BAB II
LANDASAN TEORI

2.1 Data Mining

Faktor penentu bagi usaha atau bisnis apapun pada masa sekarang ini adalah kemampuan untuk menggunakan informasi seefektif mungkin. Penggunaan data secara tepat karena penemuan fakta-fakta yang sangat berharga yang cukup sering, tersembunyi dan tidak terdeteksi sebelumnya mengenai prilaku transaksi konsumen, *retailer* dan *supplier*, tren –tren bisnis, da faktor-faktor petunjuk yang lain (Berson, 1997). Menurut Kamber (2007) secara sederhana data mining mengacu kepada mengestrak atau "menambang" pengetahuan dari sekumpulan besar data. Menambang dalam hal ini bukan diibaratkan sebagai menambang emas atau tambang pasir, tetapi lebih diibaratkan sebagai "*knowledge mining from data*" atau lebih ringkasnya menambang pengetahuan. Pengertian lain data mining juga dapat berarti proses untuk memperkerjakan satu atau lebih teknik pembelajaran terkomputerisasi untuk mengotomasi analisa dan mengestrak pengetahuan dari data didalam database (Roger and Geatz, 2003).

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database. Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan, dan *mechine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, dkk. 2005).

Kemajuan luar biasa yang terus berlanjut dalam bidang data mining didorong oleh beberapa faktor, antara lain (Larose, 2005):

1. Pertumbuhan yang cepat dalam kumpulan data
2. Penyimpanan data dalam data warehouse, sehingga seluruh perusahaan dan memiliki akses kedalam database yang handal
3. Adanya peningkatan akses data melalui navigasi web dan internet

4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi
5. Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi)
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Dari defenisi-defenisi yang telah disampaikan, hal penting yang terkait dengan data mining adalah:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada
2. Data yang akan diproses berupa data yang sangat besar
3. Tujuan data mining adalah untuk mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermafaat

Tugas *data mining* secara garis besar dibagi menjadi dua kategori utama, yaitu (Tan dkk, 2006) :

1. Tugas prediktif.

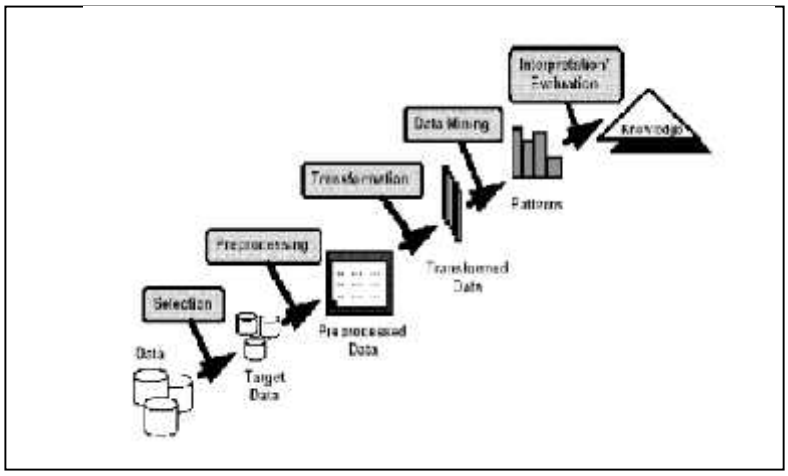
   Tujuan utama dari tugas ini adalah untuk memprediksikan nilai dari atribut tertentu berdasarkan nilai dari atribut lainnya. Atribut yang diprediksi dikenal sebagai target atau *dependent variable*, sedangkan atribut yang digunakan untuk membuat prediksi disebut penjelas atau *independent variable*.
2. Tugas deskriptif.

   Tujuan utama dari tugas ini adalah untuk memperoleh pola (*correlation, trend, cluster, trajectory, anomaly*) untuk menyimpulkan hubungan di dalam data. Tugas deskriptif merupakan tugas *data mining* yang sering dibutuhkan pada teknik *postprocessing* untuk melakukan validasi dan menjelaskan hasil proses *data mining*.

