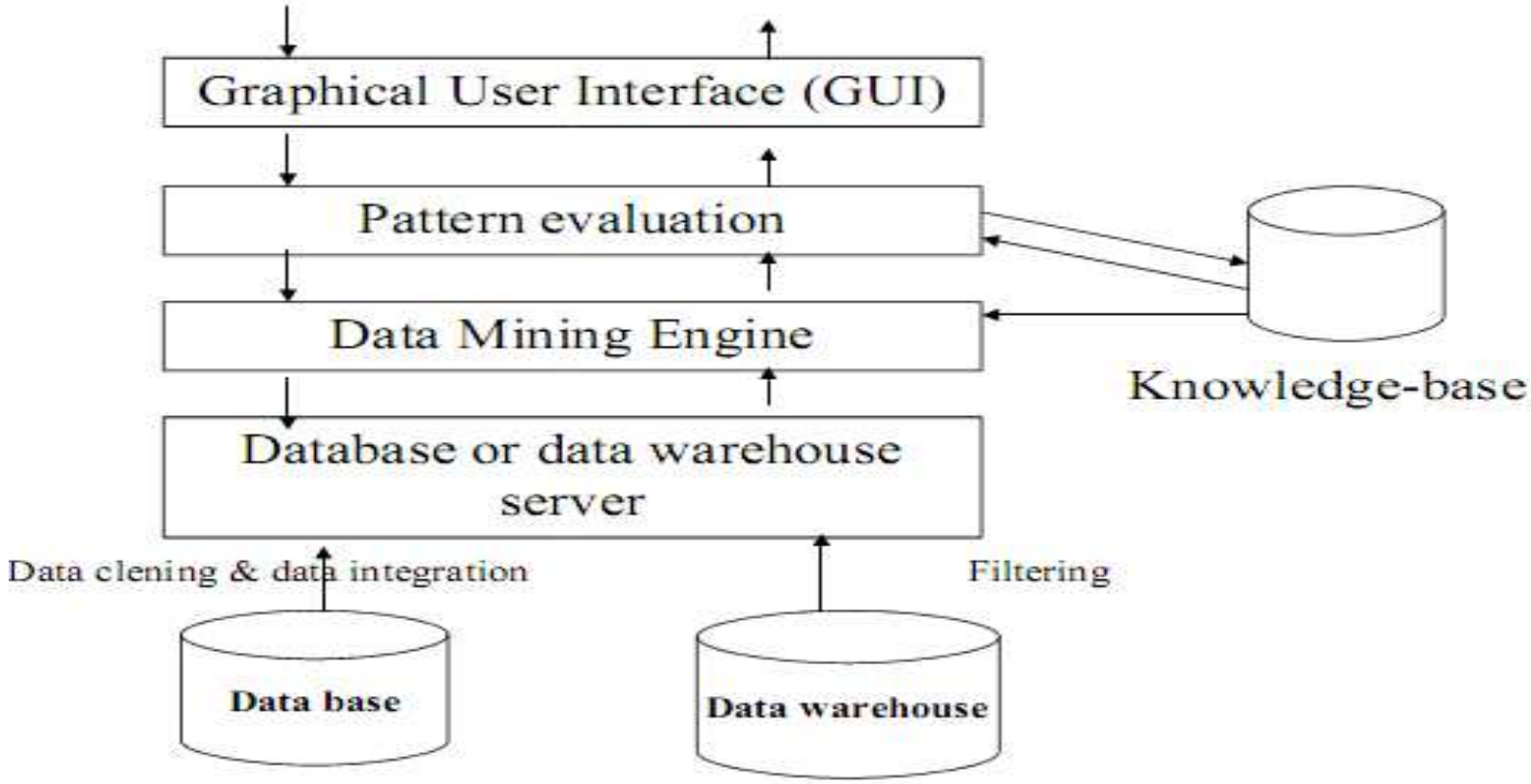
PERTEMUAN 13

ARSITEKTUR
&
MODEL DATA MINING

Arsitektur : Sistm Data Mining



Keterangan :

1. Data cleaning (Pembersihan Data) : untuk membuang data yang tidak konsisten dan noise)
2. Data integration : penggabungan data dari beberapa sumber
3. Data Mining Engine : Mentranformasikan data menjadi bentuk yang sesuai untuk di mining
4. Pattern evaluation : untuk menemukan yang bernilai melalui knowledge base
5. Graphical User Interface (GUI) : untuk end user

Semua tahap bersifat interaktif di mana user terlibat langsung atau dengan perantaraan knowledge base

Model Data Mining

- Prediction Methods

  Menggunakan beberapa variabel untuk memprediksi sesuatu atau suatu nilai yang akan datang.

