BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data Mining didefinisikan sebagai sebuah proses untuk menemukan hubungan, pola dan tren baru yang bermakna dengan menyaring data yang sangat besar, yang tersimpan dalam penyimpanan, menggunakan teknik pengenalan pola seperti teknik Statistik dan Matematika (Larose, 2005).

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dahulu. Berawal dari beberapa disiplin ilmu, *data mining* bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani:

1. Jumlah data yang sangat besar
2. Dimensi data yang tinggi
3. Data yang heterogen dan berbeda sifat

Menurut para ahli, *data mining* merupakan sebuah analisa dari observasi data dalam jumlah besar untuk menemukan hubungan yang tidak diketahui sebelumnya dan dua metode baru untuk meringkas data agar mudah dipahami serta kegunaannya untuk pemilih data (David Hand, 2001).

2.2 Pengolahan *Data Mining*

Pengolahan *data mining* terdiri dari beberapa metode pengolahan, yaitu (Larose, 2005):

(a) *Predictive modelling* yang merupakan pengolahan *data mining* dengan melakukan prediksi/peramalan. Tujuan metode ini untuk membangun model prediksi suatu nilai yang mempunyai ciri-ciri tertentu. Contoh algoritmanya *Linear Regression, Neural Network, Support Vector Machine*, dan lain-lain.

(b) *Association* (Asosiasi) merupakan teknik dalam *data mining* yang mempelajari hubungan antar data. Contoh penggunaannya seperti untuk menganalisis perilaku mahasiswa yang datang terlambat. Contohnya jika mahasiswa memiliki jadwal dengan dosen A dan B, maka mahasiswa akan datang terlambat. Contoh algoritmanya FP-*Growth, A Priori,* dan lain-lain.

(c) *Clustering* (Klastering) atau pengelompokkan merupakan teknik untuk mengelompokkan data ke dalam suatu kelompok tertentu. Contoh algoritmanya *K-Means, K-Medoids, Self-Organitation Map (SOM), Fuzzy C-Means*, dan lain-lain. Contoh untuk *clustering*: Terdapat lima pulau di Indonesia: Sumatera, Kalimantan, Jawa, Sulawesi dan Papua. Maka lima pulau tersebut dijadikan tiga klaster berdasarkan waktunya: Waktu Indonesia Barat (Sumatera, Kalimantan dan Jawa), Waktu Indonesia Tengah (Sulawesi) dan Waktu Indonesia Timur (Papua).

(d) *Classification* merupakan teknik mengklasifikasikan data. Perbedaannya dengan metode *clustering* terletak pada data, dimana pada *clustering* variabel dependen tidak ada, sedangkan pada *classification* diharuskan ada variabel dependen. Contoh algoritma yang menggunakan metode ini ID3 dan *K Nearest Neighbors*.

2.3 Pohon Keputusan (*Decision Tree*)

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami, juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel *input* dengan sebuah variabel target. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, pohon keputusan sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain (Kusrini, 2009).

2.4 Algoritma C4.5

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut (Jefri, 2013):

a. Pilih atribut sebagai akar
b. Buat cabang untuk masing-masing nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus seperti yang tertera berikut:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(Si)$$ ... Rumus 2.1

Gambar 2.1 Rumus *Gain* (sumber: Jefri, 2013)

Keterangan:

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

Sebelum mendapatkan nilai *Gain* adalah dengan mencari nilai Entropi. Entropi digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan sebuah atribut. Rumus dasar dari Entropi adalah sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$ … Rumus 2.2

Gambar 2.2 Rumus *Entropy* Total (sumber: Jefri, 2013)

Keterangan:

S : Himpunan Kasus

n : Jumlah partisi S

pi: Proporsi dari Si terhadap S