2.2 Knowledge Discovery In Database (KDD)

Knowledge Discovery In Databases (KDD) adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data.

Berikut merupakan Tahapan dari KDD: (Fayyad, 1996)



Gambar 2.1. Tahapan KDD

a. *Data Selection*
   - Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (*discovery*) akan dilakukan.
   - Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

b. *Pre-processing/ Cleaning*
   - Pemprosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan.

- 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.
- Dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain (eksternal)

c. *Transformation*
   - Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai.
   - Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

d. *Data mining*
   - Pemilihan tugas data mining, pemilihan *goal* dari proses KDD misalnya klasifikasi, regresi, clustering, dll.
   - Pemilihan algoritma data mining untuk pencarian (*searching*)
   - Proses Data mining yaitu 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*
   - Penerjemahan pola-pola yang dihasilkan dari data mining.
   - 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 mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

2.3 *Clustering*

Cluster adalah suatu kumpulan dari entitas yang hampir sama (Everit, 1993). Pengertian lain menurut Kamber (2007), cluster adalah kumpulan dari objek yang mirip dengan objek lainnya dan berada pada kelompok yang sama. Sedangkan proses untuk pengelompokkan data baik itu bersifat fisik atau abstrak kedalam suatu kelompok atau kelas yang memiliki kesamaan sifat disebut *clustering.*

Clustering dikategorikan kedalam teknik *Undirect Knowladge atau Unsupervised Learning* karena tidak membutuhkan prose pelatihan untuk klasifikasi awal data dalam masing-masing kelompok atau *cluster*. Tujuan utama *clustering* adalah untuk menentukan atau mencari pola yang bermanfaat atau yang berguna pada suatu database, kemudian merangkumnya dan membuat lebih mudah untuk dipahami.

Dalam menentukan proses analisa terhadap cluster-cluster yang telah terbentuk dan pencarian pengetahuan dengan metode tertentu disebut cluster analyse (Kamber, 2007).

Clustering adalah metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode *Data Mining*, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu ‘wilayah’ yang sama dan data dengan karakteristik yang berbeda ke ‘wilayah’ yang lain. Ada beberapa pendekatan yang digunakan dalam mengembangkan metode clustering. Dua pendekatan utama adalah clustering dengan pendekatan partisi dan clustering dengan pendekatan hirarki. Clustering dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada.

Clustering dengan pendekatan hirarki atau sering disebut dengan *hierarchical clustering* mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan. Di samping kedua pendekatan tersebut, ada juga clustering dengan pendekatan *automatic mapping* (Self-Organising Map/SOM).

2.4 K-Means

Metode K-means adalah metode pengelompokkan data dengan mengambil parameter sejumlah k cluster, dan mempartisi data kedalam cluster tersebut, dengan berpatokan pada kemiripan antara data dalam suatu cluster dan ketidakmiripan di antara cluster berbeda, pusat dari cluster adalah rata-rata dari nilai anggota cluster yang disebut dengan *centroid* atau *center of gravity* (Kamber, 2007). Selain itu K-means melakukan pengelompokkan dengan meminimalkan jumlah kuadrat dari jarak (*disance*) antara data dengan centroid cluster yang cocok (Teknomo, 2006).

Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya.

2.4.1 Algoritma K-means

Algoritma k-means dimulai dengan pembentukan prototipe cluster di awal kemudian secara iteratif prototipe cluster ini diperbaiki hingga konvergen (tidak terjadi perubahan yang signifikan pada prototipe cluster). Perubahan ini diukur menggunakan fungsi objektif J yang umumnya didefinisikan sebagai jumlah atau rata-rata jarak tiap item data dengan pusat massa kelompoknya. Secara lebih detil algoritma k-means adalah seperti berikut :

1. inisialisasi nilai J (misal MAXINT)
2. Tentukan prototipe cluster awal (bisa secara acak ataupun dipilih salah satu secara acak dari koleksi data)