- Description Methods

  Mendapatkan pola penafsiran (humaninterpretable patterns) untuk menjelaskan data.

Data Mining

Prediktif

- Klasifikasi
- Decision tree
- Analisis Time series
- Regresi
- Prediksi
- Jaringan syaraf tiruan
- Data Mining

Deskriptif

- Klastering
- Summarization
- Aturan Asosiasi
- (Assosiation Rule)
- Sequence Discovery

Klasifikasi

- Proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk dapat memprediksi kelas dari suatu objek yang labelnya tidak diketahui
- Contoh : Mendeteksi Penipuan
- Tujuan : Memprediksi kasus kecurangan transaksi kartu kredit.
- Pendekatan :

  - Menggunakan transaksi kartu kredit dan informasi dilihat dari atribut account holder

- Kapan customer melakukan pembelian, Dengan cara apa customer membayar, sebarapa sering customer membayar secara tepat waktu, dll

- Beri nama/tanda transaksi yang telah dilaksanakan sebagai transaksi yang curang atau yang baik. Ini sebagai atribut klass (the class attribute.)

- Pelajari model untuk class transaksi

- Gunakan model ini untuk mendetekdi kecurangan dengan meneliti transaksi kartu kredit pada account.

Regression

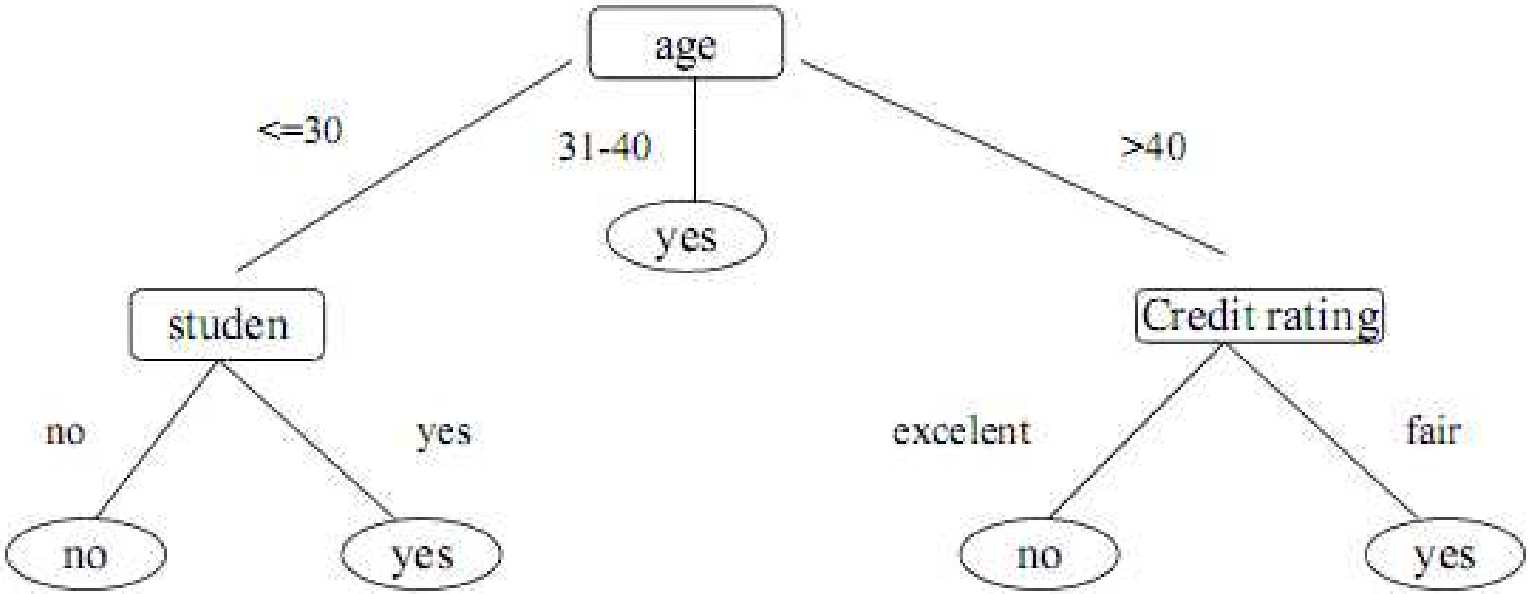
Digunakan untuk memetakan data dengan prediksi

atribut bernilai real

Contoh:

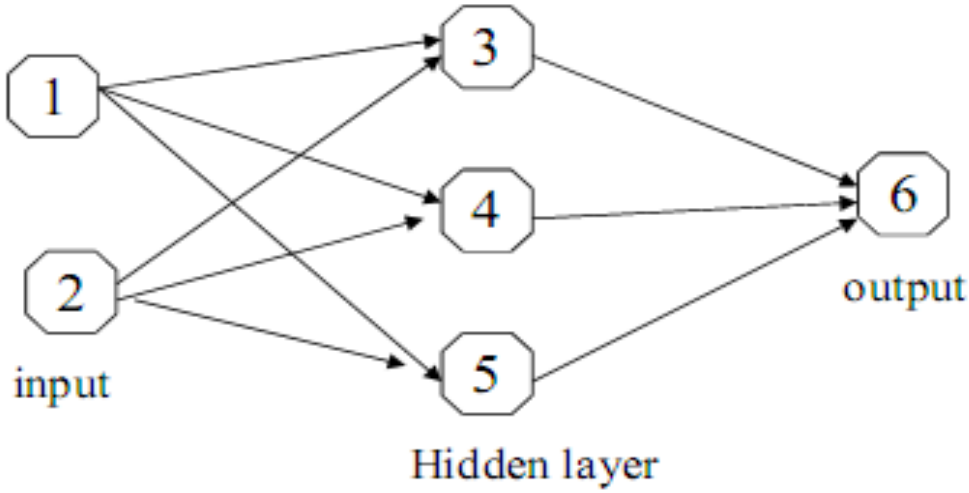
– Memprediksi jumlah penjualan produk baru pada advertising expenditure.
– Memprediksi kecepatan memutar (wind velocities) pada fungsi temperatur, tekanan udara , dll

Decision tree (Pohon keputusan)
Salah satu model klasifikasi yang mudah di interpretasikan
Contoh : identifikasi pembeli komputer (dari decision tree di bawah ini ternyata salah satu kelompok yang potensial adalah orang yang berusia < 30 dan pelajar



Neural Network (Jaringan syaraf tiruan)

- Jaringan syaraf buatan di mulai dengan layer input, dimana tiap simpul berkorespondensi dengan variabel prediktor.
- Simpul- simpul input ini terhubung kebeberapa simpul dalam hidden layer. Dan simpul dalam hidden layer dapat terhubung ke simpul lain dalam hidden layer atau ke output layer.
- Output layer terdiri dari satu atau beberapa variable respon



- Telekomunikasi

  Data mining digunakan untuk melihat jutaan transaksi yang masuk dengan tujuan menambah layanan otomatis

- Keuangan

  Data mining digunakan untuk mendeteksi transaksitransaksi keuangan yang mencurigakan dimana akan susah dilakukan jika menggunakan analisis standar.

- Asuransi

  Australian Health Insurance Commision menggunakan data mining untuk mengidentifikasi layanan kesehatan dan berhasil menghemat satu juta dollar pertahun

- Olah raga

  IBM Advanced Scout menggunakan data mining untuk menganalisis statistik permainan NBA dalam rangka competitive advantage untuk tim New York Knicks

- Astronomi

  Jet Propulsion Laboratory (JPL) di Pasadena dan Pulomar Observatory menemukan 22 quasar dengan bantuan data mining.

- Internet Web Surf-Aid

  IBM Surf-Aid menggunakan algoritma data mining untuk mendata akses halaman Web khususnya berkaitan dengan pemasaran melalui web.

Tools Data Mining

- Karateristik-karateristik penting dari tool data mining meliputi :
  - Data preparation facilities
  - Selection of data mining operation (algorithms)
  - Product scalability and performance
  - Facilities for visualization of result

- Data mining tool, meliputi :
  - Integral Solution Ltd's Clementine
  - DataMind Corp's Data Crusher
  - IBM's Intelligent Miner
  - Silicon Graphics Inc.'s MineSet
  - Informations Discovery Inc.'s Data Mining Suite
  - SAS Institute Inc.'s SAS System and Right Information System'Thought.

