MSIM4403
Edisi 1

MODUL 01

Konsep Dasar Data Mining

Gede Aditra Pradnyana, S.Kom., M.Kom.
Dr. Ketut Agustini,S.Si., M.Si.

Daftar Isi

Pendahuluan

Dewasa ini, kita sedang hidup di era data (*data age*) dimana data dihasilkan dalam jumlah yang sangat banyak setiap harinya. Pernahkah kita berpikir, dalam sehari berapa data yang kita hasilkan? Contoh sederhana adalah data berupa pesan di *whatsapp* yang kita kirimkan atau terima, status di *facebook* yang kita unggah, data transaksi belanja di supermarket, data akademik di kampus, *email*, dan masih banyak lainnya. Apakah jumlah data yang dihasilkan sebanding dengan besar manfaat dari adanya data tersebut? Apakah data yang tersimpan akan selamanya dapat digunakan? Contoh data IPK seorang alumni di sistem informasi akademik sebuah fakultas, yang jumlahnya akan terus semakin bertambah. Apakah data tersebut akan berguna atau hanya menjadi "sampah" data? Dari fenomena ini muncullah ungkapan "*We are drowning in Data, but starving for Knowledge*". Data mining muncul sebagai bidang ilmu yang berusaha menggali pengetahuan tersembunyi dari sebuah data.

Setelah mempelajari modul ini Anda diharapkan mampu menjelaskan konsep dasar Data Mining. Secara lebih rinci Anda diharapkan mampu:

1. menjelaskan definisi data mining;
2. menjelaskan pentingnya data mining;
3. menjelaskan peranan data mining;
4. menjelaskan implementasi data mining;
5. menyebutkan bidang ilmu yang berkaitan dengan data mining;
6. menyebutkan prinsip dan proses data mining;
7. menjelaskan atribut data dalam data mining;

Untuk itu Anda diharapkan mengikuti petunjuk belajar sebagai berikut.

1. **Bacalah** bagian **uraian dan contoh** dari setiap Kegiatan Belajar dengan baik sampai dimengerti, dipahami, dan diterapkan.
2. **Kerjakan Latihan** dengan baik dan jujur, penuh kesungguhan dan tanggung jawab.
3. **Bacalah Rangkuman** yang disediakan untuk memberikan ringkasan pemahaman tentang Konsep Dasar Data Mining.
4. **Kerjakan Tes Formatif** yang disediakan untuk mengecek seberapa jauh Anda mencapai tujuan pembelajaran setiap kegiatan belajar tanpa melihat rambu-rambu jawaban yang tersedia.
5. **Bacalah Glosarium** yang disediakan untuk menyamakan persepsi tentang istilah yang dipakai dalam Konsep Dasar Data Mining.
6. **Jika** Anda merasa telah menjawab Tes Formatif dengan baik dan benar, bandingkanlah jawaban tersebut dengan kriteria yang tersedia. Apabila setelah dihitung ternyata Anda telah mencapai tingkat pengusaan minimal 80%, Anda dapat meneruskan ke Kegiatan Belajar Berikutnya.

Kegiatan Belajar 1

Definisi Data Mining

Istilah Data Mining sebenarnya mulai dikenal sejak tahun 1990, ketika pekerjaan pemanfaatan data menjadi sesuatu yang dianggap penting dalam berbagai bidang, mulai dari bidang akademik, bisnis, hingga bidang medis. Munculnya data mining didasarkan pada jumlah data yang tersimpan dalam basis data semakin besar. Perkembangan yang cepat dalam teknologi pengumpulan dan penyimpanan data telah memudahkan organisasi untuk mengumpulkan sejumlah data berukuran besar sehingga menghasilkan gunung data. Misalnya dalam sebuah universitas ada berapa data mahasiswa yang tersimpan dari tahun ke tahun. Kemudian data di sebuah supermarket, ada berapa transaksi pelanggan yang terjadi dalam sehari dan ada berapa juta data yang tersimpan dalam sebulan.

Ekstraksi informasi yang berguna dari basis data tersebut menjadi pekerjaan yang cukup menantang. Seringkali alat dan teknik analisis data tradisional tidak dapat digunakan dalam mengekstrak informasi dari data berukuran besar. Data mining adalah teknologi yang merupakan campuran teknik-teknik analisis data dengan algoritma-algoritma untuk memproses data berukuran besar. Data mining telah banyak diaplikasikan dalam berbagai bidang, seperti dalam bidang bisnis dan kedokteran.

Dalam bidang bisnis, teknik data mining digunakan untuk mendukung cakupan yang luas dari aplikasi-aplikasi bisnis inteligen seperti *customer profiling*, *targeted marketing, workflow management, store layout* dan *fraud detection*. Teknik data mining dapat digunakan untuk menjawab pertanyaan bisnis yang penting seperti “Siapakah pelanggan yang akan paling banyak mendatangkan keuntungan?” dan ”Seperti apa perkiraan pendapatan perusahaan tahun depan?”. Dalam bidang kedokteran, peneliti dalam bidang biomolekuler dapat menggunakan teknik data mining untuk menganalisis sejumlah besar data *genomic* yang sekarang ini telah banyak dikumpulkan untuk menjelaskan struktur dan fungsi gen, memprediksi struktur protein, dan lain-lain.

A. DEFINISI DAN KONSEP DATA MINING

Kalau kita membahas tentang Data Mining, tentulah kita harus mengetahui terlebih dahulu definisi dari Data Mining. Secara umum Data Mining terbagi atas 2 (dua) kata berikut.

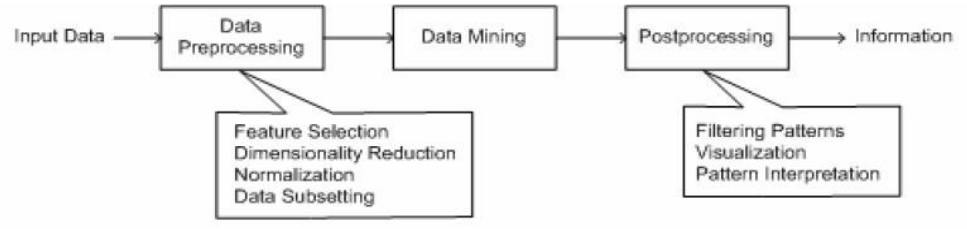
1. Data yaitu Kumpulan Fakta yang terekam atau sebuah entitas yang tidak memiliki arti dan selama ini terabaikan.
2. Mining yaitu proses Penambangan sehingga, Data Mining itu dapat diartikan sebagai proses penambangan data yang menghasilkan sebuah *ouput* (keluaran) berupa pengetahuan.

Data mining adalah sebuah proses pencarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Teknik data mining digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Namun tidak semua pekerjaan pencarian informasi dapat dinyatakan sebagai data mining.

Berikut ini adalah beberapa contoh yang merupakan data mining dan yang bukan data mining.

1. Bukan Data Mining: Pencarian informasi tertentu di internet, misalnya dengan mengetikkan *keyword* di mesin pencari.
   Data Mining: Pengelompokkan informasi yang mirip dalam konteks tertentu pada hasil pencarian dengan mesin pencari. Contoh mengelompokkan hasil pencarian Universitas berdasarkan nilai akreditasi, jumlah mahasiswa, kualitas alumni, dan lain-lain.
2. Bukan Data Mining: Membandingkan data laporan keuangan bulan Januari dan Februari untuk mencari berapa kenaikan permintaan.
   Data Mining: Hasil laporan keuangan saat ini digunakan untuk mempredikasi pengeluaran dan pembelian selanjutnya.

Istilah lain yang sering dikaitkan dengan data mining diantaranya *knowledge discovery* (mining) *in databases*, *knowledge extraction*, *data/pattern analysis*, *data archeology*, *data dredging*, *information harvesting*, dan *business intelligence*. Data mining adalah bagian integral dari *Knowledge Discovery in Databases* (KDD). Keseluruhan proses KDD untuk konversi *raw* data (data mentah) ke dalam informasi yang berguna ditunjukkan dalam Gambar 1.1.



Gambar 1.1
Proses dalam KDD

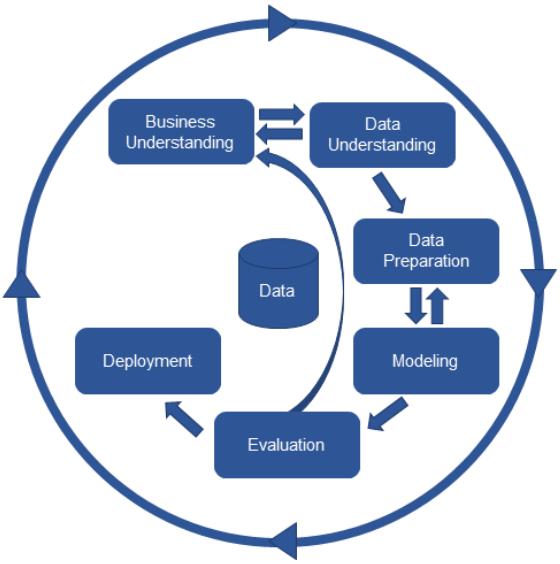
Data *input* dapat disimpan dalam berbagai format seperti *flat file*, *spreadsheet*, atau tabel-tabel relasional, dan dapat menempati tempat penyimpanan data terpusat atau terdistribusi pada banyak tempat. Selanjutnya dilakukan *prepocessing* data untuk mentransformasikan data *input* mentah ke dalam format yang sesuai untuk analisis selanjutnya. Langkah-langkah yang terlibat dalam *preprocessing* data meliputi menggabungkan data dari berbagai sumber, membersihkan (*cleaning*) data untuk membuang *noise* dan observasi duplikat, serta menyeleksi *record* dan fitur yang relevan untuk pekerjaan data mining. Karena terdapat banyak cara mengumpulkan dan menyimpan data, tahapan *preprocessing* data merupakan langkah yang banyak menghabiskan waktu dalam KDD.

Hasil dari data mining sering kali diintegrasikan dengan *decision support system* (DSS) atau sistem pendukung keputusan. Sebagai contoh, dalam aplikasi bisnis informasi yang dihasilkan oleh data mining dapat diintegrasikan dengan *tool* manajemen kampanye produk sehingga promosi pemasaran yang efektif yang dilaksanakan dan dapat diuji. Integrasi demikian memerlukan langkah *postprocessing* yang menjamin bahwa hanya hasil yang valid dan berguna yang akan digabungkan dengan DSS. Salah satu proses yang dilaksanakan pada tahap *postprocessing* adalah visualisasi yang memungkinkan analis untuk mengeksplor data dan hasil data mining dari berbagai sudut pandang. Ukuran-ukuran statistik dan teknik pengujian hipotesis juga dapat digunakan selama *postprocessing* untuk membuang hasil data mining yang palsu.

Secara khusus, data mining menggunakan ide-ide seperti (1) pengambilan contoh, estimasi, dan pengujian hipotesis dari statistika dan (2) algoritma pencarian, teknik pemodelan, dan teori pembelajaran dari kecerdasan buatan, pengenalan pola, dan *machine learning*. Data mining juga telah mengadopsi ide-ide dari area lain meliputi optimisasi, *evolutionary computing*, teori informasi, pemrosesan sinyal, visualisasi dan *information retrieval*. Sejumlah area lain juga memberikan peran pendukung dalam data mining, seperti sistem basis data yang dibutuhkan untuk menyediakan tempat penyimpanan yang efisien, *indexing* dan pemrosesan *query*.

B. PROSES DATA MINING

Cross-Industry Standard Process for Data Mining (CRISP-DM) yang dikembangkan pada tahun 1996 oleh analis dari beberapa industri seperti *Daimler Chrysler*, NCR, dan SPPSS, menyediakan proses standar dari data mining sebagai strategi pemecahan masalah secara umum dari bisnis maupun unit penelitian.



Gambar 1.2
Fase pada CRISP-DM

Dalam CRISP-DM, sebuah proses data mining memiliki siklus hidup yang terdiri dari enam fase. Keseluruhan fase berurutan yang ada tersebut bersifat adaptif. Seperti terlihat pada gambar diatas, fase berikutnya dalam urutan bergantung pada keluaran dari fase sebelumnya. Fase-fase dalam CRISP-DM dapat dijelaskan sebagai berikut.

