemberian beasiswa.

Julianto. W. *et al* (2014), penelitian yang pernah mereka lakukan adalah "Algoritma C4.5 untuk Penilaian Kinerja Karyawan". Dengan menggunakan algoritma C4.5 yang menggunakan teknik *data mining* untuk membuat pohon keputusan, algoritma ini dimulai dengan memasukkan data *training* ke dalam simpul akar pada pohon keputusan. *Data training* adalah sampel yang digunakan untuk membangun model *classifier* dalam hal ini pohon keputusan. Adapun hasil analisis sebagai berikut: berdasarkan evaluasi yang dilakukan dapat diketahui bahwa proses pembentukan pohon menggunakan teknik *pruning* memiliki kecepatan yang lebih tinggi karena penyederhanaan pohon, tetapi tidak selalu memliki akurasi yang lebih besar, dan perbedaan pohon keputusan yang dihasilkan disebabkan oleh perbedaan jumlah *data training* yang digunakan pada masing-masing partisi.

Dari ketiga peneliti yang telah dilakukan tersebut, klasifikasi *Decision Tree* dengan Algoritma C4.5 digunakan oleh para peneliti sebagai solusi untuk mengambil keputusan yang diharapkan mampu membantu dalam pengambilan keputusan dengan lebih mudah dan cepat. Begitu juga dengan penelitian ini, klasifikasi *Decision Tree* dengan Algoritma C4.5 sebagai solusi pengambilan

keputusan bagi pihak Universitas dalam menentukan status dosen, sehingga mempermudah pihak universitas dalam menentukan status dosen. Adapun perbedaan yang ada yaitu dalam penelitian ini, data yang digunakan data yang tersimpan didalam *database server* dan pengambilan data menggunakan *software Sql Server* 2014.

2.2. Landasan Teori

2.2.1. *Data Mining*

Menurut Gunadi dan Sensuse (2012) *Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. *Data mining* merupakan proses analisa data untuk menemukan suatu pola dari kumpulan data tersebut. *Data mining* mampu menganalisa data yang besar menjadi informasi berupa pola yang mempunyai arti bagi pendukung keputusan.

Menurut Hermawati (2013) *Data mining* adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan secara otomatis.

Menurut Kursini dan Luthfi (2009) *Data Mining* merupakan suatu proses otomatis terhadap data yang sudah ada. Dan data yang akan diproses berupa data yang sangat besar.

Menurut Han dan Kamber (2006), rancangan bangun dari *data mining* yang khas memiliki beberapa komponen utama yaitu:

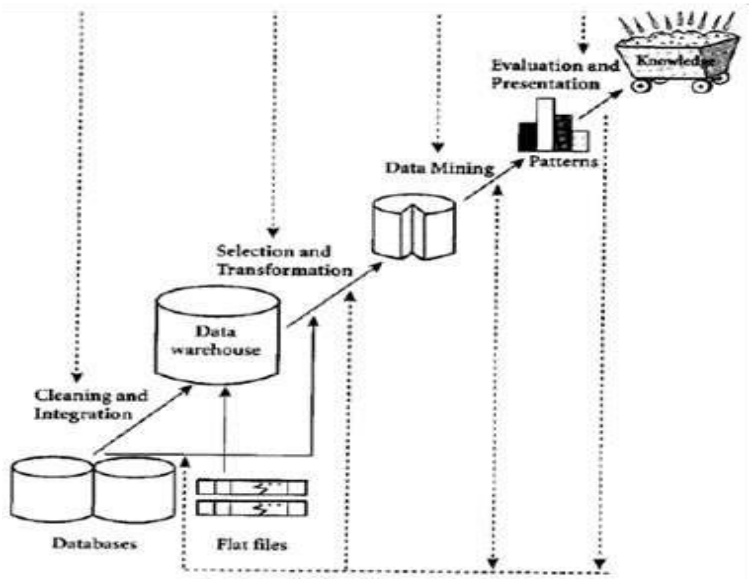
- *Database, data warehouse,* atau tempat penyimpanan informasi lainnya.
- *Server database* atau *data warehouse.*

- *Knowledge base.*
- *Data mining engine.*
- *Pattern evolution module*.
- *Graphical user interface.*