Evolusi Database

- Th 1960
  – Pengumpulan data, pembuatan data, IMS dan network DBMS
- Th 1970
  – Model data relasional, Implementasi DBMS relasional
- Th 1980
  – RDBMS, Model data lanjutan (extended-relational, OO, deductive)
- Th 1990
  – Data mining, data warehouse, database multimedia, dan Web database.
- Th 2000
  – Stream data managemen dan mining – Data mining dengan berbagai variasi aplikasi – Teknologi web dan sistem informasi global

Teknik – teknik Database

Searching

- Searching dilakukan untuk memeriksa serangkaian item yang memiliki sifatsifat yang diinginkan.
- Tindakan untuk menemukan suatu item tertentu baik yang diketahui keberadaannya maupun tidak.
- Memasukkan kata dalam suatu program komputer untuk membandingkan dengan informasi yang ada dalam database.

Indexing

- Indexing adalah struktur-struktur akses yang digunakan untuk mempercepat respon dalam mendapatkan record-record pada kondisi-kondisi pencarian tertentu.
- Indexing field adalah suatu struktur akses index yang biasanya menjelaskan field tunggal dari suatu file.
- Indexing organization memberikan efisiensi akses ke record-record secara berurut atau random.

Data Reduction

- *Data reduction adalah transformasi suatu masalah ke masalah lain dan dapat digunakan untuk mendefinisikan serangkaian masalah yang kompleks.*

- *Data reduction merupakan teknik yang digunakan untuk mentransformasi dari data mentah ke bentuk format data yang lebih berguna. Sebagai contoh groupping, summing dan averaging data.*

- *Data reduction dilakukan untuk mengatasi ukuran data yang terlalu besar. Ukuran data yang terlalu besar dapat menimbulkan ketidakefisienan proses dan peningkatan biaya pemrosesan.*

- *Data reduction dilakukan dalam tahap data preprocessing pada rangkaian proses Knowledge Discovery Databases (KDD) sebelum data mining dengan tujuan mengurangi ukuran data yang besar.*

OLAP (On-line analytical processing)

- OLAP adalah suatu sistem atau teknologi yang dirancang untuk mendukung proses analisis kompleks dalam rangka mengungkapkan kecenderungan pasar dan faktor-faktor penting dalam bisnis

- OLAP ditandai dengan kemampuannya menaikkan atau menurunkan dimensi data sehingga kita dapat menggali data sampai pada level yang sangat detail dan memperoleh pandangan yang lebih luas mengenai objek yang sedang kita analisis.

- OLAP secara khusus memfokuskan pada pembuatan data agar dapat diakses pada saat pendefinisian kembali dimensi.

- OLAP dapat digunakan membuat rangkuman dari multidimensi data yang berbeda, rangkuman baru dan mendapatkan respon secara online, dan memberikan view dua dimensi pada data cube multidimensi secara interaktif.

OLAP (ONLINE ANALYTICAL PROCESSING)

Aplikasi OLAP didominasi oleh ad hoc, query kompleks. Dalam istilah SQL, ini adalah query yang melibatkan kelompok-oleh dan operator agregasi.Cara alami untuk berpikir tentang query OLAP adalah dalam hal model data multidimensi.Kita mulai bagian ini dengan menyajikan model data multidimensi dan membandingkannya dengan representasi data relasional.

MODEL DATA MULTIDIMENSIONAL

- Dalam model data multidimensi, fokusnya adalah pada koleksi **langkah-langkah** numerik.Setiap ukuran tergantung pada set **dimensi.**
- Beberapa sistem OLAP, misalnya, Essbase dari Software Arbor, sebenarnya menyimpan data dalam array multidimensi. Sistem OLAP yang menggunakan array untuk menyimpan dataset multidimensi disebut **OLAP multidimensi (MOLAP)** sistem.

OLAP QUERY

- Operasi yang didukung oleh model ini sangat dipengaruhi oleh alat pengguna akhir seperti spreadsheet. Tujuannya adalah untuk memberikan pengguna akhir yang bukan ahli SQL antarmuka yang intuitif dan kuat untuk umum tugas analisis businessoriented. Pengguna diharapkan untuk menimbulkan ad hoc query secara langsung, tanpa bergantung pada programmer aplikasi database.