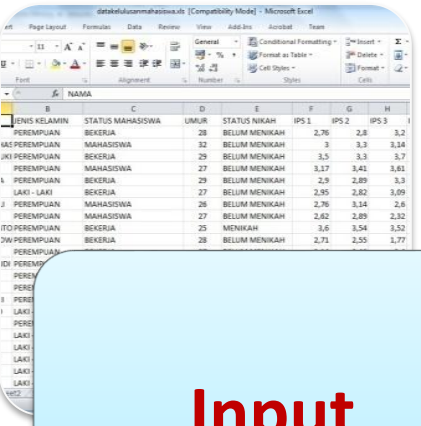
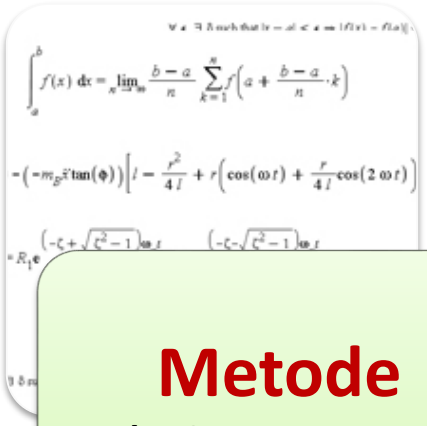
Dataset

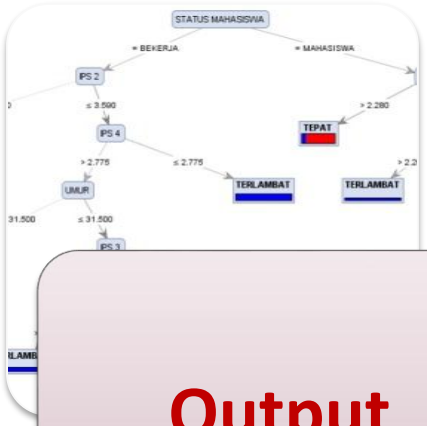
Proses Utama pada Data Mining



Input
(Data)



Metode
(Algoritma
Data Mining)



Output
(Pola/Model)

Dataset

- Obyek

  (kasus,record,titik)
- Atribut

  (fitur,field,karakteristik,variabel)

Atribut, Class dan Tipe Data

- Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi
- Class adalah atribut yang akan dijadikan target, sering juga disebut dengan label
- Tipe data untuk variabel pada statistik terbagi menjadi empat: nominal, ordinal, interval, ratio
- Tapi secara praktis, tipe data untuk atribut pada data mining hanya menggunakan dua:
  1. **Kategorikal** (Nominal,Ordinal)
  2. **Numeric** (Real,Integer)

Atribut Kategorikal

- Nominal (Distinctness)
- Ordinal(Distinctness &Order)

m_nim	jns_klmn	agama	kota_asal	pendidikan_ortu
A12.2014.05109	1	1	KABUPATEN KENDAL	D3
A12.2014.05106	1	1	KABUPATEN TEGAL	S1
A12.2014.05107	1	2	KOTA SEMARANG	S1
A12.2014.05105	1	1	KABUPATEN PATI	SMA
A12.2014.05104	1	1	KOTA SEMARANG	SD
A12.2014.05103	1	1	KABUPATEN GROBOGAN	SD
A12.2014.05102	1	1	KOTA SEMARANG	SD
A12.2014.05100	1	1	KABUPATEN KENDAL	SMA
A12.2014.05101	1	1	KABUPATEN SEMARANG	S2
A12.2014.05098	1	1	KABUPATEN KENDAL	S2
A12.2014.05099	0	1	KOTA SEMARANG	S1
A12.2014.05097	1	1	KABUPATEN REMBANG	SMA
A12.2014.05096	0	1	KOTA SEMARANG	-
A12.2014.05094	1	1	KABUPATEN DEMAK	SMA
A12.2014.05095	1	1	KABUPATEN REMBANG	D3
A12.2014.05092	1	1	KOTA SEMARANG	SMA
A12.2014.05093	0	1	KABUPATEN PATI	-
A12.2014.05090	0	1	KOTA SEMARANG	SMP