1. Fase Pemahaman Bisnis (*Business Understanding Phase*), yang terdiri dari tahapan:
   a. penentuan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau penelitian secara keseluruhan,
   b. menerjemahkan tujuan dan batasan menjadi formula dari permasalahan data mining,
   c. menyiapkan strategi awal untuk mencapai tujuan.

   Contohnya adalah saat kita menentukan permasalahan memprediksi berapa harga jual cabai di bulan depan, agar bisa memperoleh keuntungan yang maksimal.

2. Fase Pemahaman Data (*Data Understanding Phase*), yang terdiri dari tahapan:
   a. mengumpulkan data,
   b. menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal,
   c. mengevaluasi kualitas data.

   Contoh pada fase ini adalah bagaimana kita melakukan tabulasi data penjualan sertai memastikan tidak ada data yang kosong atau salah ketik.

3. Fase Persiapan Data (*Data Preparation Phase*), yang terdiri dari tahapan:
   a. mempersiapkan data awal yang akan digunakan untuk keseluruhan fase berikutnya,
   b. memilih variabel yang sesuai dan akan dianalisis,
   c. melakukan perubahan pada beberapa variabel jika diperlukan.

   Contoh pada fase ini adalah proses normalisasi data sehingga data memiliki *range* yang konsisten.

4. Fase Pemodelan (*Modeling Phase*), yang terdiri dari tahapan:
   a. mengaplikasikan teknik pemodelan yang sesuai,
   b. melakukan kalibrasi aturan model untuk mengoptimalkan hasil,
   c. jika diperlukan, proses dapat kembali ke fase persiapan data untuk menjadikan data ke dalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik data mining tertentu.

   Pada fase ini kita sudah menerapkan teknik-teknik data mining yang sesuai untuk menyelesaikan permasalahan. Contoh melakukan klasifikasi dengan membuat sebuah model berupa *decision tree*.

5. Fase Evaluasi (*Evaluation Phase*), yang terdiri dari tahapan:
   a. mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum digunakan atau disebarkan,
   b. menetapkan model yang memenuhi tujuan pada fase awal,
   c. memastikan tidak terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik,
   d. mengambil keputusan berkaitan dengan penggunaan hasil dari data mining.

   Contohnya adalah melakukan pemilihan dan evaluasi *model tree* yang diperoleh. Evaluasi dapat dilakukan dengan menggunakan *k-fold cross validation* dan melihat kinerja dari teknik terssebut.

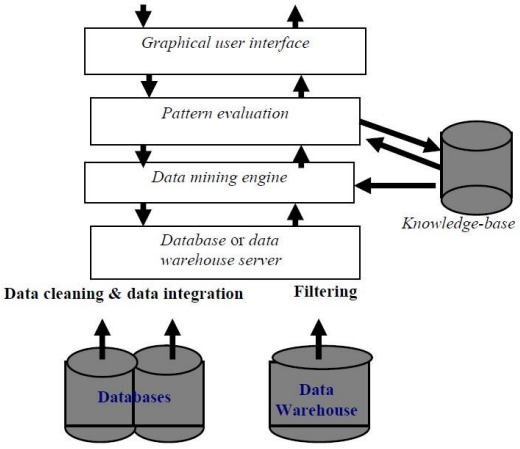
6. Fase Penyebaran (*Deployment Phase*), yang terdiri dari tahapan:
   a. menggunakan model yang dihasilkan,
   b. contoh penyebaran misalkan pembuatan laporan dan penerapan proses data mining secara paralel pada departemen lain.

C. ARSITEKTUR DATA MINING

Data mining merupakan proses pencarian pengetahuan yang menarik dari data berukuran besar yang disimpan dalam basis data, *data warehouse* atau tempat penyimpanan informasi lainnya. Dengan demikian arsitektur sistem data mining memiliki komponen-komponen utama sebagai berikut.

1. Basis data, *data warehouse* atau tempat penyimpanan informasi lainnya. Komponen ini bertanggung jawab dalam pengambilan data yang relevan berdasarkan permintaan pengguna.
2. Basis pengetahuan. Komponen ini merupakan *domain knowledge* yang digunakan untuk memandu pencarian atau mengevaluasi pola-pola yang dihasilkan. Pengetahuan tersebut meliputi hierarki konsep yang digunakan untuk mengorganisasikan atribut atau nilai atribut ke dalam level abstraksi yang berbeda. Pengetahuan tersebut juga dapat berupa kepercayaan pengguna (*user belief*), yang dapat digunakan untuk menentukan kemenarikan pola yang diperoleh. Contoh lain dari *domain knowledge* adalah *threshold* dan metadata yang menjelaskan data dari berbagai sumber yang heterogen.
3. Data mining *engine*. Bagian ini merupakan komponen penting dalam arsitektur sistem data mining. Komponen ini terdiri modul-modul fungsional data mining seperti karakterisasi, asosiasi, klasifikasi, dan analisis klaster.
4. Modul evaluasi pola. Komponen ini menggunakan ukuran-ukuran kemenarikan dan berinteraksi dengan modul data mining dalam pencarian pola-pola menarik. Modul evaluasi pola dapat menggunakan *threshold* kemenarikan untuk mem-*filter* pola-pola yang diperoleh.
5. Antarmuka pengguna grafis. Modul ini berkomunikasi dengan pengguna dan sistem data mining. Melalui modul ini, pengguna berinteraksi dengan sistem untuk menentukan *queri* atau *task* data mining. Antarmuka juga menyediakan informasi untuk memfokuskan pencarian dan melakukan eksplorasi data mining berdasarkan hasil data mining antara. Komponen ini juga memungkinkan pengguna untuk mencari (*browse*) basis data dan skema data *warehouse* atau struktur data, evaluasi pola yang diperoleh dan visualisasi pola dalam berbagai bentuk.

Arsitektur sebuah sistem data mining dapat dilihat dalam Gambar 1.3



Sumber: Han & Kamber (2001)

Gambar 1.3
Arsitektur Sistem Data Mining

Data mining dapat diaplikasikan pada berbagai jenis penyimpanan data seperti basis data relasional, *data warehouse*, *transactional database*, *object-oriented* dan *object-relational databases*, *spatial databases*, *time-series data* dan *temporal data*, *text databases* dan *multimedia databases*, *heterogeneous* dan *legacy databases*, serta *WWW*.

1. Basis Data Relasional

Basis data relasional merupakan koleksi dari tabel. Setiap tabel berisi atribut (i*ield*) dan biasanya menyimpan sejumlah besar *tuple* (*record*). Setiap *tuple* dalam tabel relasional merepesentasikan sebuah objek yang diidentifikasikan oleh kunci unik dan dideskripsikan oleh sekumpulan nilai atribut. Data relasional dapat diakses oleh *queri* basis data yang ditulis dalam bahasa *queri* relasional seperti SQL atau dengan bantuan antarmuka pengguna grafis.

2. Data *Warehouse*

Data *warehouse* merupakan tempat penyimpanan informasi yang dikumpulkan dari berbagai sumber, disimpan dalam skema yang dipersatukan (*unified schema*) dan biasanya bertempat pada tempat penyimpanan tunggal. Data *warehouse* dikonstruksi melalui sebuah proses *data cleaning*, *data transformation*, *data integration*, *data loading* dan *periodic data refreshing*. Untuk memfasilitasi proses pembuatan keputusan, data dalam data *warehouse* diorganisasikan ke dalam subjek utama seperti *customer*, *item*, *supplier* atau aktivitas. Data disimpan untuk menyediakan informasi dari perspektif sejarah (seperti 5-10 tahun yang lalu) dan biasanya data tersebut diringkas (*summarized*). Sebagai contoh, daripada menyimpan data rinci dari transaksi penjualan, data *warehouse* dapat menyimpan ringkasan dari transaksi per tipe *item* untuk setiap toko atau diringkas dalam level yang lebih tinggi seperti daerah pemasaran.

Data *warehouse* biasanya dimodelkan oleh struktur basis data multidimensional, dimana setiap dimensi berkaitan dengan sebuah atribut atau sekumpulan atribut dalam skema, dan setiap sel menyimpan nilai dari ukuran agregasi seperti *count* dan *sales_amount*. Struktur fisik dari data *warehouse* dapat berupa penyimpanan basis data relasional atau sebuah kubus data multidimensional.

Selain data *warehouse*, terdapat istilah penyimpanan data yang lain yaitu data *mart*. Sebuah data *warehouse* mengumpulkan informasi mengenai subjek-subjek yang menjangkau seluruh organisasi, dengan demikian cakupannya *enterprise-wide*. Sedangkan data *mart* merupakan sub bagian dari data *warehouse*. Fokus data *mart* adalah pada subjek yang dipilih dan dengan demikian cakupannya adalah *department-wide.*

3. Basis Data Transaksional

Secara umum, basis data transaksional terdiri dari sebuah *file* dimana setiap *record* merepresentasikan transaksi. Sebuah transaksi biasanya meliputi bilangan identitas transaksi yang unik (`trans_id`), dan sebuah daftar dari *item* yang membuat transaksi (seperti *item* yang dibeli dalam sebuah *took*). Basis data transaksi dapat memiliki tabel tambahan, yang mengandung informasi lain berkaitan dengan penjualan seperti tanggal transaksi, *customer* ID *number*, ID *number* dari *sales person* dan dari kantor cabang (*branch*) dimana penjualan terjadi.

Latihan

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

1) Apa yang Anda pahami tentang data mining?
2) Jelaskan secara singkat tujuan dari adanya data mining!
3) Sebutkan dan jelaskan secara singkat proses data mining sesuai dengan CRISP-DM!
4) Sebutkan dan jelaskan komponen-komponen dalam sebuah arsitektur data mining!

Petunjuk Jawaban Latihan

1) Data mining adalah sebuah proses pencarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Teknik data mining digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna.

2) Perkembangan yang cepat dalam teknologi pengumpulan dan penyimpanan data telah memudahkan manusia untuk mengumpulkan sejumlah data berukuran besar sehingga menghasilkan gunung data. Ekstraksi informasi yang berguna dari basis data tersebut menjadi pekerjaan yang cukup menantang. Seringkali alat dan teknik analisis data tradisional tidak dapat digunakan dalam mengekstrak informasi dari data berukuran besar. Data mining adalah teknologi yang merupakan campuran teknik-teknik analisis data dengan algoritma-algoritma untuk memproses data berukuran besar.
3) Tahapan proses data mining sesuai dengan CRISP-DM terdiri dari: Fase Pemahaman Data, Fase Persiapan Data, Fase Pemodelan, Fase Evaluasi, dan Fase Penyebaran.
4) Komponen dalam sebuah arsitektur data mining terdiri dari basis data, basis pengetahuan, mesin data mining, modul evaluasi pola, dan antar muka grafis.

Rangkuman

Data mining adalah sebuah proses pencarian secara otomatis informasi yang berguna dalam tempat penyimpanan data berukuran besar. Teknik data mining digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Secara khusus, data mining menggunakan ide-ide seperti (1) pengambilan contoh, estimasi, dan pengujian hipotesis dari statistika dan (2) algoritme pencarian, teknik pemodelan dan teori pembelajaran dari kecerdasan buatan, pengenalan pola, dan *machine learning*. Data mining juga telah mengadopsi ide-ide dari area lain meliputi optimisasi, *evolutionary computing*, teori informasi, pemrosesan sinyal, visualisasi dan *information retrieval*. Komponen dalam sebuah arsitektur data mining terdiri dari basis data, basis pengetahuan, mesin data mining, modul evaluasi pola, dan antar muka grafis. Tahapan proses data mining sesuai dengan CRISP-DM terdiri dari: Fase Pemahaman Data, Fase Persiapan Data, Fase Pemodelan, Fase Evaluasi, dan Fase Penyebaran.

Tes Formatif 1

Pilihlah satu jawaban yang paling tepat!

