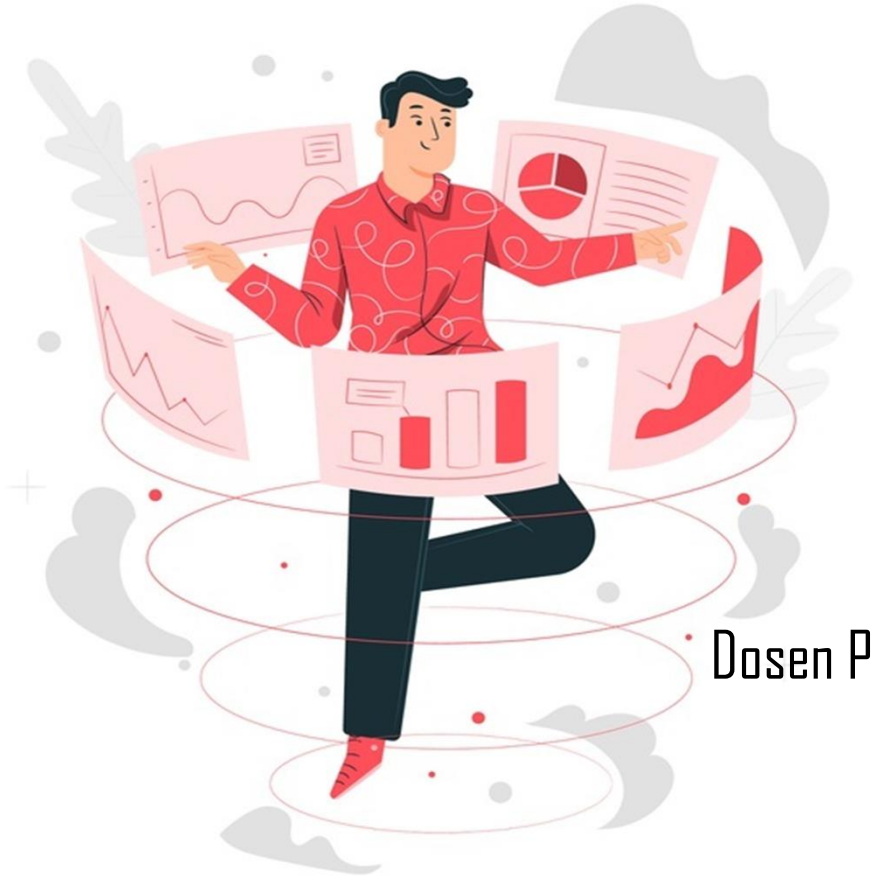
Penambangan Data
[Data Mining]



Kode : SIT5255

Bobot : 2 SKS

Dosen Pengampu : Dr. Heny Pratiwi, S.Kom., M.Pd., M.TI

Capaian Pembelajaran :

Dapat mengetahui dan memahami konsep data mining agar mampu melakukan penerapan atas teknik-teknik dan model sistem data mining dengan memanfaatkan aplikasi/tools yang digunakan dalam data mining.

DEFINISI

- Outlier/anomali ??

- Analisis outlier dikenal juga dengan analisis anomali atau deteksi anomali atau deteksi deviasi

Data Outlier disebut juga dengan data pencilan. Pengertian dari Outlier adalah data observasi yang muncul dengan nilai-nilai ekstrim, baik secara univariat ataupun multivariat.

MANFAAT MENGGUNAKAN ANALISIS OUTLIER

Studentized Residual

Outlier Univariat

Outlier Multivariat

PENYEBAB ADANYA OUTLIER

????

SKEMA ANALISIS OUTLIER

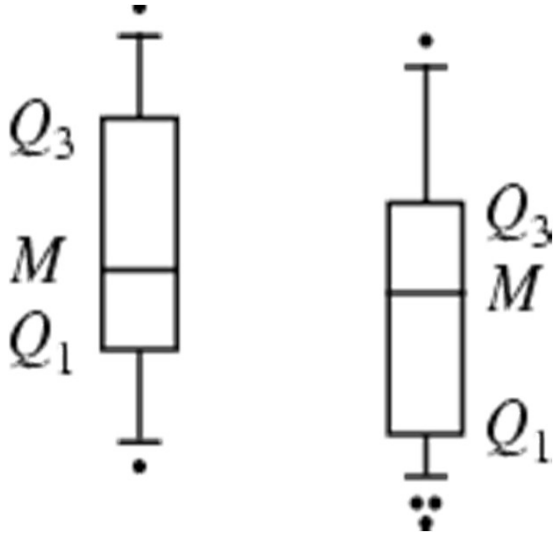
- Bangun profil dari data “normal”
- Gunakan profil tsb untuk mendeteksi anomali

