

BAB II

LANDASAN TEORI

2.1 Sistem

Sistem mengandung arti kumpulan-kumpulan dari komponen-komponen yang memiliki unsur keterkaitan antara satu dengan lainnya (Indrajit, 2001). Kumpulan-kumpulan dari elemen-elemen tersebut saling berinteraksi untuk mencapai suatu tujuan tertentu. Sistem ini menggambarkan suatu kejadian-kejadian dan kesatuan yang nyata adalah suatu objek nyata, seperti tempat, benda dan orang-orang yang betul-betul ada dan terjadi (Jogianto, 2005).

Jogianto (2005) juga mengemukakan karakteristik atau sifat-sifat tertentu yang dimiliki sebuah sistem, yakni:

a. **Komponen**

Suatu sistem terdiri dari sejumlah komponen yang saling berinteraksi, yang artinya saling bekerja sama membentuk satu kesatuan. komponen-komponen sistem atau elemen-elemen sistem dapat berupa suatu subsistem atau bagian-bagian dari sistem. setiap subsistem mempunyai sifat-sifat dari sistem untuk menjalankan suatu fungsi tertentu mempengaruhi proses sistem secara keseluruhan.

b. **Batasan sistem**

Batasan sistem (boundary) merupakan daerah yang membatasi antara suatu sistem dengan sistem yang lainnya atau dengan lingkungan luarnya. batasan suatu sistem menunjukkan ruang lingkup dari sistem tersebut.

c. **Lingkungan Luar Sistem**

Lingkungan luar (environment) dari suatu sistem adalah apapun diluar batas sistem yang mempengaruhi operasi. Lingkungan luar sistem dapat bersifat menguntungkan sistem tersebut. Lingkungan luar yang menguntungkan berupa energi dari sistem dan dengan demikian harus tetap dijaga dan dipelihara. sedang lingkungan luar yang merugikan harus ditahan dan dikendalikan, kalau tidak maka akan mengganggu kelangsungan hidup dari sistem.

d. Penghubung Sistem

Penghubung (interfance) merupakan media penghubung antara satu subsistem dengan subsistem yang lainya. melalui penghubung ini memungkinkan sumber-sumber daya mengalir dari satu subsistem ke subsistem yang lainya. Dengan penghubung satu subsistem dapat berintegrasi dengan subsistem yang lainya membentuk satu kesatuan.

2.2 Pengelompokan

Pengelompokan adalah suatu tindakan analisis dari teknik *data mining* yang bertujuan untuk mengidentifikasi sekelompok objek yang memiliki kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok objek lainnya. Objek yang berada dalam kelompok yang sama relatif lebih homogen dari pada objek yang berada pada kelompok yang berbeda (Tan, 2006). Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data objek yang digunakan.

2.2.1 Tujuan Pengelompokan

Tujuan dari pengelompokan adalah mengumpulkan objek data ke dalam beberapa kelompok yang memiliki karakteristik tertentu dan dapat dibedakan satu sama lainnya. Hal tersebut dilakukan untuk analisis lebih lanjut sesuai dengan tujuan penelitian yang dilakukan.

Beberapa contoh tujuan yang dapat dilakukan dalam melakukan pengelompokan untuk penggunaan data sebagai berikut:

a. Peringkasan

Ada banyak teknik analisis data, salah satunya adalah melakukan peringkasan. Peringkasan merupakan cara untuk meningkatkan efesiensi waktu dalam melakukan pengelompokan data dalam jumlah yang banyak. Teknik pengelompokan yang dapat dilakukan dengan membuat *prototype* yang dapat mewakili seluruh data, misalnya dengan mengambil rata-rata untuk setiap kelompok sehingga

sejumlah data yang tergabung dalam sebuah kelompok akan diwakili oleh sebuah data.

b. Kompresi

Data-data yang tergabung dalam setiap kelompok dapat dianggap berkarakter sama sehingga data dalam kelompok yang sama dapat di kompresi dengan diwakili oleh indeks *prototype* dari setiap kelompok. Setiap objek direpresentasikan dengan indeks *prototype* yang dikaitkan dengan sebuah kelompok.

c. Pencarian tetangga terdekat secara efisien

Teknik komputasi yang dilakukan dalam mencari tetangga terdekat akan semakin berat ketika jumlah data semakin besar. Hal tersebut tidak sebanding dengan jumlah data yang akhirnya digunakan sebagai tetangga terdekat. Dengan pengelompokan kita dapat membuat *prototype* di mana setiap *prototype* mewakili satu kelas.