1) Kumpulan fakta yang terekam adalah pengertian dari ….
   A. informasi
   B. data
   C. pengetahuan
   D. *record*

2) Makna kata “mining” dalam data mining adalah ….
   A. pemaknaan
   B. perubahan
   C. penggalian
   D. penggunaan

3) Berikut ini yang bukan merupakan definisi dari data mining adalah ….
   A. proses penambangan data yang menghasilkan sebuah *ouput* (keluaran) berupa pengetahuan
   B. proses pencarian pengetahuan yang menarik dari data berukuran besar yang disimpan dalam basis data, data *warehouse* atau tempat penyimpanan informasi lainnya
   C. proses menemukan fakta yang dikumpulkan dari berbagai sumber, disimpan dalam skema yang dipersatukan (*unified schema*)
   D. ekstraksi informasi yang berguna dari sebuah basis data

4) Proses berikut ini yang tergolong dalam data mining adalah ….
   A. mencari mahasiswa yang bernama Aan di dalam basis data absensi
   B. menghitung laba dan rugi perusahaan dengan merekap data transaksi
   C. memprediksi lama pengiriman makanan lewat Ojek *Online* berdasarkan jarak, kecepatan kendaraan, jenis makanan, dan jumlah *traffic light*
   D. menghitung rata-rata IPK dari sekelompok mahasiswa yang memiliki tahun kelahiran sama

5) Berikut adalah komponen-komponen pada sebuah arsitektur data mining, *kecuali* ….
   A. basis data
   B. basis pengetahuan
   C. data mining *engine*
   D. *search engine*

6) *Domain knowledge* yang digunakan untuk memandu pencarian atau mengevaluasi pola-pola yang dihasilkan adalah pengertian dari ….
   A. basis data
   B. basis pengetahuan
   C. data mining
   D. antarmuka pengguna

7) Tempat penyimpanan informasi yang dikumpulkan dari berbagai sumber, disimpan dalam skema yang dipersatukan (*unified schema*) dan biasanya bertempat pada tempat penyimpanan tunggal adalah pengertian dari ….

A. data *warehouse*
B. data *mart*
C. basis data relasional
D. basis data transaksional

8) Yang bukan merupakan tahapan pada proses *Knowledge Discovery in Database* adalah ….

A. data *preprocessing*
B. data mining
C. data *postprocessing*
D. data *reprocessing*

9) Proses mengevaluasi kualitas data dalam *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dilakukan pada fase ….

A. pemodelan data
B. pemahaman bisnis
C. pemahaman data
D. evaluasi

10) Proses yang dilakukan pada Fase Evaluasi dalam CRISP-DM antara lain, *kecuali* ….

A. memastikan kualitas model
B. memastikan kualitas data
C. memilih model yang tepat
D. memastikan semua permasalahan dapat diselesaikan

Cocokkanlah jawaban Anda dengan Kunci Jawaban Tes Formatif 1 yang terdapat di bagian akhir modul ini. Hitunglah jawaban yang benar. Kemudian, gunakan rumus berikut untuk mengetahui tingkat penguasaan Anda terhadap materi Kegiatan Belajar 1.

$$\text{Tingkat Penguasaan} = \frac{\text{Jumlah Jawaban yang Benar}}{\text{Jumlah Soal}} \times 100$$

Arti tingkat penguasaan

<70%	70% - 79%	80% - 89%	90% - 100%
kurang	cukup	baik	baik sekali

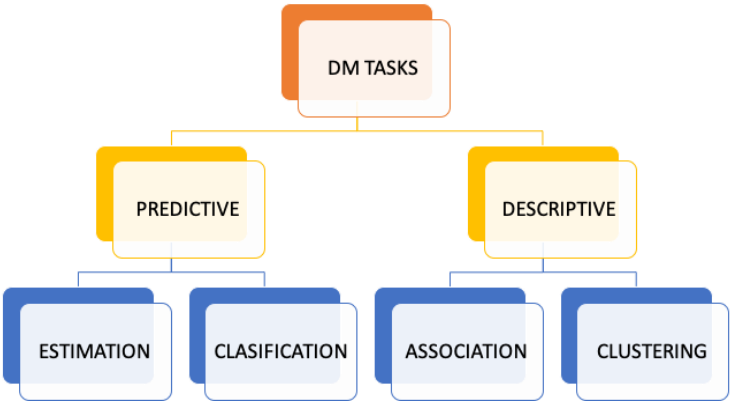
Apabila mencapai tingkat penguasaan 80% atau lebih, Anda dapat meneruskan dengan Kegiatan Belajar 2. **Bagus!** Jika masih di bawah 80%, Anda harus mengulangi materi Kegiatan Belajar 1, terutama bagian yang belum dikuasai.

Kegiatan Belajar 2

Kebutuhan akan Data Mining

Kebutuhan atau pentingnya bidang data mining tidak terlepas dari tujuan dan peran dari data mining itu sendiri. Adapun tujuan data mining secara umum dapat dibagi ke dalam dua kategori utama, yaitu :

1. **Prediktif.** Tujuannya untuk memprediksi atau memperkirakan nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variabel tak bebas, sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explanatory* atau variabel bebas.
2. **Deskriptif.** Tujuannya untuk menurunkan pola-pola (korelasi, *trend*, *cluster*, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas data mining deskriptif sering merupakan penyelidikan dan seringkali memerlukan teknik *postprocessing* untuk validasi dan penjelasan hasil.



Gambar 1.4
Peran Data Mining

Berdasarkan peran data mining dalam melakukan proses prediksi dan mendeskripsikan data, tugas data mining dapat dibagi ke dalam empat kelompok utama, yaitu:

1. Estimasi

Teknik untuk melakukan estimasi terhadap sebuah data baru yang tidak memiliki keputusan berdasarkan histori data yang telah ada, dimanavariabel 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 dari variabel prediksi. Contohnya melakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk memprediksi kasus baru lainnya.

2. Klasifikasi

Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Salah satu contoh yang mudah dan popular adalah dengan *decision tree* yaitu salah satu teknik klasifikasi yang paling populer karena mudah untuk interpretasi seperti Algoritma C4.5, ID3 dan lain-lain. Contoh pemanfaatannya misalnya pada bidang akademik terkait klasifikasi siswa yang layak masuk kedalam kelas unggulan atau akselerasi di sekolah tertentu.

3. Asosiasi

Teknik untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian. Adapun teknik pemecahan masalah yang sering digunakan seperti Algoritma Apriori. Contoh pemanfaatan Algoritma apriori yaitu pada bidang Marketing ketika sebuah Minimarket melakukan tata letak produk yang dijual berdasarkan produk-produk mana yang paling sering dibeli konsumen, selain itu seperti tata letak buku yang dilakukan pustakawan di perpustakaan.

4. Klasterisasi

Teknik untuk mengelompokkan data dan membentuk kelas objek-objek yang memiliki kemiripan. Klaster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam klaster lain. Proses klasterisasi berbeda dengan proses klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. roses klasterisasi mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), dimana kemiripan data dalam satu kelompok tinggi (maksimal) dan kemiripan data dengan data pada kolompok lain rendah (minimal). Contoh klasterisasi misalkan proses mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.

Secara garis besar terdapat 2 pendekatan untuk melakukan teknik - teknik data mining di atas, yaitu:

1. ***Supervised Learning*** yaitu pembelajaran menggunakan guru dan biasanya ditandai dengan adanya *class*/label/target pada himpunan data. *Supervised lerning* merupakan sebuah pendekatan dimana sudah terdapat data yang dilatih, dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengelompokan suatu data ke data yang sudah ada. Adapun teknik-teknik yang digunakan yang bersifat *supervised learning* seperti Teknik Prediksi dan Klasifikasi seperti Algoritma C4.5, Teknik *Rough* Set dan lain-lain.
2. ***Unsupervised Learning*** yaitu pembelajaran tanpa menggunakan guru dan biasanya ditandai pada himpunan datanya tidak memiliki atribut keputusan atau *class*/label/target. Perbedaan dengan *supervised learning*, pada *unsupervised learning* tidak memiliki data latih, sehingga dari data yang ada dikelompokan menjadi 2 bagian atau 3 bagian dan seterusnya. Adapun teknik-teknik yang bersifat *Unsupervised Learning* yaitu *Teknik Estimasi*, *Clustering*, dan Asosiasi seperti *Regresi Linier*, *Analytical Hierarchy Clustering* dan lain-lain.

Penerapan Data Mining

Berikut ini adalah beberapa contoh penerapan teknik data mining dalam berbagai bidang:

a. Kesehatan

Data mining memiliki potensi besar untuk memperbaiki sistem kesehatan. Menggunakan data dan analisis untuk mengidentifikasi praktik terbaik yang meningkatkan perawatan dan mengurangi biaya. Peneliti menggunakan pendekatan data mining seperti database multi dimensi, pembelajaran mesin, *soft computing*, visualisasi data dan statistik. Penambangan dapat digunakan untuk memprediksi volume pasien dalam setiap kategori. Proses dikembangkan untuk memastikan bahwa pasien mendapat perawatan yang tepat, di tempat yang tepat, dan pada saat yang tepat. Data mining juga dapat membantu perusahaan asuransi kesehatan untuk mendeteksi kecurangan dan penyalahgunaan.

b. Analisis Pasar

Analisis Pasar adalah teknik pemodelan berdasarkan teori bahwa jika seorang membeli kelompok item tertentu, maka cenderung membeli kelompok item lainnya. Teknik ini memungkinkan pengecer memahami perilaku pembelian dari para pembeli. Informasi ini dapat membantu pengecer mengetahui kebutuhan pembeli dan akan mengubah tata letak toko sesuai dengan hasil analisis tersebut. Perbandingan hasil antara toko yang berbeda, atau antara pelanggan dalam kelompok demografis yang berbeda dapat dilakukan dengan menggunakan analisis diferensial.

c. Pendidikan

Pada bidang pendidikan, muncul bidang baru yang disebut *Educational* Data Mining (EDM). Bidang tersebut berkaitan dengan teknik pengembangan yang menemukan pengetahuan dari data yang berasal dari lingkungan pendidikan. Tujuan EDM ebagai prediksi perilaku belajar di masa depan siswa, mempelajari dampak dukungan pendidikan, dan memajukan pengetahuan ilmiah tentang pembelajaran. Data mining dapat digunakan oleh sebuah institusi untuk mengambil keputusan yang akurat dan juga untuk memprediksi hasil siswa. Dengan hasilnya institusi bisa fokus pada apa yang harus diajarkan dan bagaimana cara mengajarnya. Pola belajar siswa dapat diambil dan digunakan untuk mengembangkan teknik mengajar mereka.

d. Rekayasa Manufaktur

Pengetahuan adalah aset terbaik yang dimiliki perusahaan manufaktur. Alat data mining bisa sangat berguna untuk menemukan pola dalam proses manufaktur yang kompleks. Data mining dapat digunakan dalam perancangan tingkat sistem untuk mengekstrak hubungan antara arsitektur produk, portofolio produk, dan data kebutuhan pelanggan. Ini juga bisa digunakan untuk memprediksi perkembangan produk *span time*, *cost*, dan *dependencies* antar tugas lainnya.

e. CRM

Customer Relationship Management adalah tentang mengakuisisi dan mempertahankan pelanggan, juga meningkatkan loyalitas pelanggan dan menerapkan strategi yang berfokus pada pelanggan. Untuk menjaga hubungan yang benar dengan pelanggan bisnis, perlu untuk mengumpulkan data dan menganalisa informasi. Di sinilah data mining berperan, dimana dengan teknologi data mining, maka data yang terkumpul dapat digunakan untuk melakukan analisis. Hasil analisis dapat digunakan untuk pengelompokkan target pemasaran, prediksi pelanggan, *profilling* pelanggan, dan lainnya.

f. *Fraud Detection* (Deteksi Penipuan)

Miliaran dolar bisa hilang akibat aksi penipuan. Metode tradisional untuk mendeteksi kecurangan memakan waktu dan kompleks. Data mining membantu dalam memberikan pola yang berarti dan mengubah data menjadi informasi. Setiap informasi yang valid dan berguna adalah pengetahuan. Sistem deteksi kecurangan yang sempurna harus melindungi informasi semua pengguna. Metode yang diawasi mencakup pengumpulan catatan sampel. Catatan ini akan digolongkan sebagai catatan curang atau tidak palsu. Sebuah model dibangun dengan menggunakan data ini dan algoritma dibuat untuk mengidentifikasi apakah rekaman tersebut salah atau tidak.

g. *Intrusion Detection*

Setiap tindakan yang akan membahayakan integritas dan kerahasiaan sumber daya adalah gangguan. Langkah-langkah defensif untuk menghindari gangguan mencakup otentikasi pengguna, menghindari kesalahan pemrograman, dan perlindungan informasi. Data mining dapat membantu memperbaiki deteksi intrusi dengan menambahkan tingkat fokus pada deteksi anomali. Hal tersebut membantu analis dalam membedakan aktivitas dari aktivitas jaringan biasa sehari-hari. Data mining juga membantu mengekstraksi data yang lebih relevan dengan masalah.