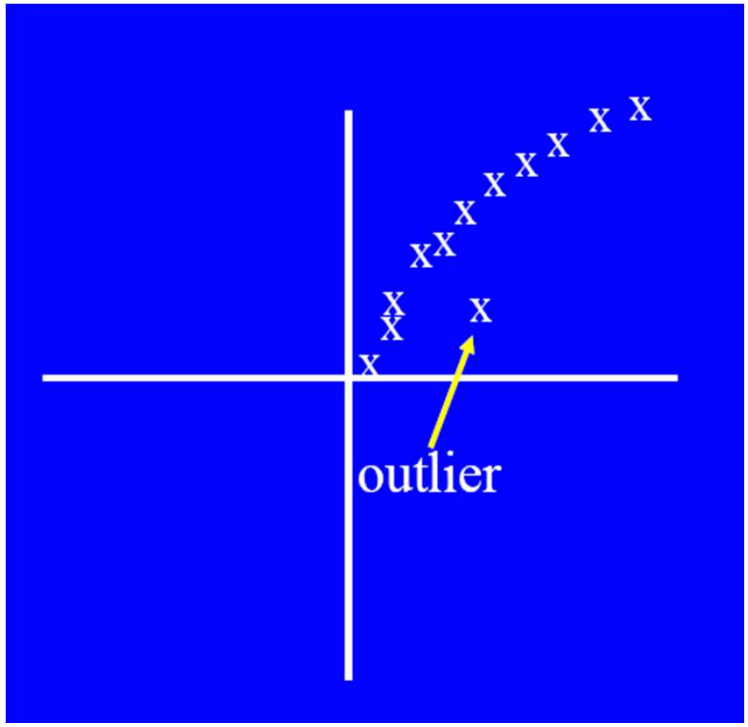
PENDEKATAN ANALISIS OUTLIER

a. Pendekatan Grafis
b. Model Based
c. Distance Based
d. Deviation Based

PENDEKATAN GRAFIS

- Misalkan dengan menggunakan Box Plot (1D), scatter plot (2 D) spin plot (3D)



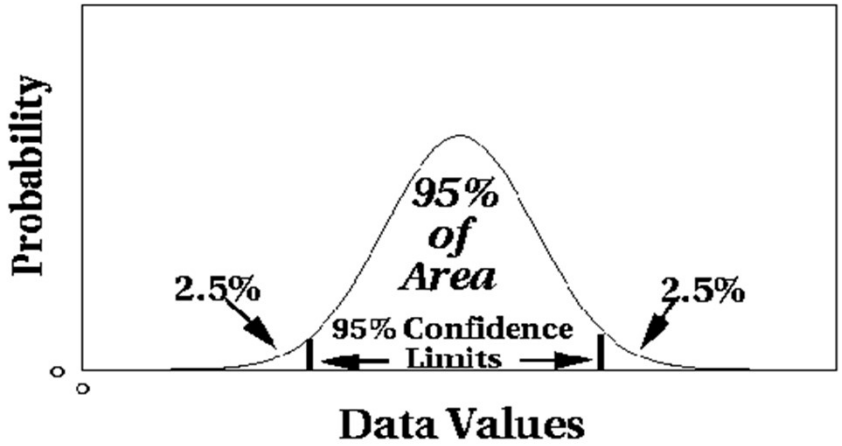


KEKURANGAN PENDEKATAN GRAFIS

????

PENDEKATAN STATISTIK

- Asumsikan fungsi distribusi data yang dimiliki (mis Distribusi Normal, distribusi Poison, distribusi Gamma,dsb)
- Gunakan Uji Statistik



KELEBIHAN & KEKURANGAN PENDEKATAN STATISTIK

- Jenis distribusi data dan jenis uji yang diperlukan.
- Fungsi distribusi dan jenis uji yang tepat
- Single attribut
- Data berdimensi tinggi

NEAREST-NEIGHBOR BASED

- Tentukan jarak dari tiap pasang titik (data)
- Sebuah titik dikatakan outlier jika (pilih salah satu):

  - Banyaknya titik tetangga di sekitarnya lebih sedikit dari ***p*** dalam jarak ***D***

  - Titik tsb merupakan top n titik yang jaraknya paling jauh dari k tetangga terdekatnya

  - Titik tsb merupakan top n titik rata-rata jaraknya paling besar dari k tetangga terdekatnya

KELEBIHAN & KEKURANGAN NEAREST_NEIGHBOR APPROACH

- Pendekatannya
- Basis data
- Nilai parameter
- Kasus himpunan data yang memiliki kepadatan

OUTLIERS PADA PROYEKSI DIMENSI RENDAH

- Ruang berdimensi tinggi, data menjadi jarang dan jarak menjadi tidak memberi “arti“.
- Solusi : Metoda outlier pada Proyeksi pada Ruang yang berdimensi lebih rendah.