3. Masukkan tiap satuan data ke dalam kelompok yang *jarak* dengan pusat massa-nya paling dekat
4. ubah nilai pusat massa tiap cluster sebagai rata-rata (mean) dari seluruh anggota kelompok tersebut
5. Hitung fungsi objektif J
6. jika nilai J sama dengan sebelumnya, berhenti atau ulangi langkah 3

2.4.2 *Distance Space*

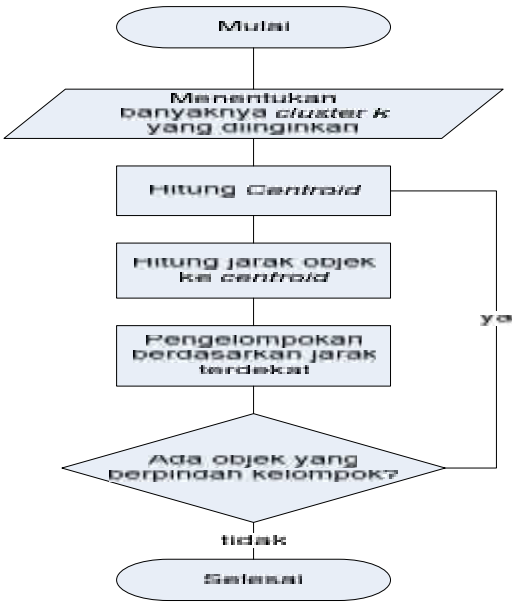
Beberapa *distance space* telah diimplementasikan dalam menghitung jarak (*distance* antara data dan *centroid* termasuk di antaranya L1 (*Manhattan/ City Block*) *distance space*, L2 (*Euclidean*) *distance space*, dan Lp (*Minkowski*) *distance space*. Jarak antara dua titik x1 dan x2 pada *Manhattan/City Block distance space* dihitung dengan menggunakan rumus sebagai berikut (Yudi Agusta, 2007)

$$D_{1}(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^{p} |x_{2j} - x_{1j}| \qquad (2.1)$$

Sedangkan untuk L2 (*Euclidean*) *distance space*, jarak antara dua titik dihitung menggunakan rumus sebagai berikut:

$$D_{l_2}(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} (x_{2j} - x_{1j})^2} \qquad (2.2)$$

Jadi, algoritma *K-Means* adalah algoritma partitional (*Non Hierarchical*) *clustering yang* mempartisi atau membagi sekumpulan data ke dalam sejumlah *cluster*. Setiap *cluster* mempunyai titik pusat *cluster/centroid. Centroid* adalah rata-rata (*mean*) dari setiap titik angota *cluster*. Untuk lebih mengetahui alur algoritma dari metode *K-Means*, dapat dilihat pada gambar berikut :



Gambar 2.11 Flowchart Algoritma *Clustering K-Means* (Teknomo, 2006)

Algoritma dasar dalam K-Means adalah :

1. Tentukan jumlah klaster (k), tetapkan pusat klaster sembarang.
2. Bangkitkan k centroid (titik pusat klaster) awal secara random.
3. Hitung jarak setiap data ke pusat klaster.
4. Kelompokkan data ke dalam klaster dengan jarak yang paling pendek.
5. Hitung pusat klaster yang baru dengan mencari nilai rata-rata dari data-data yang menjadi anggota pada klaster tersebut.
6. Ulangi langkah 3 jika sudah tidak ada lagi data yang berpindah ke klaster yang lain.

2.5 Klasifikasi *Fuzzy* RFM

Analisa RFM adalah proses mengeanalisis prilaku pelanggan. Hal ini umumnya digunakan dalam pemasaran database dan pemasaran langsung. RFM singkatan dari Recency, Frequency, Monetary. Menggunakan informasi tentang prilaku pelanggan pada masa lalu yang mudah dilacak dan tersedia. Tujuan dari RFM adalah untuk meramalkan prilaku konsumen di masa yang

depan (mengarahkan keputusan segmentasi yang lebih baik). Oleh karena itu, perlu menterjemahkan prilaku konsumen dalam ‘angka’ sehingga dapat digunakan sepanjang waktu.

Analisa RFM terdiri dari tiga dimensi, yaitu Recency, Frequency, Monetary.

1. *Recency*

   Recency adalah mengukur nilai pelanggan dengan melihat prilaku konsumen yang berkenaan dengan pembelian yang dilakukannya paling akhir. Informasi terpenting yang tidak boleh dilewatkan berkenaan dengan recency adalah tanggal pembelian terakhir yang merupakan barometer pengukuran recency.

2. *Frequency*

   Recency adalah mengukur nilai pelanggan dengan melihat prilaku konsumen yang berkenaan dengan aktifitas transaksi yang dilakukan oleh konsumen dalam satu periode. Satu periode yaitu dalam rentang waktu di tentukan, misalnya dalam 2 tahun berapakah rata2 transaksi yang dilakukan oleh konsumen.

3. *Monetary*

   Recency adalah mengukur nilai pelanggan dengan melihat prilaku konsumen yang berkenaan dengan rata-rata transaksi yang dilakukan oleh konsumen dalam satu kali bertransaksi.

Metode *sharp* RFM mendeskripsikan atribut *recency, frequency,* dan *monetary* dengan variabel linguistik. Konteks dari masing-masing atribut didefinisikan menggunakan tabel sebagai berikut :

Tabel 2.1 Rentang jarak variable linguistik

ATRIBUT	VARIABEL LINGUISTIK	DOMAIN NILAI
RECENCY	BARU SAJA	0 <= r < 14 Hari
	AGAK LAMA	42 < r < 56 Hari
	LAMA	56 Hari < r

ATRIBUT	VARIABEL LINGUISTIK	DOMAIN NILAI
FREQUENCY	JARANG	0 <= t < 10 Transaksi
	AGAK SERING	30 < t < 40 Transaksi
	SERING	60 Transaksi < r
MONETARY	RENDAH	0 <= m < 15 Juta Rupiah
	SEDANG	35 juta< m < 45 juta
	TINGGI	65 juta < mS

Tabel 2.2 Rentang jarak variabel recency

Recency	0	14	42	56	84	100
Baru saja	1	1	0	0	0	0
Agak lama	0	0	1	1	0	0
Lama	0	0	0	0	1	1

Tabel 2.3 Rentang jarak variabel frequency

Frequency	0	10	30	40	60	80
Jarang	1	1	0	0	0	0
Agak Jarang	0	0	1	1	0	0
Sering	0	0	0	0	1	1

Tabel 2.4 Rentang jarak variable monetary

Monetary	-	15.000.000	35.000.000	45.000.000	65.000.000	80.000.000
Rendah	1	1	0	0	0	0
Sedang	0	0	1	1	0	0
Tinggi	0	0	0	0	1	1

Pada Tabel 2.2 sampai 2.4 diberikan contoh nilai *recency*, *frequency*, dan *monetary* dari empat konsumen. Nilai yang diperoleh oleh masing-masing konsumen diberikan berdasarkan ketentuan pada Tabel 2.1.

Tabel 2.5 Contoh nilai *recency*, *frequency*, dan *monetary* dari empat konsumen

Customer	***Class***	***RFM attributes, (equivalence classes) and terms***					
		Recency		***Frequency***		***Monetary value***	
		Terakhir transaksi	***kelas***	***Rata-rata transaksi***	***kelas***	***Rata-rata sekali transaksi***	***kelas***
Riko	C3	378	Long ago	11	Frequent	92.000	Low value
Toni	C4	723	Long ago	7	Rare	46.000	Low value
Ari	C5	342	Very recent	13	Frequent	175.000	High value
Rian	C5	14	Very recent	38	Frequent	323.000	High value