h. Deteksi Kebohongan

Menangkap penjahat merupakan hal yang mudah, namun membawa keluar kebenaran dari penjahat tersebut adalah hal yang sulit. Penegakan hukum bisa menggunakan teknik penambangan, misalkan untuk menyelidiki kejahatan atau memantau komunikasi tersangka teroris. Hal ini termasuk penambangan teks juga. Proses tersebut berusaha untuk menemukan pola yang berarti dalam data, yang biasanya berupa teks tidak terstruktur. Kemudian sampel data yang dikumpulkan dari penelitian sebelumnya dibandingkan dan dibuat sebuah model untuk deteksi kebekuan.

i. Segmentasi Pelanggan

Penelitian pasar tradisional dapat membantu untuk meng-segmentasikan pelanggan, namun data mining dapat digunakan untuk meningkatkan efektivitas pasar. Data mining merupakan alat bantu dalam upaya menyelaraskan pelanggan menjadi segmen yang berbeda dan dapat menyesuaikan kebutuhan sesuai pelanggan. Pasar selalu mempertahankan konsumen, sehingga penggunaan data mining memungkinkan untuk menemukan segmen pelanggan berdasarkan kerentanan dan bisnis dapat menawarkannya dengan penawaran khusus dan meningkatkan kepuasan.

j. Perbankan/Keuangan

Dengan komputerisasi perbankan di mana-mana sejumlah besar data seharusnya dihasilkan dengan transaksi baru. Data mining dapat berkontribusi untuk memecahkan masalah bisnis di bidang perbankan dan keuangan dengan menemukan pola, sebab-akibat, dan korelasi dalam informasi bisnis dan harga pasar yang tidak segera terlihat oleh manajer karena data volume terlalu besar atau dihasilkan terlalu cepat untuk disaring oleh para ahli. Para manajer dapat menemukan informasi ini untuk segmentasi, penargetan, perolehan, penahanan, dan pemeliharaan pelanggan yang lebih baik.

k. Pengawasan Perusahaan

Pengawasan perusahaan merupakan pemantauan perilaku seseorang atau kelompok oleh perusahaan. Data yang dikumpulkan paling sering digunakan untuk tujuan pemasaran atau dijual ke perusahaan lain, namun

dapat juga dibagi secara reguler dengan instansi pemerintah. Hal ini dapat digunakan oleh bisnis untuk menyesuaikan produk mereka dengan yang diinginkan oleh pelanggan. Data tersebut dapat digunakan untuk tujuan pemasaran langsung, seperti iklan bertarget di Google dan Yahoo, di mana iklan ditargetkan ke pengguna mesin pencari dengan menganalisis riwayat pencarian dan *email* mereka.

l. Analisis Riset

Sejarah menunjukkan bahwa kita telah menyaksikan perubahan revolusioner dalam penelitian. Data mining sangat membantu dalam pembersihan data, pra-pengolahan data dan integrasi database. Para peneliti dapat menemukan data serupa dari *database* yang mungkin membawa perubahan dalam penelitian. Identifikasi sekuens *co-occurring* dan korelasi antara aktivitas apapun dapat diketahui. Visualisasi data dan data mining visual memberi kita gambaran yang jelas tentang data.

m. Investigasi Kriminal

Kriminologi adalah proses yang bertujuan untuk mengidentifikasi karakteristik kejahatan. Sebenarnya analisis kejahatan mencakup menjajaki dan mendeteksi kejahatan dan hubungannya dengan penjahat. Tingginya volume *dataset* kejahatan dan juga kompleksitas hubungan antar data semacam ini membuat kriminologi menjadi bidang yang tepat untuk menerapkan teknik data mining. Laporan kejahatan berbasis teks dapat diubah menjadi *file* pengolah kata. Informasi ini bisa digunakan untuk melakukan proses pencocokan kejahatan.

n. Bioinformatika

Pendekatan Data Mining nampaknya ideal untuk Bioinformatika, karena kaya akan data. Data biologi hasil penambangan membantu untuk mengekstrak pengetahuan yang berguna dari kumpulan data besar yang dikumpulkan dalam biologi, dan bidang ilmu kehidupan lainnya yang terkait seperti kedokteran dan ilmu saraf. Aplikasi data mining untuk bioinformatika meliputi penemuan gen, inferensi fungsi protein, diagnosis penyakit, prognosis penyakit, optimasi pengobatan penyakit, rekonstruksi jaringan interaksi protein dan gen, pembersihan data, dan prediksi lokasi sub-seluler protein.

Latihan

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

1) Sebutkan dan jelaskan 2 peran utama data mining!
2) Berikan sebuah contoh pemanfaatan data mining!

Petunjuk Jawaban Latihan

1) Peran data mining secara umum dapat dibagi ke dalam dua kategori utama, yaitu Prediktif dan Deskriptif. Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend, cluster*, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data.
2) Pemanfaatan data mining dapat dilakukan pada berbagai bidang, misalkan pada bidang pendidikan, kesehatan, perbankan, riset, rekayasa manufaktur, sampai dengan bioinformatika.

Rangkuman

Kebutuhan atau pentingnya bidang data mining tidak terlepas dari peran data mining itu sendiri. Adapun peran data mining secara umum dapat dibagi ke dalam dua kategori utama, yaitu: Prediktif dan Deskriptif. Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain. Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend*, *cluster*, trayektori, dan anomali) yang meringkas hubungan yang pokok dalam data. Berdasarkan peran data mining dalam melakukan proses prediksi dan mendeskripsikan data, tugas data mining dapat dibagi ke dalam empat kelompok utama, yaitu : estimasi, klasifikasi, asosiasi, dan klasterisasi. Secara garis besar terdapat 2 pendekatan untuk melakukan teknik - teknik data mining yaitu *supervised learning* dan *unsupervised learning*. Pemanfaatan data mining dapat dilakukan pada berbagai bidang, misalkan pada bidang pendidikan, kesehatan, perbankan, riset, rekayasa manufaktur, sampai dengan bioinformatika.

Tes Formatif 2

Pilihlah satu jawaban yang paling tepat!

1) Peran utama dari data mining adalah ….
   A. prediktif dan deskriptif
   B. prediktif dan persuasif
   C. deskriptif dan persuasif
   D. analisis dan visualisasi

2) Teknik data mining yang tergolong prediktif adalah ….
   A. klasifikasi dan klasterisasi
   B. klasifikasi dan asosiasi

C. estimasi dan klasifikasi
D. asosiasi dan klasterisasi

3) Teknik data mining yang tergolong deskriptif adalah ….
A. klasifikasi dan klasterisasi
B. klasifikasi dan asosiasi
C. estimasi dan klasifikasi
D. asosiasi dan klasterisasi

4) *Supervised learning* adalah ….
A. pendekatan dimana sudah terdapat data yang dilatih, dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan suatu data ke data yang sudah ada
B. pendekatan dimana sudah terdapat data yang dilatih, dan belum terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan suatu data ke data yang belum ada.
C. pembelajaran tanpa menggunakan guru dan biasanya ditandai pada himpunan datanya tidak memiliki atribut keputusan.
D. pembelajaran tanpa menggunakan guru dan biasanya ditandai pada himpunan datanya sudah memiliki atribut keputusan.

5) *Unsupervised learning* adalah ….
A. pendekatan dimana sudah terdapat data yang dilatih, dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan suatu data ke data yang sudah ada
B. pendekatan dimana sudah terdapat data yang dilatih, dan belum terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokan suatu data ke data yang belum ada.
C. pembelajaran tanpa menggunakan guru dan biasanya ditandai pada himpunan datanya tidak memiliki atribut keputusan.
D. pembelajaran tanpa menggunakan guru dan biasanya ditandai pada himpunan datanya sudah memiliki attribut keputusan.

6) Dalam data mining, teknik yang berupa proses pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (*homogen*), dimana kemiripan data dalam satu kelompok tinggi (*maksimal*) dan kemiripan data dengan data pada kolompok lain rendah (*minimal*) disebut dengan ….
A. klasifikasi
B. estimasi

C. klasterisasi
D. asosiasi

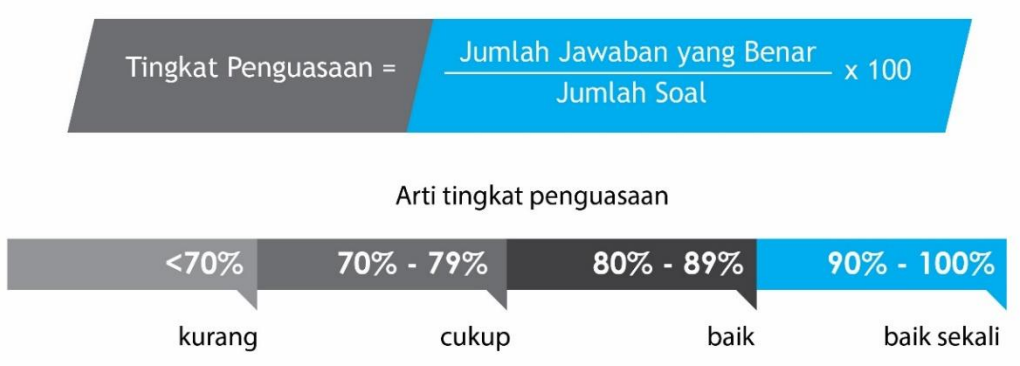
7) Memprediksi nilai dari atribut tertentu berdasarkan pada nilai dari atribut-atribut lain merupakan contoh peran dari data mining pada jenis ….
A. prediktif
B. asosiatif
C. deskriptif
D. klasterisasi

8) Jika ingin melakukan prediksi apakah seorang mahasiswa akan lulus tepat waktu atau tidak berdasarkan data alumni yang ada, maka teknik yang lebih cocok diterapkan adalah ….
A. klasterisasi
B. asosiasi
C. estimasi
D. klasifikasi

9) Klasterisasi merupakan teknik data mining yang berjenis ….
A. *unsupervised learning*
B. *supervised learning*
C. pembelajaran dengan target *class*
D. pembelajaran dengan data latih

10) Perbedaan dari teknik estimasi dan klasifikasi dalam data mining adalah ….
A. estimasi digunakan untuk melakukan prediksi, sedangkan klasifikasi untuk melakukan deskripsi data
B. klasifikasi digunakan untuk melakukan prediksi, sedangkan estimasi untuk melakukan deskripsi data
C. target data estimasi umumnya berupa data numerik, sedangkan klasifikasi berupa data kategorikal
D. target data estimasi umumnya berupa data kategorikal, sedangkan klasifikasi berupa data numerik

Cocokkanlah jawaban Anda dengan Kunci Jawaban Tes Formatif 2 yang terdapat di bagian akhir modul ini. Hitunglah jawaban yang benar. Kemudian, gunakan rumus berikut untuk mengetahui tingkat penguasaan Anda terhadap materi Kegiatan Belajar 2.

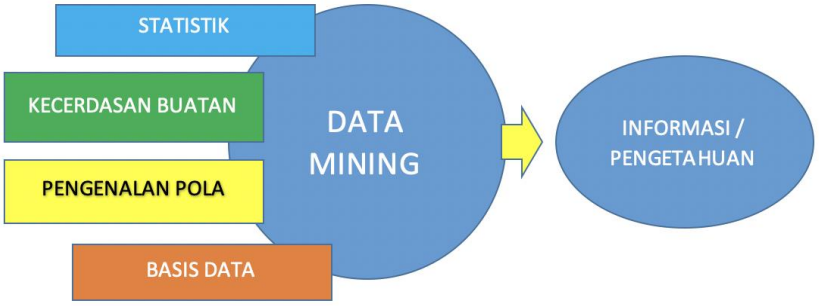


Apabila mencapai tingkat penguasaan 80% atau lebih, Anda dapat meneruskan dengan Kegiatan Belajar 3. **Bagus!** Jika masih di bawah 80%, Anda harus mengulangi materi Kegiatan Belajar 2, terutama bagian yang belum dikuasai.