Cara ini dapat menghemat efisiensi waktu yang dilakukan dalam komputasi secara keseluruhan. Konsekuensinya adalah mungkin cara mendapatkan tetangga terdekat menjadi kurang representatif karena diukur berdasarkan kelompok data bukan individu data sehingga ada kemungkinan ketidaksesuaian tetangga pada data tersebut.

2.2.2 Kegiatan Sistem Pengelompokan

Berdasarkan uraian yang telah dijelaskan di atas dapat diambil sebuah pengelompokan dari sebuah sub-sub sistem pengelompokan yang terdiri dari:

a. Input

Merupakan suatu kegiatan untuk menyediakan serta memasukkan data untuk diproses.

b. Proses

Menggambarkan bagaimana suatu data diproses agar dapat menghasilkan suatu informasi yang bernilai lebih.

c. *Output*

Suatu kegiatan untuk menghasilkan laporan dari proses yang telah dijalankan.

d. Analisis Hasil Proses

Suatu kegiatan untuk menganalisa hasil keluaran dari kegiatan proses dan aktivitas untuk menjamin bahwa sistem Pengelompokan tersebut berjalan sesuai dengan yang diharapkan.

2.3 *Data Mining*

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban dkk. 2005).

Data mining adalah salah satu teknik penelusuran data untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat menggali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga memanfaatkan teknik ini. Dalam *data mining*, pengelompokan data juga dilakukan. Tujuannya dapat mengetahui pola dan tindak lanjut yang diambil. Semua hal tersebut bertujuan untuk mendukung kegiatan pengelompokan agar sesuai dengan yang diharapkan (Prasetyo, 2012).

Terdapat beberapa istilah lain yang memiliki makna sama dengan *data mining*, yaitu *Knowledge discovery in databases* (KDD), ekstraksi pengetahuan (*knowledge extraction*), Analisa data/pola (*data/pattern analysis*), kecerdasan bisnis (*business intelligence*) dan *data archaeology* dan *data dredging* (Larose, 2005).

Dari beberapa definisi tersebut dapat diambil kesimpulan bahwa *Data Mining* adalah proses menemukan sesuatu yang bermakna dari suatu kolerasi baru, pola dan tren yang ada dengan cara memilah-milah data berukuran besar yang disimpan dalam repositori menggunakan teknologi pengenalan pola serta teknik matematika dan statistik.

2.3.1 Fungsi *Data Mining*

Data mining mempunyai fungsi yang penting untuk membantu mendapatkan informasi yang berguna serta meningkatkan pengetahuan bagi pengguna. Pada dasarnya, *data mining* mempunyai empat fungsi dasar, diantaranya:

- a. Fungsi Prediksi (*prediction*). Proses untuk menemukan pola dari data dengan menggunakan beberapa variabel untuk memprediksikan variabel lain yang tidak diketahui jenis atau nilainya.
- b. Fungsi Deskripsi (*description*). Proses untuk menemukan suatu karakteristik penting dari data dalam suatu basis data.
- c. Fungsi Klasifikasi (*classification*). Klasifikasi merupakan suatu proses untuk menemukan model atau fungsi untuk menggambarkan *class* atau konsep dari suatu data. Proses yang digunakan untuk mendeskripsikan data yang penting serta dapat meramalkan kecenderungan data pada masa depan.
- d. Fungsi Asosiasi (*association*). Proses ini digunakan untuk menemukan suatu hubungan yang terdapat pada nilai atribut dari sekumpulan data.