Tahap-tahap *data mining* salah satu tuntutan dari *data mining* ketika diterapkan pada data berskala besar adalah diperlukan metodologi sistematis tidak hanya ketika melakukan analisa saja tetapi juga ketika mempersiapkan data dan juga melakukan interpretasi dari hasilnya sehingga dapat menjadi aksi ataupun keputusan yang bermanfaat. Karenanya *data mining* seharusnya dipahami sebagai suatu proses, yang memiliki tahapantahapan tertentu dan juga ada umpan balik dari setiap tahapan ke tahapan sebelumnya.

Pada umumnya proses *data mining* berjalan interaktif karena tidak jarang hasil *data mining* pada awalnya tidak sesuai dengan harapan analisnya sehingga perlu dilakukan desain ulang prosesnya (Kusnawi, 2007).

Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan *knowledge base*.



Gambar 1. Tahapan *Data Mining* (Han dan Kamber, 2006)

Keterangan:

1. Pembersihan Data *(Data Cleaning)*

Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isianisian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa *data mining* yang dimiliki.

Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik *data mining* karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi Data (*Data Integration*)

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Tidak jarang data yang diperlukan untuk *data mining*

tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau *file* teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya.

Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. Seleksi Data (*Data Selection*)

Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.

Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang membeli dalam kasus market basket analisis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

4. Transformasi Data (*Data Transformation*)

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan.

Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.

5. Proses *Mining*

Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi Pola (*Pattern Evaluation*)

Untuk mengidentifikasi pola-pola menarik kedalam *knowledge base* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses *data mining*, mencoba metode *data mining* lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses *data mining* adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami *data mining*. Karenanya presentasi hasil *data mining* dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses *data mining*. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil *data mining* (Han dan Kamber, 2006)

2.2.2. Klasifikasi

Klasifikasi adalah suatu *fungsionalitas data mining* yang menghasilkan model untuk memprediksi kelas atau kategori dari objek - objek didalam basis data. Klasifikasi merupakan proses yang terdiri dari dua tahap, yaitu tahap pembelajaran dan tahap pengklasifikasian.

Pada tahap pembelajaran, sebuah algoritma klasifikasi akan membangun sebuah model klasifikasi dengan cara menganalisis training data. Tahap pembelajaran dapat juga dipandang sebagai tahap pembentuakan fungsi atau pemetaan Y=F(X) dimana Y adalah kelas hasil prediksi dan X adalah *tuple* yang ingin diprediksi kelasnya. Selanjutnya pada tahap pengklasifikasian, model yang telah dihasilkan akan digunakan untuk melakukan pengklasifikasian.

Menurut Herman Aldino, Naam, Julfriadif (2012), klasifikasi adalah proses pencarian sekumpulan model yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu obyek yang belum diketahui kelasnya.

2.2.3. *Decision Tree* (Pohon Keputusan)

Pohon *(tree)* adalah sebuah struktur data yang terdiri dari simpul *(node)* dan rusuk *(edge).* Simpul pada sebuah pohon dibedakan menjadi tiga, yaitu simpul akar *(root node),* simpul percabangan/internal *(branch/ internal node)* dan simpul daun *(leaf node),* (Hermawati, 2013).

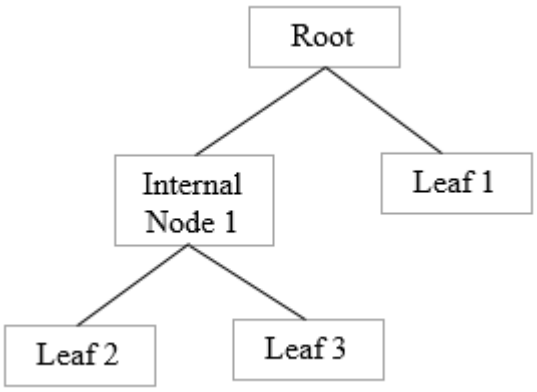
Pohon keputusan merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, dimana simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknya diberi label nilai atribut yang

mungkin dan simpul daun ditandai dengan kelas-kelas yang berbeda (Hermawati, 2013).