Kegiatan Belajar 3

Ilmu yang berkaitan dengan Data Mining

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan yang dialami dalam mendeskripsikan data mining adalah fakta bahwa data mining mewarisi banyak aspek dan teknik-teknik dari bidang ilmu yang sudah ada dan mapan terlebih dahulu. Gambar 1.5 menunjukkan bahwa data mining memiliki akar yang panjang dari bidang ilmu seperti statistik, kecerdasan buatan, basis data dan pengenalan pola.



Gambar 1.5
Ilmu yang Berkaitan Erat dengan Data Mining

Data Mining bertujuan untuk memanfaatkan data dalam basis data dengan mengolahnya sehingga menghasilkan informasi baru yang berguna. Jika dilacak dari akar keilmuwannya, Data Mining memiliki 4 buah bidang ilmu yang mendasari yatu Statistik, Kecerdasan Buatan, Pengenalan Pola, dan Basis Data.

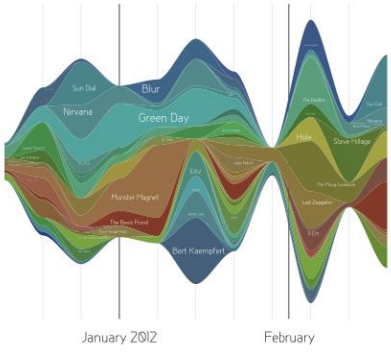
A. STATISTIK

Bidang ini merupakan akar paling tua, karena tanpa ada statistik maka data mining mungkin tidak ada. Dengan menggunakan Statistik Klasik ternyata data yang diolah dapat diringkas sebagai *Explanatory Data Analysis* (EDA). EDA berguna untuk mengidentifikasi hubungan sistematis antara variabel/fitur ketika tidak ada cukup informasi alami yang dibawanya.

Teknik EDA klasik yang digunakan dalam data mining diantaranya:

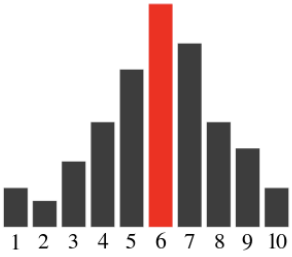
1. Teknik Komputasional, meliputi statistik deskriptif, korelasi, tabel frekuensi, teknik eksplorasi *multivariate*, model linear/non linear lanjutan, dan lain-lain;
2. Visualisasi Data, lebih mengarah pada representasi informasi dalam bentuk visual. Visualisasi data merupakan teknik eksplorasi data yang atraktif, dengan teknik visualisasi yang paling umum yang dikenal adalah histogram semua jenis (kolom, silinder, kerucut, pidamida, batang, dan sebagainya), kotak, *scatter*, kontur, matriks, ikon dan sebagainya.

Berikut ini adalah beberapa contoh diagram yang digunakan untuk melakukan visualisasi data.

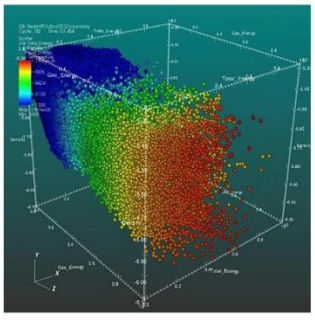


(a)

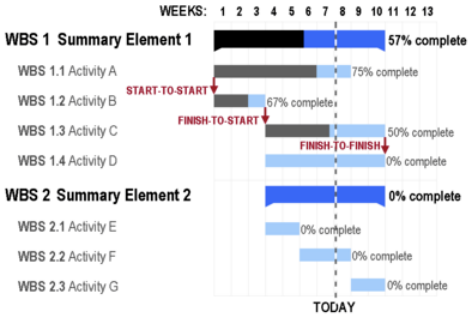
(b)



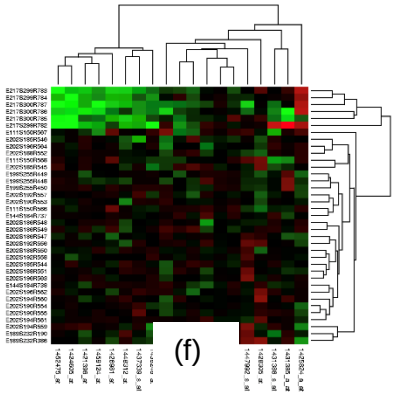
(c)



(d)



(e)



(f)

Gambar 1.6
Diagram (a) *Streamgraph*, (b) *Network*, (c) Batang, (d) *Scatter Plot* 3D, (e) *Gantt Chart*, dan (f) *Heat Map*

B. KECERDSASAN BUATAN (*ARTIFICIAL INTELLIGENCE*)

Kecerdasan Buatan (*Artificial Intelligence*) merupakan salah satu bagian dari ilmu komputer yang mempelajari bagaimana membuat mesin (komputer) dapat melakukan pekerjaan seperti dan sebaik yang dilakukan oleh manusia bahkan bisa lebih baik daripada yang dilakukan manusia. *Artificial Intelligence* (AI) digunakan untuk mengetahui dan memodelkan proses–proses berpikir manusia dan mendesain mesin agar dapat menirukan perilaku manusia. Cerdas, berarti memiliki pengetahuan ditambah pengalaman, penalaran (bagaimana membuat keputusan dan mengambil tindakan), moral yang baik.

Manusia cerdas (pandai) dalam menyelesaikan permasalahan karena manusia mempunyai pengetahuan dan pengalaman. Pengetahuan diperoleh dari belajar. Semakin banyak bekal pengetahuan yang dimiliki tentu akan lebih mampu menyelesaikan permasalahan. Namun bekal pengetahuan saja tidak cukup, karena manusia juga diberi akal untuk melakukan penalaran, sehingga dapat mengambil kesimpulan berdasarkan pengetahuan dan pengalaman yang dimiliki. Tanpa memiliki kemampuan untuk menalar dengan baik, manusia dengan segudang pengalaman dan pengetahuan tidak akan dapat menyelesaikan masalah dengan baik. Demikian juga dengan kemampuan menalar yang sangat baik, namun tanpa bekal pengetahuan dan pengalaman yang memadai, manusia juga tidak akan bisa menyelesaikan masalah dengan baik.

Oleh karena itu, agar mesin bisa cerdas (bertindak seperti dan sebaik manusia) maka harus diberi bekal pengetahuan, sehingga mempunyai kemampuan untuk menalar. Untuk membuat aplikasi kecerdasan buatan ada 2 bagian utama yang sangat dibutuhkan.

1. Basis Pengetahuan (*Knowledge Base*), bersifat fakta-fakta, teori, pemikiran dan hubungan antar satu dengan yang lainnya.
2. Motor Inferensi (*Inference Engine*), kemampuan menarik kesimpulan berdasarkan pengetahuan dan pengalaman.

Kecerdasan buatan memiliki beberapa kelebihan dan kekurangan jika dibandingkan dengan kecerdasan alami. Kelebihan kecerdasan buatan sebagai berikut.

1. Lebih bersifat permanen
   Kecerdasan alami bisa berubah karena sifat manusia pelupa. Kecerdasan buatan tidak berubah selama sistem komputer & program tidak mengubahnya.
2. Lebih mudah diduplikasi & disebarkan
   Memindahkan (*transfer*) pengetahuan manusia dari satu orang ke orang lain membutuhkan proses yang sangat lama & keahlian tidak akan pernah dapat diduplikasi dengan lengkap. Jadi jika pengetahuan terletak pada suatu sistem komputer, pengetahuan tersebut dapat disalin dari komputer tersebut & dapat dipindahkan dengan mudah ke komputer yang lain.
3. Lebih murah
   Menyediakan layanan komputer akan lebih mudah & murah dibandingkan mendatangkan seseorang untuk mengerjakan sejumlah pekerjaan dalam jangka waktu yang sangat lama. Bersifat konsisten karena kecerdasan buatan adalah bagian dari teknologi komputer sedangkan kecerdasan alami senantiasa berubah-ubah
4. Dapat didokumentasikan
   Keputusan yang dibuat komputer dapat didokumentasi dengan mudah dengan cara melacak setiap aktivitas dari sistem tersebut, sedangkan kecerdasan alami sangat sulit untuk direproduksi.
5. Cara kerja lebih cepat
6. Hasil lebih baik.

Sedangkan kelebihan dari kecerdasan alami dibandingkan kecerdasan buatan adalah sebagai berikut.

1. Kreatif, karena manusia memiliki kemampuan untuk menambah pengetahuan, sedangkan pada kecerdasan buatan untuk menambah pengetahuan harus dilakukan melalui sistem yang dibangun.
2. Memungkinkan orang untuk menggunakan pengalaman secara langsung, sedangkan pada kecerdasan buatan harus bekerja dengan input-input simbolik.
3. Pemikiran manusia dapat digunakan secara luas, sedangkan kecerdasan buatan sangat terbatas.

C. PENGENALAN POLA

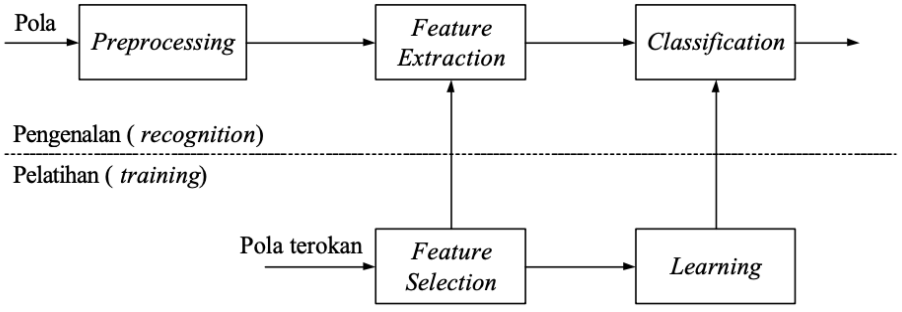
Pola adalah entitas yang terdefinisi dan dapat diidentifikasi melalui ciri-cirinya (*features*). Ciri-ciri tersebut digunakan untuk membedakan suatu pola dengan pola lainnya. Ciri yang bagus adalah ciri yang memiliki daya pembeda yang tinggi, sehingga pengelompokan pola berdasarkan ciri yang dimiliki dapat dilakukan dengan keakuratan yang tinggi. Sebagai contoh misalkan: pola huruf akan memiliki ciri berupa tinggi, tebal, titik sudut, lengkungan garis; pola suara akan memiliki ciri berupa amplitude, frekuensi, nada, intonasi, dan warna.

Pengenalan pola bertujuan menentukan kelompok atau kategori pola berdasarkan ciri-ciri yang dimiliki oleh pola tersebut. Dengan kata lain, pengenalan pola untuk membedakan suatu objek dengan objek lain. Ada dua fase dalam sistem pengenalan pola, yaitu: (a) fase pelatihan dan (b) fase pengenalan. Pada fase pelatihan, beberapa contoh objek data dipelajari untuk menentukan ciri yang akan digunakan dalam proses pengenalan serta prosedur klasifikasinya. Pada fase pengenalan, objek data diambil cirinya kemudian ditentukan kelas kelompoknya.

Terdapat dua pendekatan yang dilakukan dalam pengenalan pola, yaitu: pendekatan secara statistik dan pendekatan secara sintaktik atau *structural*.

1. Pengenalan Pola Secara Statistik

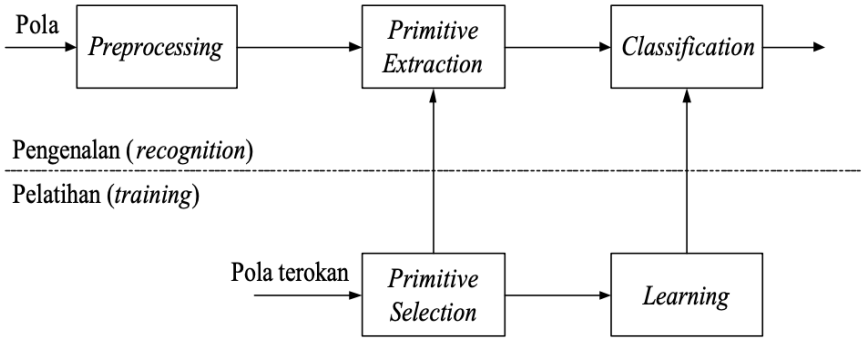
Pendekatan ini menggunakan teori-teori ilmu peluang dan statistik. Ciri-ciri yang dimiliki oleh suatu pola ditentukan oleh distribusi statistiknya. Pola yang berbeda akan memiliki distribusi yang berbeda pula. Dengan menggunakan teori keputusan di dalam statistik, kita menggunakan distribusi ciri untuk mengklasifikasikan pola. Sistem pengenalan pola dengan pendekatan statistik ditunjukkkan oleh Gambar 1.7.



