DATA CLEANING, DATA PREPROCESSING, SERTA PENENTUAN ATRIBUT DAN DATA

Dalam Data Mining

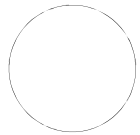
PENDAHULUAN

- Data mining digunakan untuk penyelesaian suatu masalah atau pengambilan keputusan berdasarkan informasi dalam database.
- Pola atau hasil prediksi terhadap suatu parameter diperoleh dari perhitungan-perhitungan data mining menggunakan berbagai metode dan algoritma.

PERMASALAHAN

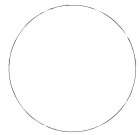
- Data yang terkoleksi masih kotor
  - Incomplete : tidak ada atribut, tidak lengkap
  - Noise : terdapat error di dalam data tersebut
  - Inconsistency : tidak konsisten antara data yang satu dengan yang lain

Jika data tidak bersih maka hasil mining tidak bisa dipastikan kualitasnya



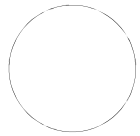
Data cleaning

- Data cleaning tasks
  - Akuisisi data
  - Melengkapi nilai yang hilang
  - Menyeragamkan format
  - Konversi data nominal ke numerik atau numerik ke nominal
  - Membenahi data yang tidak konsisten



Data akuisisi

- Data bisa disimpan di DBMS,
- Data di dalam flat file
  - Fixed format
  - Delimited format : tab, comma, “, dll
- Verifikasi nomor urut suatu kolom



Contoh

- **Original data (fixed column format)**

00000000130.06.19971979-10-3080145722 #000310 111000301.01.00010000000004
000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.0000
0000000000. 000000000000000.000000000000000.0000000......
000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.00
000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.0000
0000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000000000000.000000
00000000.000000000000000.000000000000000.000000000000000.00 000000000300.00 000000000300.00

- **Clean data**

0000000001,199706,1979.833,8014,5722 , ,#000310 ....
,111,03,000101,0,04,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0300,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0300,0300.00

REFORMATTING

- Mengkonfersi data kedalam bentuk yang standar cth : csv, arff
- *Missing data*, karena:
  - Kerusakan peralatan
  - Tidak konsisten terhadap data yang lain, sehingga dihapus
  - Data tidak dimasukkan akibat kesalahpahaman
  - Data tertentu dianggap tidak penting pada waktu menginputkan
  - Tidak dapat menunjukkan riwayat atau perubahan data

PROBLEM SOLVING

Contoh studi kasus:

- Seorang siswa yang telah lulus SMA berencana akan melanjutkan studi S1 ke sebuah perguruan tinggi (PT). Ada banyak pilihan PT yang akan dituju. Masing-masing PT memiliki kelebihan dan kekurangan, dan siswa tersebut juga memiliki latar belakang pendidikan tertentu.

PROBLEM SOLVING

Pertanyaan:

- Kriteria / variabel apa sajakah yang mempengaruhi pemilihan sebuah PT bagi calon mahasiswa?
- Variabel apakah yang paling signifikan dalam menentukan PT bagi seorang calon mahasiswa?
- Apakah calon siswa dengan kriteria tertentu dapat diterima pada sebuah PT tertentu?

JENIS ATRIBUT / VARIABEL

Sebelum melanjutkan penyelesaian kasus sebelumnya, perlu diketahui jenis-jenis variabel dalam data mining :

- Variabel *Independent* / Bebas (X) → *Regular Attribute*

  Variabel yang nilainya tidak dipengaruhi oleh nilai variabel lainnya.

- Variabel *Dependent* / Terikat (Y) → *Label*

  Variabel yang nilainya dipengaruhi oleh nilai variabel lainnya.

Dalam sebuah studi kasus, jumlah variabel X maupun variabel Y bisa lebih dari satu.