Atribut Numeric

- Discrete
- Continue

m_nim	sks ditempuh	IPK
A12.2014.05109	20	2.6
A12.2014.05106	24	3.6
A12.2014.05107	22	2.9
A12.2014.05105	22	2.8
A12.2014.05104	24	3.7
A12.2014.05103	22	2.8
A12.2014.05102	24	2.6
A12.2014.05100	24	3.6
A12.2014.05101	22	2.9
A12.2014.05098	20	2.6
A12.2014.05099	24	3.7
A12.2014.05097	24	2.8
A12.2014.05096	24	3.7
A12.2014.05094	24	3.6
A12.2014.05095	20	2.5
A12.2014.05092	24	3.8
A12.2014.05093	24	3.7
A12.2014.05090	18	2.4

Karakteristik Dataset

- Dimensionality
- Sparsity
- Resolution

Tipe Dataset

- Record Data
- Matrix
- Dokumen
- Transaksi
- Graph
- Data Terurut

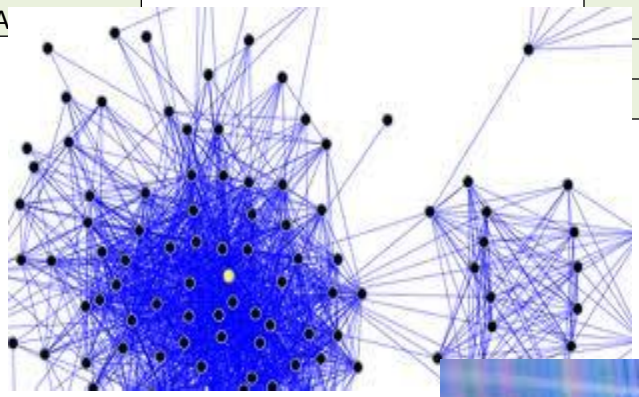
Tipe Dataset

m_nim	jns_klmn	agama	kota_asal
A12.2014.05109	Laki-laki	Islam	KABUPATEN KENDAL
A12.2014.05106	Laki-laki	Islam	KABUPATEN TEGAL
A12.2014.05107	Laki-laki	Kristen	KOTA SEMARANG
A12.2014.05105	Laki-laki	Islam	KABUPATEN PATI
A12.2014.05104	Laki-laki	Islam	KOTA SEMARANG
A12.2014.05103	Laki-laki	Islam	KABUPATEN GROBOGAN
A12.2014.05102	Laki-laki	Islam	KOTA SEMARANG
A12.2014.05100	Laki-laki	Islam	KABUPATEN KENDAL
A12.2014.05101	Laki-laki	Islam	KABUPATEN SEMARANG
A12.2014.05098	Laki-laki	Islam	KABUPATEN KENDAL
A12.2014.05099	Perempuan	Islam	KOTA SEMARA

	Team	Coach	play	ball
Document 1	3	0	5	0
Document 2	0	7	0	2
Document 3	1	6	4	1

TID	Item
1	Roti,Fanta,Susu
2	Aqua,Roti
3	Daging,Softdrink,Susu
4	Aqua,Roti,Susu,popok bayi
5	fanta,popok bayi,susu

$$= \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$



Kualitas Data

- Kesalahan Pengukuran
- Kesalahan Pengumpulan

1. Input (Dataset)

- Jenis dataset ada dua: Private dan Public
- Private Dataset: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
  - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- Public Dataset: data set dapat diambil dari repositori pubik yang disepakati oleh para peneliti data mining
  - UCI Repository (*http://www.ics.uci.edu/~mlearn/MLRepository.html*)
  - ACM KDD Cup (*http://www.sigkdd.org/kddcup/*)
- Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: comparable, repeatable dan verifiable

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation