Data Mining

Pertemuan 4
Feature Extraction

Data Discretization

- Mengubah variabel kontinu pada data menjadi discrete (kata) atau index (angka berurutan), sehingga evaluasi dan pengelolaan data menjadi lebih mudah.

Rule

- Jika **suhu panas** dan **kelembaban udara tinggi** maka **potensi hujan tinggi**

Tabel

Variabel Kontinu	Discrete
X<= 25	Dingin
25 < x <=29	Sejuk
x>=30	Panas

Midian

- Gunakan submidian dan midian untuk menentukan batasan.
- Submidian umumnya 25% dan 75% tetapi bisa dibuat lebih rinci sesuai variasi data.

Keuntungan Discretization

- Mengurangi Row
- Mempermudah pembuatan pivot table

String Indexer

- Mengubah variabel discrete atau categori menjadi integer sehingga bisa di kalkulasi correlation, covariance dan sebagainya

Data Normalization

- Pada metode berbasis jarak (distance-based method), normalisasi mencegah atribut dengan rentang yang besar menyebabkan atribut dengan rentang kecil menjadi tidak “terlihat”.
- Menyamakan skala

Jenis Normalisasi

- Formula normalisasi untuk:
  - Min-max normalization:

$$v' = \frac{v - min_v}{max_v - min_v}(new\_max_v - new\_min_v) + new\_min_v$$

  - Z-score normalization:

$$v' = \frac{v - \bar{v}}{\sigma_v}$$

  - Normalisasi dengan *decimal scaling*:

$$v' = \frac{v}{10^j}$$

Di mana j adalah integer terkecil sehingga Max(|v'|) < 1