DENSITY BASED

- Berdasarkan pendekatan density-based, outlier adalah titik yang berada pada daerah dengan kepadatan rendah (tidak padat)

$$density(x,k) = \left(\frac{\sum_{y \in N(x,k)} dist(x,y)}{|N(x,k)|} \right)^{-1}$$

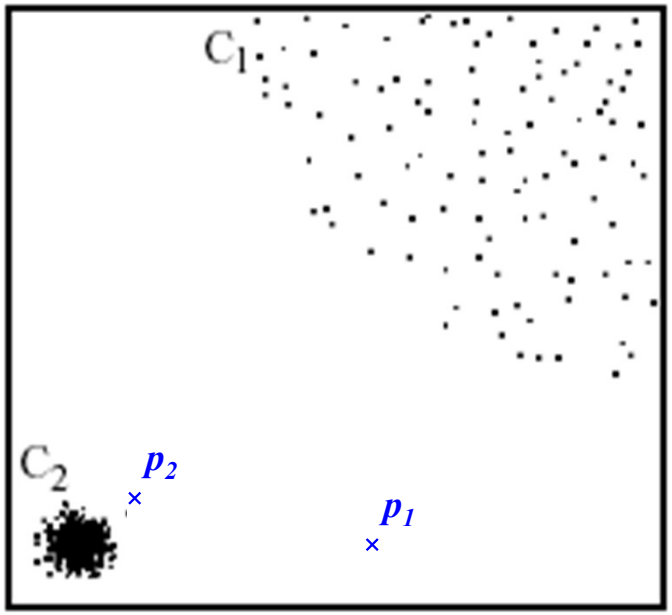
N(x,k) adalah himpunan yang berisi k tetangga terdekat x, y adalah tetangga terdekat dari x dan |N(x,k)| adalah banyaknya anggota himpunan N(x,k)

DENSITY-BASED: LOF APPROACH

- Untuk setiap titik, hitunglah kepadatan lokal dengan average relative density

$$average\_relative\_density(x,k) = \frac{density(x,k)}{\sum_{y \in N(x,k)} density(y,k) / |N(x,k)|}$$

- Outlier adalah titik dengan nilai LOF (ard) terbesar



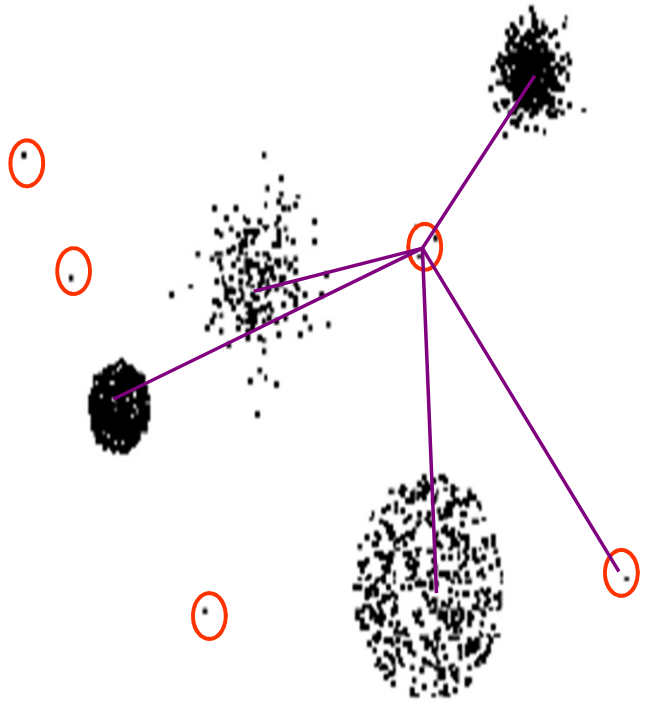
Dengan pendekatan NN, p_2 tidak akan dianggap sbg outlier, sedangkan dengan pendekatan LOF, p_1 dan p_2 akan dianggap sebagai outlier

KELEBIHAN & KEKURANGAN DENSITY BASED

- Kepadatan Data
- Pemilihan Parameter

CLUSTERING-BASED

- Ide dasar:
  - Kelompok-kelompok yang kepadatannya berbeda-beda
  - Pilih titik-titik yang berada pada klaster yang kecil sebagai kandidat outlier
  - Hitung jarak antara titik-titik kandidat outlier dengan titik-titik yg berada pada klaster non-kandidat.
    - Jika titik-titik kandidat terletak jauh dari semua titik-titik non kandidat, maka titik kandidat tsb adalah outlier



KELEBIHAN & KEKURANGAN CLUSTERING BASED

- Dapat menggunakan berbagai teknik clustering.
- Pemilihan nilai parameter.
- Hanya sesuai dengan tipe data tertentu

DEVIATION-BASED APPROACH

1. Mengidentifikasi outliers dengan menentukan karakteristik utama dari objek-objek dalam sebuah grup
2. Objek yang memiliki “deviasi” dari deskripsi ini, akan dianggap sebagai outlier
3. Teknik sequential exception
4. Teknik OLAP data cube

Sekian &
Terima Kasih