

DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi | UNIKOM | 2016

Nizar Rabbi Radliya | nizar.radliya@yahoo.com

Nama Mahasiswa	
NIM	
Kelas	
Kompetensi Dasar	
Memahami definisi, proses serta teknik data mining.	
Pokok Bahasan	
Pengenalan Data Mining: 1. Definisi data mining 2. Proses data mining 3. Teknik data mining	

I. Definisi Data Mining

Munculnya data mining didasarkan pada jumlah data yang tersimpan dalam basis data semakin besar. Dari fenomena tersebut data mining berusaha menjawab dengan melakukan proses yang dapat menemukan suatu informasi baru yang berguna. Istilah data mining kadang disebut juga dengan *knowledge discovery*.

Pada perkuliahan ini data mining akan diimplementasikan pada data-data yang terkumpul di dalam sebuah sistem informasi. Data-data tersebut disimpan dengan menggunakan teknologi basis data berupa data. Basis data merupakan salah satu bidang dari bidang lainnya yang berkaitan dengan data mining. Bidang-bidang yang berkaitan dengan data mining sudah kita bahas pada materi sebelumnya.

Sudah banyak yang mendefinisikan data mining melalui buku yang ditulisnya, diantaranya:

Tan (2006) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan.

Darly Pregibon (2011) menyatakan bahwa data mining adalah campuran dari statistik, kecerdasan buatan, dan riset basis data.

Pramudiono (2006) mengartikan data mining sebagai serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

Definisi-definisi di atas mengenai data mining hanyalah sebagian dari banyaknya definisi yang ada. Akan tetapi pada intinya adalah data mining merupakan proses mendapatkan pola atau informasi (pengetahuan) baru dengan teknik tertentu (dengan melibatkan beberapa bidang lainnya) dari kumpulan data yang besar.

Salah satu teknik yang dibuat dalam data mining adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain, yang tidak/belum berada dalam basis data yang tersimpan. Kebutuhan untuk proses prediksi juga dapat menggunakan pola tersebut. Dalam data mining, pengelompokan data juga dapat dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil.

Jadi pengetahuan yang dihasilkan oleh data mining dapat kita manfaatkan dalam sistem informasi pengambilan keputusan guna menciptakan bisnis cerdas (*business intelligence*) pada sebuah organisasi. Selanjutnya kita akan melihat beberapa contoh yang masuk dalam kategori data mining dan bukan data mining.

1. Bukan data mining: Pencarian informasi tertentu di internet (misalnya mencari informasi masakan menggunakan google).

Termasuk data mining: Pengelompokan informasi yang mirip dalam konteks tertentu pada hasil pencarian (misalnya masakan sunda, masakan padang dan sebagainya yang ditemukan google).

2. Bukan data mining: Petugas medis mencari data medis untuk menganalisis catatan pasien dengan penyakit tertentu.

Termasuk data mining: Peneliti medis mencari cara pengelompokan data penyakit pasien berdasarkan data diagnosis, umur, alamat.

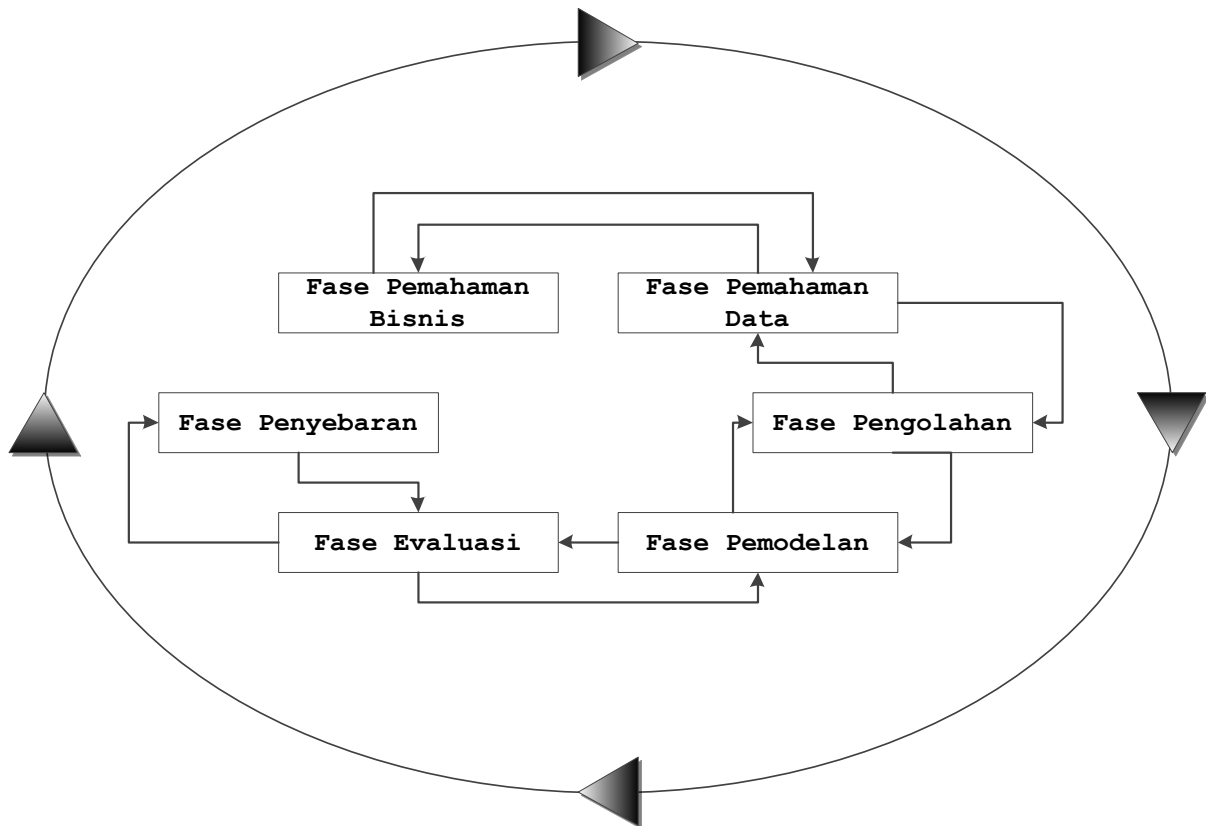
3. Bukan data mining: Pembuatan laporan tahunan penjualan perusahaan dengan merekap semua data selama setahun.

Termasuk data mining: Pemanfaatan data penjualan perusahaan untuk mendapatkan pola prediksi stok yang sebaiknya disediakan pada tahun berikutnya.

II. Proses Data Mining

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan standar yang dapat kita gunakan dalam melaksanakan proses data mining. CRIPS-DM

CRIPS-DM merupakan siklus hidup proyek data mining yang terbagi dalam enam fase seperti terlihat pada gambar 1 di bawah ini. Pengerjaan fase bersifat adaptif, dimana fase berikutnya bergantung pada keluaran fase sebelumnya. Fase yang sudah dikerjakan mungkin dapat kembali pada fase sebelumnya apabila diperlukan evaluasi atau kesalahan pada fase sebelumnya.



Berikut penjelasan dari setiap fase CRIPS-DM (Larose, 2005):

1. Fase Pemahaman Bisnis (*Business Understanding Phase*)
 - a. Penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan.
 - b. Menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining.
 - c. Menyiapkan strategi awal untuk mencapai tujuan.
2. Fase Pemahaman Data (*Data Understanding Phase*)
 - a. Mengumpulkan data.

- b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
 - c. Mengevaluasi kualitas data.
 - d. Jika diinginkan, pilih sebagian kecil group data yang mungkin mengandung pola dari permasalahan.
3. Fase Pengolahan Data (*Data Preparation Phase*)
- a. Siapkan dari data awal, kumpulan data yang akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif.
 - b. Pilih kasus dan variable yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan.
 - c. Lakukan perubahan pada beberapa variable jika dibutuhkan.
 - d. Siapkan data awal sehingga siap untuk perangkat pemodelan.
4. Fase Pemodelan (*Modeling Phase*)
- a. Pilih dan aplikasi teknik pemodelan yang sesuai.
 - b. Kalibrasi aturan model untuk mengoptimalkan hasil.
 - c. Perlu diperhatikan bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama.
 - d. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.
5. Fase Evaluasi (*Evaluation Phase*)
- a. Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan.
 - b. Menetapkan apakah terdapat model yang memenuhi tujuan pada fase awal.
 - c. Menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik.
 - d. Mengambil keputusan berkaitan dengan penggunaan hasil dari data mining.
6. Fase Penyebaran (*Deployment Phase*)
- a. Menggunakan model yang dihasilkan. Terbentuknya model tidak menandakan telah terselesaikannya proyek.
 - b. Contoh sederhana penyebaran: pembuatan laporan.