Gambar 2. Konsep Pohon Keputusan

Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule*. Manfaat utama dari penggunaan pohon keputusan adalah kemampuannya untuk *membreak down* proses pengambilan keputusan yang *kompleks* menjadi lebih simpel sehingga pengambil keputusan akan lebih menginterpretasikan solusi dari permasalahan. Pohon Keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.

Pohon keputusan merupakan himpunan aturan *IF...THEN*. Setiap *path* dalam *tree* dihubungkan dengan sebuah aturan, di mana premis terdiri atas sekumpulan *node-node* yang ditemui, dan kesimpulan dari aturam terdiri atas kelas yang terhubung dengan *leaf* dari *path* (Wibowo, 2011).



Gambar 3. Konsep Dasar Pohon Keputusan

Bagian awal dari pohon keputusan ini adalah titik akar *(root)*, sedangkan setiap cabang dari pohon keputusan merupakan pembagian berdasarkan hasil uji, dan titik akhir *(leaf)* merupakan pembagian kelas yang dihasilkan.

Pohon keputusan banyak mengalami perkembangan, beberapa *algoritma* yang populer dan sering dipakai adalah ID3, C4.5 dan *CART*.

Tabel 1. Frekuensi Penggunaan Algoritma Pohon Keputusan

Algoritma Pohon Keputusan	Frekuensi
ID3	68 %
C4.5	54.55 %
CART	40.9 %
SPRINT	31.84 %
SLIQ	27.27 %
PUBLIC	13.6 %
C5.0	9 %
CLS	9 %
RANDOM FOREST	9 %
RANDOM *TREE*	4.5 %
ID3+	4.5 %
OCI	4.5 %
CLOUDS	4.5 %

2.2.4. Algoritma C4.5

Menurut Luthfi (2009), algoritma C4.5 adalah algoritma klasifikasi data dengan teknik pohon keputusan yang memiliki kelebihan-kelebihan. Kelebihan ini misalnya dapat mengolah data numerik *(kontinyu)* dan *diskret*, dapat menangani nilai atribut yang hilang, menghasilkan aturan - aturan yang mudah diintrepetasikan dan tercepat diantara algoritma-algoritma yang lain.

Keakuratan prediksi yaitu kemampuan model untuk dapat memprediksi label kelas terhadap data baru atau yang belum diketahui sebelumnya dengan baik. Dalam hal kecepatan atau efisiensi waktu komputasi yang diperlukan untuk membuat dan menggunakan model. Kemampuan model untuk memprediksi dengan benar walaupun data ada nilai dari atribut yang hilang. Dan juga skalabilitas yaitu kemampuan untuk membangun model secara *efisien* untuk data berjumlah besar (aspek ini akan mendapatkan penekanan). Terakhir *interpretabilitas* yaitu model yang dihasilkan mudah dipahami.

Dalam algoritma C4.5 untuk membangun pohon keputusan hal pertama yang dilakukan yaitu memilih atribut sebagai akar. Kemudian dibuat cabang untuk tiap-tiap nilai didalam akar tersebut. Langkah berikutnya yaitu membagi kasus dalam cabang. Kemudian ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Menurut Luthfi (2009), untuk memilih atribut dengan akar, didasarkan pada nilai *gain* tertinggi dari atribut - tribut yang ada. Untuk menghitung gain digunakan rumus seperti tertera dalam persamaan 1 berikut:

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

Keterangan:

S : Himpunan kasus
A : Atribut
N : Jumlah partisi atribut A
| Si | : Jumlah kasus pada partisi ke-i
| S | : Jumlah kasus dalam S

Sehingga akan diperoleh nilai gain dari atribut yang paling tertinggi. Gain adalah salah satu *atribute selection measure* yang digunakan untuk memilih *test atribute* tiap *node* pada *tree*. Atribut dengan *information gain* tertinggi dipilih sebagai *test atribute* dari suatu *node*.

Sementara itu, penghitungan nilai entropi dapat dilihat pada persamaan 2.

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

Keterangan:

S : Himpunan kasus
A : Atribut
N : Jumlah partisi S
Pi : Proporsi dari Si terhadap S