



### **Hak cipta dan penggunaan kembali:**

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

### **Copyright and reuse:**

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

## BAB II

### TINJAUAN PUSTAKA

#### 2.1 Data Mining

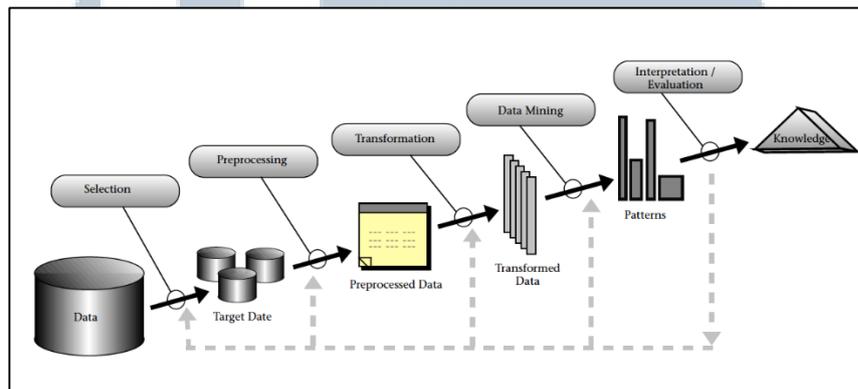
*Data mining* adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode *data mining* ini dapat digunakan untuk mengambil keputusan di masa depan. *Data mining* ini juga dikenal dengan istilah *pattern recognition* (Santosa, 2007).

Menurut Fayyad dkk. (1996), *data mining* juga merupakan tahap analisis dalam proses *Knowledge Discovery in Databases* yang merupakan proses mencari dan menggali pola baru dari sebuah data kumpulan yang besar dan tidak terstruktur dan melibatkan metode-metode dari ilmu statistika, intelegensia buatan, dan juga manajemen *database*.

Tugas utama *data mining* adalah melakukan analisis terhadap sejumlah kumpulan data dengan tujuan untuk mencari pola yang menarik dan penyajian kepada *end-user* melalui penggambaran secara visual yang sebelumnya tidak diketahui sebagai sebuah potensi lainnya, seperti dependensi (*association rule mining*), mencari data yang tidak umum (*anomaly detection*), dan pengelompokan data (*cluster analysis*) (Fayyad dkk., 1996). Pola ini kemudian dipelajari sebagai suatu acuan kesimpulan dari *input*-an data dan digunakan dalam penelitian lebih lanjut. Hal yang termasuk dalam *data mining* diantaranya adalah pengumpulan data, persiapan data, interpretasi hasil, dan pelaporan.

*Data mining* merupakan metode pengolahan data berskala besar. Oleh karena itu, *data mining* ini memiliki peranan penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi. *Data mining* membantu perusahaan untuk mendapatkan

pola dari data-data yang tersimpan di dalam basis data perusahaan. Pengetahuan yang diperoleh tersebut akan menjadi acuan pedoman dalam mengambil tindakan bisnis sebagai upaya pemeliharaan dan peningkatan tingkat kompetitif bisnis perusahaan (Yusuf , 2007).



Gambar 2.1 Tahapan Penyusunan *Knowledge Discovery in Databases* (Sumber: Fayyad dkk, 1996)

Gambar 2.1 menunjukkan proses tahapan penyusunan *Knowledge Discovery in Databases*(KDD). Menurut Chandrawati (2010), tahapan penyusunan KDD didefinisikan sebagai berikut.

### 1. *Selection*

Proses pemilihan dan penyaringan data berdasarkan beberapa kriteria sesuai dengan kebutuhan peneliti, misalkan berdasarkan jenis barang.

### 2. *Pre-processing*

Mempersiapkan data dengan cara membersihkan data yang akan diproses serta menghapus informasi atau *field-field* yang tidak dibutuhkan, yang nantinya akan memperlambat kinerja proses *query*.

### 3. *Transformation*

Proses untuk melakukan transformasi data ke bentuk yang *navigable* dan *useable*.

#### 4. *Data mining*

Pada bagian ini, dipusatkan untuk mendapatkan pola-pola dari data. Secara umum, kajian *data mining* memiliki metode-metode yang biasa digunakan seperti, *clustering*, klasifikasi, regresi, seleksi variabel, dan market basket analysis.

#### 5. *Interpretation and evaluation*

Pada proses ini *pattern* atau pola-pola yang telah diidentifikasi oleh sistem kemudian diterjemahkan ke dalam bentuk *knowledge* yang lebih mudah dimengerti oleh *end-user* untuk mengambil keputusan.

### 2.2 Clustering

Menurut Fayyad dkk. (1996) bahwa *clustering* mengacu pada pengelompokan catatan, pengamatan, atau kasus ke dalam kelas yang serupa. Sebuah *cluster* adalah kumpulan dari catatan yang mirip satu dengan lain dan berbeda dengan catatan dalam kelompok lainnya. *Clustering* berbeda dengan klasifikasi, di dalamnya tidak ada variabel target untuk *clustering*. Tugas dari *clustering* tidak mencoba untuk mengklasifikasikan, memperkirakan, atau memprediksi nilai variabel sasaran. Namun sebaliknya, algoritma ini mencari segmen seluruh data yang ditetapkan menjadi subkelompok yang relatif homogen atau *cluster*, dimana kesamaan catatan dalam *cluster* dimaksimalkan dan kesamaan catatan luar *cluster* diminimalkan (Fayyad dkk., 1996).

*Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised*), maksudnya metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru (*teacher*) serta tidak memerlukan target keluaran (*output*). Dalam *data mining* ada 2 (dua) jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical*

*clustering* (Santosa, 2010). Analisis *cluster* menurut Han dkk. (2011) ialah metode yang dipakai untuk membagi rangkaian data menjadi beberapa kelompok berdasarkan kesamaan-kesamaan yang telah ditentukan sebelumnya. Menurut Gorunescu (2011) secara umum dapat dikatakan bahwa.

- a. Data dalam suatu *cluster* memiliki tingkat kesamaan yang tinggi.
- b. Data dalam *cluster* yang berbeda memiliki tingkat kesamaan yang rendah.

*Hierarchical clustering* adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua (Santoso, 2010). Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hirarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster* (Santoso, 2010).

*Non-hierarchical clustering* dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan. Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hirarki. Metode ini biasa disebut dengan *K-Means Clustering* (Santoso, 2010).

### **2.3 K-Means Clustering**

Menurut Agusta (2007), *k-means clustering* merupakan salah satu metode data *clustering non-hierarki* yang mengelompokkan data dalam bentuk satu atau lebih *cluster* atau kelompok. Data-data yang memiliki karakteristik yang sama dikelompokkan dengan *cluster* atau kelompok yang lain sehingga data yang berada dalam satu *cluster* atau kelompok memiliki tingkat variasi yang kecil (Agusta, 2007). Menurut Fayyad dkk. (1996) *k-means clustering* adalah algoritma sederhana

dan efektif untuk menemukan *cluster* di data. Menurut Fayyad dkk. (1996) langkah-langkah untuk algoritma *k-means clustering* sebagai berikut.

1. Pengguna diminta menentukan berapa banyak *cluster k* yang akan dibagi dalam sebuah *data set*. Nilai *k* dapat diambil menggunakan angka *random*.
2. Menetapkan nilai *k* secara acak untuk menjadi lokasi pusat *cluster* awal.
3. Untuk setiap *record*, menemukan pusat *cluster* terdekat. Dengan demikian, masing-masing pusat *cluster* memiliki *subset* dari *records*, sehingga mewakili partisi dari kumpulan data. Maka dihasilkan *k clusters* yaitu,  $C_1, C_2, \dots, C_k$ . Teori jarak Euclidean pada Rumus 2.1 digunakan untuk menghitung jarak semua data ke setiap titik pusat *cluster* (Santosa, 2007).

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \dots + (X_{ki} - X_{kj})^2}$$

... Rumus 2.1

$D(i,j)$  = Jarak data ke-*i* ke pusat *cluster j*

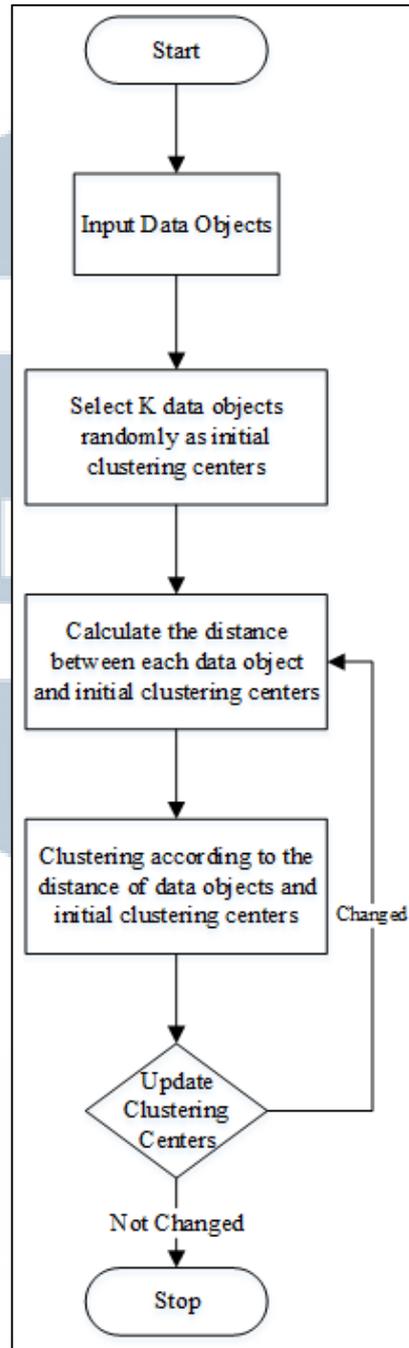
$X_{ki}$  = Data ke-*i* pada atribut data ke-*k*

$X_{kj}$  = Titik pusat ke-*j* pada atribut ke-*k*

4. Pada setiap *k clusters*, temukan *cluster centroid*, dan perbaharui lokasi setiap pusat *cluster* pada nilai baru dari setiap *centroid*.
5. Ulangi langkah ke-3 sampai 5 sampai pusat *cluster* tidak berubah lagi.

Diagram alur algoritma *k-means clustering* ditunjukkan pada Gambar 2.2.

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



Gambar 2.2 Diagram Alur *K-Means Clustering*  
(Sumber: Rahmani, 2014)

## 2.4 Sistem Rekomendasi

Sistem rekomendasi adalah suatu alat dan teknik yang menyediakan saran terkait pada sebuah hal untuk dapat dimanfaatkan oleh *user* (Ricci dkk., 2011).

Sistem rekomendasi diterapkan dengan tujuan-tujuan tertentu. Tujuan pertama

untuk meningkatkan jumlah penjualan barang ataupun jasa dengan memberikan penawaran produk lain yang diperkirakan akan menarik minat pengguna layanan. Tujuan lain adalah untuk menjual produk yang lebih beragam dengan menawarkan produk yang sulit ditemukan apabila tidak direkomendasikan. Sistem rekomendasi dapat pula meningkatkan kepuasan dan loyalitas pengguna yang memperoleh rekomendasi sesuai dengan minat. Selain itu, sistem rekomendasi memungkinkan penyedia layanan untuk lebih memahami keinginan pengguna (Ricci dkk., 2011).