- c. Contoh kompleks penyebaran: penerapan proses data mining secara parallel pada departemen lain.

Informasi lebih lanjut mengenai CRIPS-DM dapat dilihat di www.crisp-dm.org. Selain CRIPS-DM terdapat langkah sistematis lainnya yang kurang lebih intinya sama. Berikut tiga langkah utama dalam proses data mining (Gonunescu, 2011):

1. Ekplorasi/pemrosesan awal data

Ekplorasi/pemrosesan awal data terdiri dari normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

2. Membangun model dan melakukan validasi terhadapnya

Membangun model dan melakukan validasi terhadapnya berarti melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis cluster, deteksi anomali, analisis asosiasi, analisis pola sekuensial, dan sebagainya. Dalam beberapa referensi, deteksi anomali juga masuk dalam langkah eksplorasi. Akan tetapi, deteksi anomali juga dapat digunakan sebagai algoritma utama, terutama untuk mencari data-data yang special.

3. Penerapan

Penerapan berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang diinvestigasi.

Penjelasan dari masing-masing langkah akan dijelaskan pada materi-materi berikutnya.

III. Teknik Data Mining

Teknik data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan diantaranya: *classification, clustering, association, anomaly, regression, sequential pattern, deviation detection*. Pada perkuliahan ini tidak semua teknik data mining akan dibahas karena didasari oleh keterbatasan waktu. Hanya empat teknik data mining yang akan dibahas yaitu diantaranya:

1. Klasifikasi (*classification*)

Klasifikasi merupakan salah satu teknik dari model prediksi. Teknik ini digunakan untuk pembuatan model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan variabel baru yang didapat.

Beberapa contoh masalah yang dapat diselesaikan dengan klasifikasi:

- a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
- b. Melakukan deteksi penyakit pasien berdasarkan sejumlah nilai parameter penyakit yang diderita.
- c. Memprediksi pelanggan mana yang akan berpindah ke kompetitor kita atau tetap setia.

2. Pengklusteran (*clustering*)

Teknik ini digunakan untuk melakukan pengelompokan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan karakteristik masing-masing data pada kelompok-kelompok yang ada. Data-data yang masuk dalam batas kesamaan dengan kelompoknya akan bergabung dalam kelompok tersebut, begitu juga sebaliknya.

Beberapa contoh masalah yang dapat diselesaikan dengan pengklusteran:

- a. Mengetahui pola pembelian barang oleh para konsumen pada waktu-waktu tertentu.
- b. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran.

3. Asosiasi (*association*)

Teknik ini digunakan untuk menemukan pola yang menggambarkan kekuatan hubungan fitur dalam data. Mendeteksi kumpulan atribut-atribut yang muncul bersamaan dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut. Teknik ini sering digunakan untuk melakukan analisis keranjang belanja.

Beberapa contoh masalah yang dapat diselesaikan dengan asosiasi:

- a. Untuk mengenali item-item yang dibeli bersama-sama oleh cukup banyak pelanggan.
- b. Untuk mengetahui suku cadang yang membutuhkan perbaikan setelah suku cadang yang lainnya mengalami kerusakan.

4. Anomali (*anomaly*)

Teknik ini berkaitan dengan pengamatan sebuah data dari sejumlah data yang secara signifikan mempunyai karakteristik yang berbeda dari sisa data yang lain. Algoritma deteksi anomali yang baik harus mempunyai laju eror yang rendah.

Beberapa contoh masalah yang dapat diselesaikan dengan anomali:

- a. Untuk mengetahui pola data yang memasuki jaringan sehingga penyusupan bisa ditemukan jika pola kerja data yang datang berbeda.
- b. Diteksi perilaku kondisi cuaca yang mengalami anomali.

IV. Daftar Pustaka

- [1] Astuti, F.A. 2013. Data Mining. Yogyakarta: Andi.
- [2] Kusriani & Taufiz, E.L. 2009. Algoritma Data Mining. Yogyakarta: Andi.
- [3] Prasetyo, E. 2012. Data Mining: Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: Andi.
- [4] Prasetyo, E. 2014. Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB. Yogyakarta: Andi.

V. Materi Berikutnya

Pokok Bahasan	Set Data
Sub Pokok Bahasan	1. Definisi dan tipe data 2. Kualitas data 3. Similaritas dan dissimilaritas