2.3.2 Proses *Data Mining*

Proses yang umumnya dilakukan oleh data mining antara lain: deskripsi, prediksi, estimasi, klasifikasi, clustering dan asosiasi. Secara rinci proses data mining dijelaskan sebagai berikut (Larose, 2005):

a. Deskripsi

Deskripsi bertujuan untuk mengidentifikasi pola yang muncul secara berulang pada suatu data dan mengubah pola tersebut menjadi aturan dan kriteria yang dapat mudah dimengerti oleh para ahli pada domain aplikasinya. Aturan yang dihasilkan harus mudah dimengerti agar dapat dengan efektif meningkatkan tingkat pengetahuan (knowledge) pada sistem. Tugas deskriptif merupakan tugas data mining yang sering dibutuhkan pada teknik

postprocessing untuk melakukan validasi dan menjelaskan hasil dari proses data mining. Postprocessing merupakan proses yang digunakan untuk memastikan hanya hasil yang valid dan berguna yang dapat digunakan oleh pihak yang berkepentingan.

b. Prediksi

Prediksi memiliki kemiripan dengan klasifikasi, akan tetapi data diklasifikasikan berdasarkan perilaku atau nilai yang diperkirakan pada masa yang akan datang. Contoh dari tugas prediksi misalnya untuk memprediksikan adanya pengurangan jumlah pelanggan dalam waktu dekat dan prediksi harga saham dalam tiga bulan yang akan datang.

c. Estimasi

Estimasi hampir sama dengan prediksi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi.

d. Klasifikasi

Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang mendeskripsikan dan membedakan data ke dalam kelas-kelas. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya.

e. Clustering

Clustering merupakan pengelompokan data tanpa berdasarkan kelas data tertentu ke dalam kelas objek yang sama. Sebuah kluster adalah

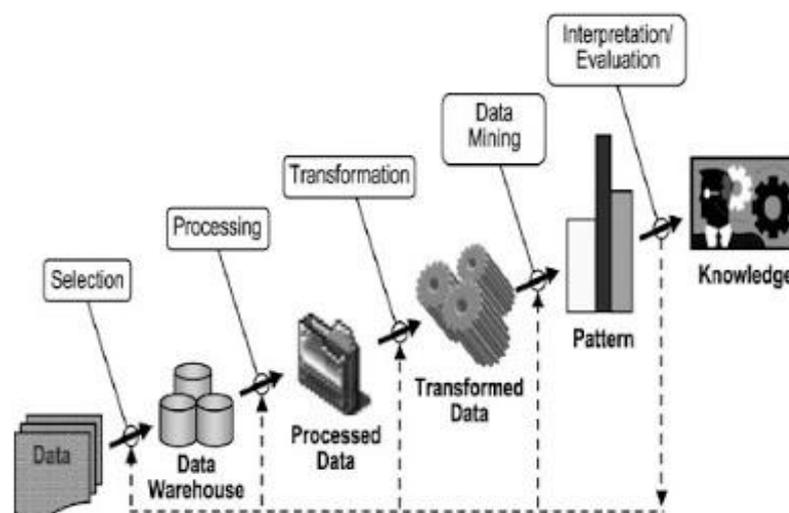
kumpulan record yang memiliki kemiripan suatu dengan yang lainnya dan memiliki ketidakmiripan dengan record dalam kluster lain. Tujuannya adalah untuk menghasilkan pengelompokan objek yang mirip satu sama lain dalam kelompok-kelompok. Semakin besar kemiripan objek dalam suatu cluster dan semakin besar perbedaan tiap cluster maka kualitas analisis cluster semakin baik.

f. Asosiasi

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (market basket analisis). Tugas asosiasi berusaha untuk mengungkap aturan untuk mengukur hubungan antara dua atau lebih atribut.

2.3.3 Tahapan *Data Mining*

Tahapan yang dilakukan pada proses *data mining* diawali dari seleksi data dari data sumber ke data target, tahap *preprocessing* untuk memperbaiki kualitas data, transformasi, *data mining* serta tahap interpretasi dan evaluasi yang menghasilkan *output* berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik. Secara detail dijelaskan pada gambar 2.1 sebagai berikut:



Gambar 2.1 Tahapan *Data Mining*

a. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *data warehouse* dimulai. Data hasil seleksi yang digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

b. Pre-processing / Cleaning

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

c. Transformation

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

d. Data mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

e. Interpretation / Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.3.4 *Clustering data*

Clustering atau analisis *cluster* adalah proses pengelompokan satu set benda-benda fisik atau abstrak ke dalam kelas objek yang sama (Han and Kamber, 2006).