Gambar 1.7
Diagram Sistem Pengenalan Pola dengan Pendekatan Statistik

2. Pengenalan Pola Secara Sintaktik

Pendekatan ini menggunakan teori bahasa formal. Ciri-ciri yang terdapat pada suatu pola ditentukan primitif dan hubungan struktural antara primitif kemudian menyusun tata bahasanya. Berdasarkan aturan produksi pada tata bahasa tersebut, kita dapat menentukan kelompok pola. Pengenalan pola secara sintaktik lebih dekat ke strategi pengenalan pola yang dilakukan manusia, namun secara praktek penerapannya relatif sulit dibandingkan pengenalan pola secara statistik. Gambar 1.8 memperlihatkan sistem pengenalan pola dengan pendekatan sintaktik.



Gambar 1.8
Diagram Sistem Pengenalan Pola dengan Pendekatan Sintaktik

D. SISTEM BASIS DATA

Basis Data terdiri dari kata basis dan data. Basis dapat diartikan sebagai markas atau gudang, sedangkan data adalah catatan atas kumpulan fakta dunia nyata yang mewakili objek seperti manusia, barang, hewan, konsep, peristiwa dan sebagainya yang diwujudkan dalam bentuk huruf, angka, simbol, gambar, teks, bunyi atau kombinasinya. Sebagai suatu kesatuan maka pengertian basis data atau biasa disebut *database* adalah sebagai berikut.

1. Himpunan kelompok data yang saling terhubung dan diorganisasi sedemikian rupa supaya kelak dapat dimanfaatkan kembali secara cepat dan mudah.
2. Kumpulan data dalam bentuk *file*/tabel/arsip yang saling berhubungan dan tersimpan dalam media penyimpanan elektronik, untuk kemudahan dalam pengaturan, pemilahan, pengelompokan dan pengorganisasian data sesuai tujuan.

Dengan basis data seseorang dapat menyimpan sebuah informasi, seperti data mahasiswa, kepegawaian atau produk ke dalam media penyimpanan elektronik seperti cakram magnetis (*disk*) melalui perangkat komputer. Kemudian data-data tersebut dapat digunakan sesuai keperluan.

Sebelum muculnya konsep basis data, untuk organisasi/perusahaan menggunakan pendekatan *flat file* untuk manajemen datanya. Pada pendekatan ini, *user* pada tiap departemen dalam suatu organisasi memiliki program aplikasi sendiri dan setiap program aplikasi menyimpan data mereka masing-masing.

Pendekatan *flat file* ini memiliki berbagai kelemahan, antara lain:

1. menghasilkan biaya penyimpanan tinggi untuk dokumen kertas/magnetik form yang digunakan;
2. adanya kerangkapan data, dimana data yang sama disimpan di berbagai tempat penyimpanan yang berbeda sehingga organisasi memiliki banyak data yang rangkap;

3. perubahan (*update*) terhadap data harus dilakukan berulang kali, mengingat data yang sama terdapat di berbagai tempat penyimpanan;
4. dapat memiliki potensi masalah jika gagal untuk memperbarui data pada semua *file* yang terpengaruh;
5. pengguna tidak dapat memperoleh tambahan informasi saat membutuhkan informasi tambahan.

Pendekatan basis data dilakukan dengan mengubah model penyimpanan data yang ada pada pendekatan *flat file*, dimana data disimpan di setiap departemen menjadi terkumpul dalam satu basis data yang dapat dipakai secara bersama oleh seluruh pengguna dalam organisasi/perusahaan. Pendekatan basis data memberikan banyak keuntungan (kelebihan), antara lain berikut ini.

1. Pengendalian terhadap kerangkapan data
   Caranya dilakukan dengan data hanya disimpan sekali. Hal ini mengurangi kerangkapan data dan megurangi biaya untuk tempat penyimpanan.
2. Konsistensi data
   Data disimpan hanya sekali dalam basis data sehingga jika terjadi perubahan pada nilai data tersebut, maka perubahan hanya dilakukan satu kali dan nilai baru tersebut akan tersedia untuk semua pengguna.
3. Dapat memperoleh lebih banyak informasi dari data yang sama.
   Pengguna basis data dapat memperoleh informasi selain dari informasi rutin yang dikelolanya karena semua data lain berada dalam basis data yang sama. Dengan demikian kebutuhan akan informasi selain dari informasi rutin dapat terpenuhi.
4. Data dapat dipakai secara bersama (*shared*)
   Data yang ada pada basis data menjadi milik seluruh organisasi dan dapat dipakai secara bersama oleh pengguna yang berwenang pada saat bersamaan.
5. Memperbaiki integritas data
   Intergritas data mengacu pada validitas dan konsistensi dari data yang disimpan. Integritas biasanya diekspresikan dalam batasan (*constraints*) yang merupakan aturan yang konsisten dan tidak dapat dilanggar. Jika kerangkapan data dikontrol dan kekonsistenan data dapat dijaga maka data menjadi akurat.
6. Meningkatkan keamanan data
   Keamanan basis data melindungi basis data dari pengguna yang tidak memiliki otorisasi. DBA (*Data Base Administration*) dapat menentukan batasan-batasan pengaksesan data, misalnya dengan memberikan *password* dan pemberian hak akses bagi pemakai (misal: *modify, delete, insert, retrieve*).
7. *Economy of scale*
   Dengan menggabungkan semua data operasional organisasi ke dalam satu basis data dengan aplikasi yang dibutuhkan dapat menghasilkan penghematan biaya. Anggaran yang biasanya dialokasikan ke setiap departemen untuk pengembangan dan pemeliharaan dari sistem file mereka dapat digabung sehingga menurunkan total biaya dan menciptakan *economy of scale*.

8. Meningkatkan aksesibilitas terhadap data dan respon yang lebih baik
   Akibat dari integrasi data yang melewati batasan-batasan departemen dapat langsung diakses oleh pengguna. Ini berarti menyediakan sistem dengan fungsi yang lebih baik. Pengguna dapat memperoleh data yang dibutuhkan dengan cepat dengan menggunakan *query language*.
9. Dapat meningkatkan data *independence* (kemandirian data). Dapat digunakan untuk bermacam-macam program aplikasi tanpa harus merubah format data yang sudah ada

Latihan

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

1) Sebutkan dan jelaskan secara singkat bidang ilmu yang terkait erat dengan data mining!
2) Jelaskan bagaimana peran bidang ilmu statistik dalam mendukung data mining!

Petunjuk Jawaban Latihan

1) Terdapat empat buah bidang ilmu yang mendasari yaitu Statistik, Kecerdasan Buatan, Pengenalan Pola, dan Basis Data.
2) Dengan menggunakan statistik, data yang diolah oleh teknik data mining dapat diringkas sebagai *Explanatory Data Analysis* (EDA). EDA berguna untuk mengidentifikasi hubungan sistematis antara variabel/fitur ketika tidak ada cukup informasi alami yang dibawanya.

Rangkuman

Data Mining bertujuan untuk memanfaatkan data dalam basis data dengan mengolahnya sehingga menghasilkan informasi baru yang berguna. Jika dilacak dari akar keilmuwannya, Data Mining memiliki 4 buah bidang ilmu yang mendasari yatu Statistik, Kecerdasan Buatan, Pengenalan Pola, dan Basis Data. Kecerdasan buatan merupakan bidang ilmu yang digunakan untuk mengetahui dan memodelkan proses–proses berpikir manusia dan mendesain mesin agar dapat menirukan perilaku manusia Pengenalan pola merupakan bidang ilmu yang bertujuan menentukan kelompok atau kategori pola berdasarkan ciri-ciri yang dimiliki oleh pola tersebut. Dengan kata lain, pengenalan pola membedakan suatu objek dengan objek lain. Basis data merupakan himpunan kelompok data yang saling terhubung dan diorganisasi sedemikian rupa supaya kelak dapat dimanfaatkan kembali secara cepat dan mudah.

Tes Formatif 3

Pilihlah satu jawaban yang paling tepat!

1) EDA dalam statistik adalah singkatan dari ….
   A. *Explanatory Data Analysis*
   B. *Expository Data Analysis*
   C. *Explanatory Data Algorithm*
   D. *Expository Data Algorithm*

2) Visualisasi data merupakan topik yang terdapat dalam bidang ….
   A. statistika
   B. kecerdasan buatan
   C. pengenalan pola
   D. basis data

3)

Gambar berikut merupakan contoh visualisasi data dengan diagram yang berjenis ….
   A. *Gant chart*
   B. *Scatter Plot*
   C. *Network* Diagram
   D. *Streamgraph*

4) Pada kecerdasan buatan, Motor Inferensi (*Inference Engine*) berperan dalam ….
   A. menyimpan fakta-fakta, teori, pemikiran dan hubungan antar satu data dengan yang lainnya
   B. menarik kesimpulan berdasarkan pengetahuan dan pengalaman
   C. ekstrasi data dari sebuah masalah
   D. visualisasi data

5) Kelebihan dari kecerdasan buatan dibandingkan kecerdasan alami antara lain, *kecuali* ….

A. lebih mudah diduplikasi & disebarkan
B. lebih cepat
C. bersifat tidak permanen
D. dapat didokumentasikan

6) Ciri-ciri yang dimiliki oleh suatu pola ditentukan distribusi statistiknya, merupakan mekanisme pengenalan pola dengan pendekatan ….

A. statistik
B. sintaktik
C. struktural
D. distributif

7) Dalam setiap pendekatan pada pengenalan pola akan melalui dua fase, yaitu ….

A. fase persiapan dan fase pemrosesan
B. fase pelatihan dan fase pengenalan
C. fase pelatihan dan fase pemrosesan
D. fase persiapan dan fase pengenalan

8) Kelemahan pendekatan *flat file* antara lain, *kecuali* ….

A. kerangkapan data
B. konsistensi data
C. biaya mahal
D. potensi kesalahan besar

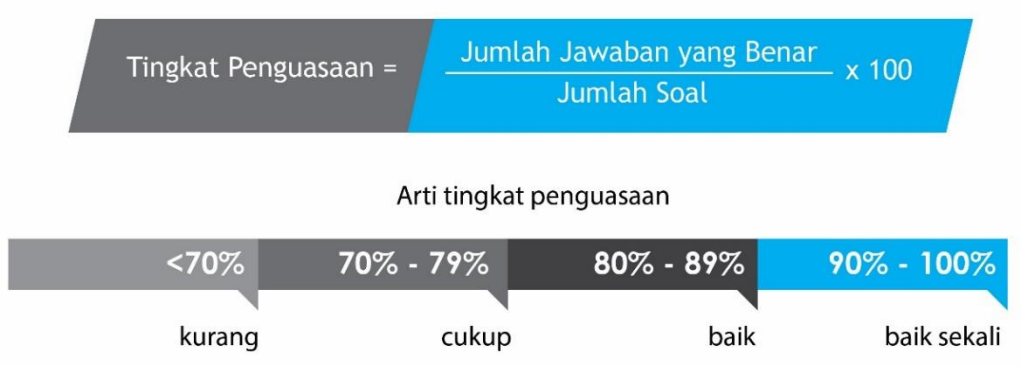
9) Pendekatan pengenalan pola yang lebih mendekati strategi pengenalan pola yang dilakukan oleh manusia adalah ….

A. pendekatan statistik
B. pendekatan sintaktik
C. pendekatan humanis
D. pendekatan algoritmik

10) Validitas dan konsistensi dari data yang disimpan dalam basis data adalah cerminan dari ….

A. integritas data
B. kemutakhiran data
C. kerangkapan data
D. keamanan data

Cocokkanlah jawaban Anda dengan Kunci Jawaban Tes Formatif 3 yang terdapat di bagian akhir modul ini. Hitunglah jawaban yang benar. Kemudian, gunakan rumus berikut untuk mengetahui tingkat penguasaan Anda terhadap materi Kegiatan Belajar 3.



Apabila mencapai tingkat penguasaan 80% atau lebih, Anda dapat meneruskan dengan Kegiatan Belajar 4. **Bagus!** Jika masih di bawah 80%, Anda harus mengulangi materi Kegiatan Belajar 3, terutama bagian yang belum dikuasai.