Jenis Data

- CATEGORICAL (Qualitative)
  - Nominal (Named Categories)

    Contoh: merah, kuning, biru, apel, rambutan
  - Ordinal (Categories with an implied order)

    Contoh: rendah, tinggi; kecil, sedang, besar;
- NUMERICAL (Quantitative)
  - Discrete (Only particular numbers)

    Contoh: 0, 1; 1, 2, 3; 10, 20, 50, 100;
  - Continuous (Any numeric value)

    Contoh: 0, 0.25, 1, 2.5, 5 – 10, 25%, 3^4

Data Class

- NUMERIC / INTEGER

  Jika kelas data berupa bilangan bulat (tanpa pengelompokan)
- REAL

  Jika kelas data berupa angka secara riil (tanpa pengelompokan), termasuk bilangan pecahan
- BINOMINAL

  Jika data hanya terdiri dari 2 kelas
- POLYNOMINAL

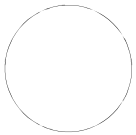
  Jika data terdiri lebih dari 2 kelas

Penggunaan Data

- Data Pelatihan (*Training Data*)
  - Data yang digunakan untuk melatih sebuah sistem untuk menemukan pola-pola tertentu atau informasi terhadap suatu data.
  - Data pelatihan diambil dari data yang sudah terjadi di masa lalu.
- Data Uji (*Testing Data*)
  - Data yang digunakan untuk menguji tingkat keberhasilan dari suatu proses pencarian informasi terhadap data pelatihan.
  - Biasanya data uji diambil dari data yang belum terjadi (prediksi).

Data Collecting

- Populasi
- Sampel
- Sensus
- Survei



Studi Kasus

Contoh penentuan kriteria / variabel yang mempengaruhi pemilihan sebuah PT bagi calon mahasiswa.

1. Jenis PT
2. Biaya kuliah (SPP)
3. Biaya hidup selama kuliah
4. Keputusan Memilih

Dll.

STUDI KASUS

Penentuan Atribut / Variabel:

- Variabel X
  1. Jenis PT → X1
  2. Biaya kuliah (SPP) → X2
  3. Biaya hidup selama kuliah → X3
- Variabel Y
  1. Keputusan Memilih → Y

Collecting Data (Sampling)

X1	X2	X3	Y
UGM	5 juta/smt	1,5 juta/bln	Ya
UMS	2 juta/smt	2,5 juta/bln	Ya
UMY	6 juta/smt	2 juta/bln	Tidak
UNIBRAW	3 juta/smt	3 juta/bln	Tidak
ITB	5 juta/smt	3 juta/bln	Ya
UII	7 juta/smt	1,5 juta/bln	Tidak
UMJ	5 juta/smt	4 juta/bln	Tidak
Trisakti	7 juta/smt	4 juta/bln	Tidak
UNDIP	6 juta/smt	1,5 juta/bln	Tidak
UNIBRAW	3 juta/smt	2,5 juta/bln	Ya

STUDI KASUS

Penentuan Jenis Data:

1. Jenis PT (X1) → Nominal
2. Biaya kuliah (SPP) (X2) → Ordinal
3. Biaya hidup selama kuliah (X3) → Ordinal
4. Keputusan Memilih (Y) → Nominal

STUDI KASUS

Penentuan Kelas Data:

1. Jenis PT (X1) → Binominal

   Isi data = PTN; PTS

2. Biaya kuliah (SPP) (X2) → Polynominal

   Isi data =

   Murah, jika X3 ≤ 3 juta/smt;

   Sedang, jika 3 juta/smt < X3 < 6 juta/smt

   Mahal, jika X3 ≥ 6 juta/smt

STUDI KASUS

Penentuan Kelas Data:

4. Biaya hidup selama kuliah (X3) → Polynominal

   Isi data =

   Murah, jika X4 < 2 juta/bln;

   Sedang, jika 2 juta/bln ≤ X4 < 3 juta/bln

   Mahal, jika X4 ≥ 3 juta/bln

5. Keputusan Memilih (Y) → Binominal

   Isi data = Ya; Tidak

Contoh Data Training

X1	X2	X3	Y
PTN	Sedang	Murah	Ya
PTS	Murah	Sedang	Ya
PTS	Mahal	Sedang	Tidak
PTN	Murah	Mahal	Tidak
PTN	Sedang	Mahal	Ya
PTS	Mahal	Murah	Tidak
PTS	Sedang	Mahal	Tidak
PTS	Mahal	Mahal	Tidak
PTN	Mahal	Murah	Tidak
PTN	Murah	Sedang	Ya

Contoh Data Testing

X1	X2	X3	Y
PTN	Sedang	Mahal	?
PTS	Mahal	Murah	?
PTS	Sedang	Mahal	?

Format Data

- Database
- CSV
- Excel
- Access DB
- ARFF
- SPSS
- URL

Dan lain-lain.