Clustering atau *clusterisasi* adalah salah satu alat bantu pada *data mining* yang bertujuan mengelompokkan objek-objek ke dalam *cluster-cluster*. *Cluster* adalah sekelompok atau sekumpulan objek-objek data yang *similar* satu sama lain dalam *cluster* yang sama dan *dissimilar* terhadap objek-objek yang berbeda *cluster*. Objek akan dikelompokkan ke dalam satu atau lebih *cluster* sehingga objek-objek yang berada dalam satu *cluster* akan mempunyai kesamaan yang tinggi antara satu dengan lainnya.

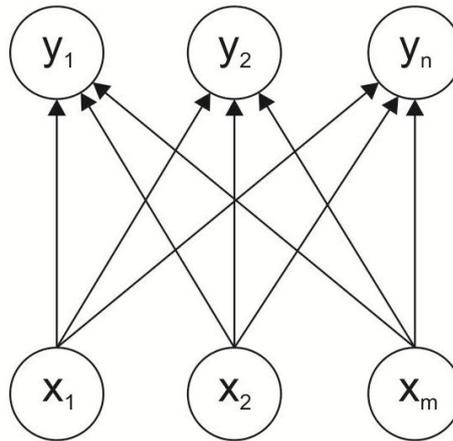
Dengan menggunakan *clusterisasi*, kita dapat mengidentifikasi tingkat potensi bahaya yang mungkin terjadi pada karyawan berdasarkan laporan kejadian hampir celaka dan kondisi tidak selamat yang ditemukan di area kerja.

2.4 *Self Organizing Map (SOM)*

Self Organizing Map (SOM) adalah salah satu tool yang sangat baik dalam penanganan data yang sangat besar (*data-mining*). SOM merupakan suatu metode jaringan syaraf tiruan yang diperkenalkan oleh Prof. Teuvo Kohonen seorang ilmuwan Finlandia pada tahun 1982. SOM adalah metode terkemuka pendekatan jaringan syaraf tiruan untuk *clustering* setelah *competitive learning*. SOM merupakan salah satu bentuk topologi dari *Unsupervised Artificial Neural Network (Unsupervised ANN)* dimana dalam proses pelatihannya tidak memerlukan pengawasan (target output). SOM digunakan untuk mengelompokkan (*clustering*) data berdasarkan karakteristik/ fitur-fitur data. Ide dasar metode SOM diilhami dari bagaimana proses otak manusia menyimpan gambar/pola yang telah dikenali melalui mata, kemudian mampu mengungkapkan kembali gambar/pola tersebut.

SOM memperlihatkan tiga karakteristik: kompetisi yaitu setiap vektor bobot saling berlomba untuk menjadi simpul pemenang, kooperasi yaitu setiap simpul pemenang bekerjasama dengan lingkungannya dan adaptasi yaitu perubahan warna

simpul pemenang. Dalam nilai bobot terdapat pola nilai input yang dikumpulkan dalam cluster. Selama proses SOM, unit cluster yang mempunyai nilai bobot akan dicocokkan dengan pola input yang terdekat dan dipilih sebagai pemenang. Adapun arsitektur SOM dapat dilihat pada gambar 2.2 berikut:



Gambar 2.2 Arsitektur SOM

2.5 Algoritma SOM

Berikut ini adalah langkah-langkah yang perlu dilakukan dalam menerapkan metode SOM dalam pengolahan data (Prasetyo, 2012):