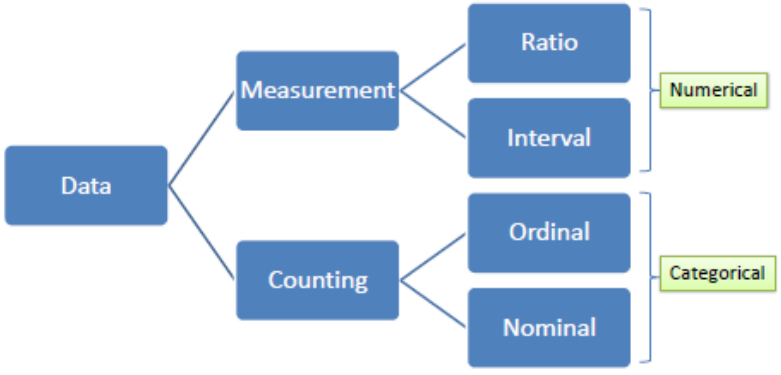
Kegiatan Belajar 4

Teknik Basis Data

Atribut adalah bagian data, yang mewakili karakteristik atau fitur dari objek data. Atribut, dimensi, fitur, dan variabel sering digunakan secara bergantian dalam literatur. Istilah dimensi ini umumnya digunakan dalam literatur data *warehouse*. Dalam literatur *machine learning* cenderung menggunakan istilahfitur, sementara statistik lebih menggunakan istilah variabel.

Data mining dan para profesional *database* biasanya menggunakan istilah atribut, dan disini akan kita gunakan istilah atribut juga. Atribut yang menggambarkan objek pelanggan dapat mencakup, misalnya, ID pelanggan, nama, dan alamat. Nilai-nilai yang diamati untuk atribut tertentu disebut dengan observasi. Satu set atribut yang digunakan untuk menggambarkan suatu objek tertentu disebut atribut vektor (atau fitur vektor). Distribusi data yang melibatkan satu atribut (atau variabel) disebut *univariate*. Kemudian distribusi data yang melibatkan dua atribut disebut *bivariate*, dan seterusnya.

Jenis atribut ditentukan oleh himpunan nilai yang mungkin muncul, yaitu nominal, biner, ordinal, atau numerik. Dalam bagian berikut, kita bahas tentang masing-masing jenis.



Gambar 1.9
Jenis Atribut dalam Data Mining

A. ATRIBUT NOMINAL

Nominal berarti "yang berkaitan dengan nama-nama". Nilai-nilai atribut nominal adalah simbol atau nama-nama dari suatu benda. Setiap nilai merupakan semacam kategori, kode, atau status dan sebagainya sehingga atribut nominal juga disebut sebagai kategorikal. Nilai-nilai di dalamnya tidak memiliki urutan. Dalam ilmu komputer, nilai-nilai tersebut disebut juga dengan *enumerasi*.

Contoh 1.1.

Misalkan warna rambut dan status perkawinan adalah dua atribut yang menggambarkan objek orang. Dalam contoh ini, nilai yang mungkin untuk warna rambut adalah hitam, coklat, pirang, merah, cokelat kemerahan, abu-abu, dan putih. Atribut status perkawinan dapat berisi nilai-nilai *single*, menikah, bercerai, dan janda/duda. Baik warna rambut dan status perkawinan adalah atribut nominal. Contoh lain dari atribut nominal adalah pekerjaan, dengan nilainya berisi misalnya guru, dokter gigi, programmer, petani, dan sebagainya.

Meskipun nilai-nilai dalam atribut nominal adalah simbol atau "nama-nama benda", tetapi bisa saja mewakili simbol atau "nama" dengan angka. Untuk warna rambut, misalnya, kita dapat menetapkan kode 0 untuk hitam, 1 untuk coklat, dan sebagainya. Contoh lain adalah customor ID, dengan nilai yang mungkin saja semuanya berupa angka. Namun, dalam kasus tersebut, angka-angka tersebut tidak dimaksudkan untuk digunakan secara kuantitatif. Artinya, operasi matematika tidak bisa dilakukan pada nilai atribut nominal. Jadi tidak bisa untuk mengurangi satu nomor ID pelanggan dari yang lain, tidak seperti, katakanlah, mengurangkan nilai usia dari yang lain (andaikan usia adalah atribut numerik). Meskipun atribut nominal mungkin saja berisi integer (bilangan bulat), tetapi itu tidak dianggap sebagai atribut numerik karena integer (bilangan bulat) tidak dimaksudkan untuk digunakan secara kuantitatif. Karena nilai atribut nominal tidak memiliki urutan dan tidak kuantitatif, jadi tidak bisa digunakan untuk menghitung nilai rata-rata (*mean*) atau nilai tengah (*median*) dari atribut seperti itu.

B. ATRIBUT BINER

Sebuah atribut biner adalah atribut nominal yang hanya berisi dua jenis nilai saja: 0 atau 1, di mana 0 biasanya berarti bahwa atribut tidak ada, dan 1 berarti bahwa itu ada. Contoh lain adalah atribut yang nilainya hanya berisi 'ya' dan 'tidak'. Atribut biner disebut sebagai *Boolean* jika kedua status berkaitan dengan *true* dan *false*.

Contoh 1.2.

Misalkan atribut 'perokok' yang menggambarkan objek pasien, 1 menunjukkan bahwa pasien merokok, sedangkan 0 menunjukkan bahwa pasien tidak merokok.

Demikian pula, misalkan pasien mengalami tes medis yang memiliki dua hasil yang mungkin. Tes medis atribut biner, di mana nilai 1 berarti hasil tes untuk pasien positif, sedangkan 0 berarti hasilnya negatif.

Sebuah atribut biner adalah simetris jika kedua status dianggap sama-sama penting dan sama-sama memiliki bobot (jumlah *tuple*) yang seimbang; artinya, tidak ada preferensi yang mana yang harus dikodekan dengan 0 atau 1. Salah satu contohnya adalah atribut *gender* yang memiliki status laki-laki dan perempuan.

Sebuah atribut biner asimetris jika salah satu statusnya dianggap lebih penting dan menjadi perhatian dari pengamatan dan memiliki jumlah bobot (jumlah *tuple*) yang tidak seimbang, misalnya hasil yang positif dan negatif dari tes medis untuk HIV. Berdasarkan konvensi, kode status berdasarkan nilai yang dianggap lebih penting, yang biasanya merupakan salah satu yang paling langka, misalnya 1 untuk HIV positif dan 0 untuk HIV negatif.

C. ATRIBUT ORDINAL

Sebuah atribut ordinal adalah atribut dengan nilai-nilai yang memiliki urutan atau peringkat, tapi besaran nilai-nilai yang berurutan tidak diketahui.

Contoh 1.3.

Misalkan ukuran minuman yang tersedia di restoran cepat saji. Atribut nominal ini memiliki tiga kemungkinan nilai: kecil, menengah, dan besar. Nilai-nilai itu memiliki urutan (yang sesuai dengan ukuran minuman); Tetapi, kita tidak bisa mengatakan seberapa lebih besar nilainya, misalkan, yang menengah dibandingkan dengan yang besar. Contoh lain dari atribut ordinal adalah nilai (misalnya, A+, A, A-, B+, dan sebagainya) dan peringkat profesional. Jajaran profesional bisa dibilang (di-*enumerasi*) secara berurutan: misalnya, asisten, *associate*, dan dosen penuh, dan prajurit, prajurit tingkat pertama, spesialis, kopral, dan sersan untuk jajaran militer.

Atribut ordinal bermanfaat dalam mengukur kualitas penilaian subjektif yang tidak dapat diukur secara objektif; sehingga atribut ordinal sering digunakan dalam survei untuk penilaian. Dalam satu survei, pada umumnya partisipan diminta untuk menilai seberapa puas mereka sebagai pelanggan. Kepuasan pelanggan memiliki kategori berikut ordinal: 0: sangat tidak puas, 1: agak tidak puas, 2: netral, 3: puas, dan 4: sangat puas.

Atribut ordinal juga bisa didapatkan dengan teknik diskritisasi dalam tipe atribut numerik dengan memisahkan rentang nilai ke dalam kategori yang berurutan. Nilai-nilai tengah (tendensi sentral) dari atribut ordinal dapat diwakili oleh *modus* dan *median*nya (nilai yang sering muncul dan nilai tengah), tetapi untuk *mean* (rata-rata) tidak bisa dihitung.

Perhatikan bahwa nominal, biner, dan atribut ordinal bersifat kualitatif. Artinya, mereka menggambarkan fitur dari sebuah objek tanpa memberikan ukuran sebenarnya atau kuantitas. Nilai-nilai dari atribut kualitatif tersebut biasanya berupa kata-kata yang mewakili kategori. Jika bilangan bulat yang digunakan, mereka mewakili kode komputer untuk kategori, dan bukan sebagai jumlah yang bisa dihitung (misalnya, 0 untuk ukuran minuman kecil, 1 untuk medium, dan 2 untuk besar).

D. ATRIBUT NUMERIK

Atribut numerik adalah kuantitatif; artinya, nilai atribut itu bisa diukur dan disajikan dalam bentuk integer atau desimal. Atribut numerik bisa berupa *interval-scale* (berskala interval) atau *ratio-scale* (berskala rasio).

1. *Atribut Interval-Scale*

Atribut interval-scale diukur dengan basis skala unit dengan ukuran yang sama. Nilai-nilai atribut *interval-scale* memiliki urutan dan bisa berupa positif, 0, atau negatif. Dengan demikian, selain bisa memberikan urutan nilai-nilai, atribut tersebut memungkinkan kita untuk menghitung perbedaan/selisih antara nilai-nilai itu.

Contoh 1.4.

Atribut temperatur adalah termasuk *interval-scale*. Misalkan kita memiliki beberapa nilai temperatur di luar ruangan untuk beberapa hari yang berbeda-beda, di mana masing-masing hari adalah sebagai objek data. Dengan mengurutkan nilai-nilainya, kita bisa mendapatkan urutan objek berdasarkan nilai temperatur. Selain itu, kita dapat mengukur perbedaan/selisih antara nilai-nilai tersebut. Sebagai contoh, suhu 20^{o}C adalah lima derajat lebih tinggi dari suhu 15^{o}C. Contoh lain adalah pada Kalender atau penanggalan. Misalnya, tahun 2002 dan 2010 adalah delapan tahun selisihnya.

Temperatur dalam Celcius dan Fahrenheit tidak memiliki *'true zero'* atau titik nol absolut, artinya, pada 0^{o}C atau 0^{o}F tidak berarti bahwa "tidak ada temperatur". Meskipun kita dapat menghitung perbedaan/selisih antara nilai temperatur, tetapi kita tidak bisa mengatakan bahwa suatu nilai temperatur adalah sebagai hasil perkalian dari nilai yang lain. Tanpa *'true- zero'* (titik nol absolut), kita tidak bisa mengatakan, misalnya, bahwa 10^{o}C adalah dua kali lebih hangat dari 5^{o}C. Artinya, kita tidak bisa mengatakan nilai-nilai itu dalam bentuk rasio atau perbandingan. Demikian pula, tidak ada *'true-zero'* (nol absolut) untuk kalender penanggalan. Dalam hal ini,Tahun 0 tidak berarti bahwa itu adalah awal dari waktu. Karena atribut *interval-scale* adalah numerik, maka kita bisa menghitung nilai rata-rata (*mean*) dari nilai-nilai tersebut, dan juga menghitung nilai-nilai tengah yang lain seperti *median* (nilai tengah) serta *modus* (nilai yang paling sering muncul).

2. *Atribut Rasio-Scale*

Atribut ratio-scale adalah atribut numerik dengan titik nol absolut. Artinya, jika sistem pengukurannya menggunakan *rasio-scale*, maka kita dapat menghitung perkalian atau perbandingan antara suatu nilai dengan nilai yang lain. Selain itu, nilai-nilai tersebut juga bisa diurutkan, dihitung perbedaan/selisihnya, bisa dihitung *mean* (rata-rata), *median* (nilai tengah), dan *modus* (yang paling sering muncul).

Contoh 1.5.