1. Inialisasi nilai bobot awal (W_{ij}) secara acak, tentukan parameter laju pembelajaran dan jumlah iterasi yang digunakan dalam melakukan perhitungan.
2. Selama jumlah iterasi maksimal belum tercapai, lakukan pembaruan bobot nilai pada langkah 3-6 sampai tidak ada perubahan bobot pada iterasi selanjutnya.
3. Menghitung jarak ke setiap neuron keluaran dengan menggunakan *euclidean* kuadrat pada persamaan berikut:

$$d_{ij} = \| \mathbf{W}_i - \mathbf{X}_j \|_2 = \sqrt{\sum (\mathbf{W}_i - \mathbf{X}_j)^2} \dots\dots\dots (2.1)$$

4. Mencari indeks dari sejumlah neuron, yaitu D_j yang mempunyai nilai terkecil sebagai neuron pemenang.

5. Menghitung perubahan bobot terhadap neuron terkecil atau pemenang menggunakan persamaan berikut:

$$W_{ij}(\text{baru}) = W_{ij}(\text{lama}) + \dot{\eta} (X_i - W_{ij}(\text{lama})) \dots\dots\dots (2.2)$$

6. Perbarui nilai laju pembelajaran setiap pergantian iterasi menggunakan persamaan berikut:

$$\dot{\eta}(\text{baru}) = \text{fungsi_pembelajaran} \times \dot{\eta}(\text{lama}) \dots\dots\dots (2.3)$$

Algoritma SOM di atas menjelaskan inisialisasi bobot awal bisa menggunakan nilai acak dengan jangkauan -0.5 sampai +0.5 atau menggunakan nilai acak dengan jangkauan nilai seperti pada data masukan. Parameter jarak untuk perbedaan atau kemiripan yang digunakan adalah *Euclidean* kuadrat (*Square Euclidean*). Hal ini dimaksudkan untuk mengurangi waktu komputasi dengan menyederhanakan kinerja algoritma. Neuron terkecil akan terpilih sebagai neuron pemenang dan dilakukan perhitungan persamaan untuk mengubah bobot nilainya. Nilai laju pembelajaran yang digunakan adalah jangkauan nilai 0 sampai 1. Tetapi, nilai ini akan terus diturunkan setiap kali ada kenaikan iterasi dengan sebuah fungsi pembelajaran.

2.6 *Davies Bouldin Index*

Davies-Bouldin Index (DBI) merupakan salah satu metode validasi *cluster* untuk evaluasi kuantitatif dari hasil *clustering*. Pengukuran ini bertujuan memaksimalkan jarak *intercluster* antara satu cluster dengan cluster yang lain. Dalam penelitian ini DBI akan digunakan untuk mendeteksi *outlier* pada masing-masing *cluster* yang terbentuk.

Pendekatan perhitungan validasi DBI meliputi beberapa persamaan, *Sum of Square Between cluster* (SSB) merupakan persamaan yang digunakan untuk mengetahui separasi antar *cluster* yang dihitung menggunakan persamaan:

$$SSB_{ij} = \| C_i - C_j \|^2 \dots\dots\dots (2.4)$$

Sum of Square Within cluster (SSW) merupakan persamaan yang digunakan untuk mengetahui matrik kohesi dalam sebuah *cluster* ke-*i* yang dirumuskan sebagai berikut:

$$SSW_i = \frac{1}{M_i} \sum_{j=1}^{M_i} d(X_j, C_{pi}) \dots \dots \dots (2.5)$$

Setelah nilai kohesi dan separasi diperoleh, kemudian dilakukan pengukuran rasio (R_{ij}) untuk mengetahui nilai perbandingan antara cluster ke-*i* dan cluster ke-*j*. Cluster yang baik adalah cluster yang memiliki nilai kohesi sekecil mungkin dan separasi yang sebesar mungkin. Nilai rasio dihitung menggunakan persamaan berikut:

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \dots \dots \dots (2.6)$$

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai *Davies-Bouldin Index* (DBI) dari persamaan berikut:

$$DBI = \frac{1}{N} \sum_{i=1}^N \max_{i \neq j} (R_{i,j}) \dots \dots \dots (2.7)$$

Dari persamaan tersebut, semakin kecil nilai DBI yang diperoleh (non-negatif ≥ 0), maka *cluster* yang diperoleh semakin bagus/ *valid*.