Pada temperatur, skala Kelvin (K) berbeda dengan skala dalam Celcius dan Fahrenheit. Skala Kelvin (K) memiliki nilai *'true-zero'* (atau titik nol absolut), yaitu 0^o K = $–273.15^o$ C. Ini adalah titik di mana partikel-partikel yang terdiri materi memiliki nol energi kinetik. Contoh lain dari atribut *rasio-scaled* adalah atribut untuk menghitung, misalnya tahun pengalaman (misalnya, objek data adalah karyawan) dan jumlah kata (misalnya, objek data adalah dokumen). Contoh lain adalah atribut untuk mengukur berat badan, tinggi badan, koordinat lintang dan bujur (misalnya, ketika mengelompokkan rumah), dan jumlah moneter (misalnya, seseorang adalah 100 kali lipat lebih kaya bila memiliki $ 100 dibandingkan dengan $ 1).

E. ATRIBUT DISKRIT VS ATRIBUT KONTINU

Sebelumnya, kita telah melihat atribut dalam nominal, biner, ordinal, dan jenis numerik. Ada banyak cara untuk mengorganisasi jenis-jenis atribut. Jenis-jenis atribut tersebut tidak saling eksklusif. Algoritma klasifikasi yang dikembangkan dari bidang disiplin *machine learning* sering membahas atribut sebagai diskrit atau kontinu. Setiap jenis dapat diproses secara berbeda.

Suatu atribut diskrit adalah atribut yang memiliki himpunan nilai-nilai yang berhingga (*finite*) atau nilai-nilai tak-hingga tetapi yang bisa dihitung (*countably infinite*), yang mungkin saja disajikan dalam bentuk integer atau mungkin juga bukan integer. Atribut-atribut seperti warna rambut, perokok, tes kesehatan, dan ukuran minuman ukuran seperti contoh-contoh di atas tadi masing-masing memiliki nilai-nilai yang jumlahnya 'berhingga' (*finite*), jadi atribut- atribut itu adalah diskrit.

Perhatikan bahwa atribut diskrit mungkin memiliki nilai-nilai numerik, seperti 0 dan 1 untuk atribut biner atau, nilai-nilai 0 hingga 110 untuk atribut usia. Suatu atribut adalah tak-hingga tetapi bisa dihitung (*countably infinite*) apabila himpunan nilai-nilainya dapat ditempatkan ke dalam relasi *one-to-one* dengan bilangan alami. Contohnya, atribut ID pelanggan adalah *countably infinite* (tak-hingga tetapi bisa dihitung/dibilang/di-enumerasi). Jumlah pelanggan dapat tumbuh hingga tak-hingga, tetapi dalam kenyataannya, kumpulan nilai-nilainya bisa dihitung/dibilang (dimana nilai-nilainya dapat ditempatkan ke dalam relasi *one-to-one* dengan himpunan bilangan bulat). Contoh lain adalah kode pos.

Bila suatu atribut tidak diskrit, berarti atribut tersebut kontinu (*continous*). Istilah atribut numerik dan atribut kontinu sering digunakan secara bergantian dalam literatur. (Hal ini dapat membingungkan karena, dalam pengertian klasik, nilai-nilai kontinu adalah bilangan real, sedangkan nilai numerik dapat berupa integer atau bilangan real.) Dalam prakteknya, nilai real disajikan dalam bentuk angka-angka. Atribut kontinu biasanya disajikan sebagai *variabel floating-point* (desimal).

Latihan

Untuk memperdalam pemahaman Anda mengenai materi di atas, kerjakanlah latihan berikut!

1) Apa yang dimaksud dengan atribut?
2) Sebutkan dan jelaskan jenis-jenis atribut dalam data mining!

Petunjuk Jawaban Latihan

1) Atribut adalah bagian data, yang mewakili karakteristik atau fitur dari objek data. Atribut, dimensi, fitur, dan variabel sering digunakan secara bergantian dalam literatur. Istilah dimensi ini umumnya digunakan dalam literatur datawarehouse. Dalam literatur, biasanya *machine learning* cenderung menggunakan istilah fitur, sementara statistik lebih menggunakan istilah variabel.
2) Jenis atribut pada data mining antara lain : nominal, biner, ordinal, atau numerik. Nilai-nilai atribut nominal adalah simbol atau nama-nama dari suatu benda. Setiap nilai merupakan semacam kategori, kode, atau status dan sebagainya sehingga atribut nominal juga disebut sebagai kategorikal. Sebuah atribut biner adalah atribut nominal yang hanya berisi dua jenis nilai saja: 0 atau 1, di mana 0 biasanya berarti bahwa atribut tidak ada, dan 1 berarti bahwa itu ada. Atribut ordinal adalah atribut dengan nilai-nilai yang memiliki urutan atau peringkat, tapi besaran nilai-nilai yang berurutan tidak diketahui. Atribut numerik adalah kuantitatif; artinya, nilai atribut itu bisa diukur, disajikan dalam bentuk integer atau desimal. Atribut numerik bisa berupa *interval-scale* (berskala interval) atau *ratio-scale* (berskala rasio).

Rangkuman

Atribut adalah bagian data, yang mewakili karakteristik atau fitur dari objek data. Atribut, dimensi, fitur, dan variabel sering digunakan secara bergantian dalam literatur. Satu set atribut yang digunakan untuk menggambarkan suatu objek tertentu disebut atribut vektor (atau fitur vektor). Distribusi data yang melibatkan satu atribut (atau variabel) disebut *univariate*. Distribusi *bivariate* melibatkan dua atribut, dan seterusnya.

Jenis atribut pada data mining antara lain: nominal, biner, ordinal, atau numerik. Nilai-nilai atribut nominal adalah simbol atau nama-nama dari suatu benda. Setiap nilai merupakan semacam kategori, kode, atau status dan sebagainya sehingga atribut nominal juga disebut sebagai kategorikal. Sebuah atribut biner adalah atribut nominal yang hanya berisi dua jenis nilai saja: 0 atau 1, di mana 0 biasanya berarti bahwa atribut tidak ada dan 1 berarti bahwa atributnya ada. Atribut ordinal adalah atribut dengan nilai-nilai yang memiliki urutan atau peringkat, tapi besaran nilai-nilai yang berurutan tidak diketahui. Atribut numerik adalah kuantitatif; artinya, nilai atribut itu bisa diukur dan disajikan dalam bentuk integer atau desimal. Atribut numerik bisa berupa *interval-scale* (berskala interval) atau *ratio-scale* (berskala rasio).

Tes Formatif 4

Pilihlah satu jawaban yang paling tepat!

1) Bagian dari data yang mewakili karakteristik atau fitur dari objek data, disebut dengan ….
   A. kelengkapan data
   B. atribut data
   C. identitas data
   D. nilai data

2) Satu set atribut yang digunakan untuk menggambarkan suatu objek tertentu disebut ….
   A. atribut set
   B. atribut himpunan
   C. atribut vektor
   D. set objek

3) Sebuah atribut biner adalah simetris jika ….
   A. kedua nilai dianggap sama-sama penting dan sama-sama memiliki bobot (jumlah *tuple*) yang seimbang
   B. kedua nilai dianggap memiliki tingkat kepentingan berbeda dan sama-sama memiliki bobot (jumlah *tuple*) yang seimbang
   C. salah satu nilai dianggap lebih penting dan menjadi perhatian dari pengamatan dan memiliki jumlah bobot (jumlah *tuple*) yang tidak seimbang
   D. kedua nilai dianggap sama-sama penting dan menjadi perhatian dari pengamatan dan memiliki jumlah bobot (jumlah *tuple*) yang tidak seimbang

4) Atribut status perkawinan yang dapat berisi nilai-nilai *single*, menikah, bercerai, dan janda/duda, berjenis atribut ….
   A. nominal
   B. ordinal
   C. biner
   D. rasio

5) Kesamaan dari atribut berjenis nominal, ordinal, dan biner adalah ….
   A. atribut memiliki urutan yang bermakna
   B. berupa data kualitatif
   C. berupa data kuantitif
   D. dapat dioperasikan secara matematis

6) Contoh data yang merupakan *interval scaled* adalah data ….
   A. tanggal lahir
   B. nomor telepon
   C. tinggi badan
   D. berat badan

7) Atribut yang memiliki himpunan nilai-nilai yang berhingga (*finite*) atau nilai-nilai tak-hingga tetapi yang bisa dihitung (*countably infinite*) adalah atribut ….
   A. diskrit
   B. kontinu
   C. numerik
   D. abstrak

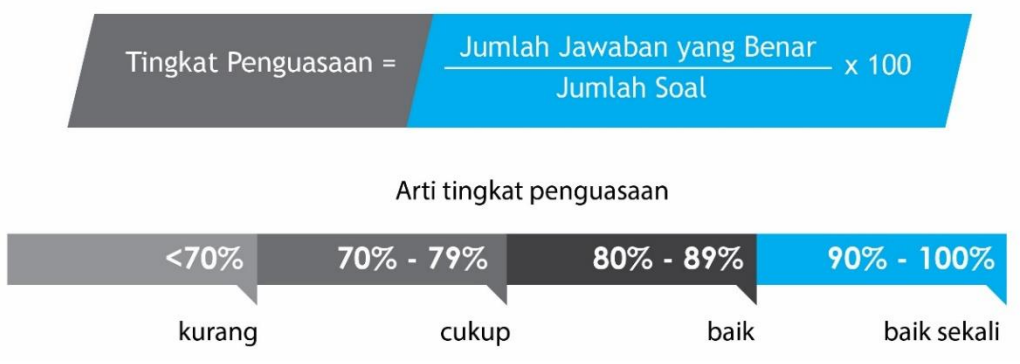
8) Atribut dengan nilai-nilai yang memiliki urutan atau peringkat, tapi besaran nilai-nilai yang berurutan tidak diketahui adalah atribut ….
   A. ordinal
   B. nominal
   C. interval
   D. rasio

9) Contoh dari atribut biner asimetris adalah ….
   A. jenis kelamin : laki-laki atau perempuan
   B. status jawaban : ya atau tidak
   C. status HIV : positif atau negatif
   D. status bilangan : ganjil, genap

10) Atribut yang bisa diukur dan disajikan dalam bentuk integer atau desimal, adalah ciri dari atribut ….

A. ordinal
B. kuantitatif
C. kualitatif
D. biner

Cocokkanlah jawaban Anda dengan Kunci Jawaban Tes Formatif 4 yang terdapat di bagian akhir modul ini. Hitunglah jawaban yang benar. Kemudian, gunakan rumus berikut untuk mengetahui tingkat penguasaan Anda terhadap materi Kegiatan Belajar 4.



Apabila mencapai tingkat penguasaan 80% atau lebih, Anda dapat meneruskan dengan modul selanjutnya. **Bagus!** Jika masih di bawah 80%, Anda harus mengulangi materi Kegiatan Belajar 4, terutama bagian yang belum dikuasai.

Kunci Jawaban Tes Formatif

Tes Formatif 1		*Tes Formatif 2*		*Tes Formatif 3*		*Tes Formatif 4*	
1)	B	1)	A	1)	A	1)	B
2)	C	2)	C	2)	A	2)	C
3)	C	3)	D	3)	C	3)	A
4)	C	4)	A	4)	B	4)	A
5)	D	5)	C	5)	C	5)	B
6)	B	6)	C	6)	A	6)	A
7)	A	7)	A	7)	B	7)	A
8)	D	8)	D	8)	B	8)	A
9)	C	9)	A	9)	B	9)	C
10)	B	10)	C	10)	A	10)	B

Daftar Pustaka

Connoly, T. M., Begg, C. E., dan Strachan, A. D. (1999). *Database system. A practical approach to design, implementation, and management*. Addison Wesley Company.

Witten, I.H., Frank, E., dan Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Elsevier.

Han, J., Kamber, M., dan Pei, J. (2012). *Data mining: concepts and techniques.* (3rd ed.). Elsevier.

Kristanto, A. (2004). *Kecerdasan buatan*, Yogyakarta: Graha Ilmu.

Kusumadewi, S. (2003). *Artificial intelligence (teknik dan aplikasinya).* Yogyakarta: Graha Ilmu.

Kuswadi, S. (2007). *Kendali cerdas (teori dan aplikasi praktisnya)*. Yogyakarta: Penerbit ANDI.

Tan, P. N., Steinbach, M., Kumar, V. (2015). *Introduction to data mining* (2nd ed.). Pearson Education, Inc.