



# KMMI 2021

## Eksplorasi dan Visualisasi Data

Pertemuan 7:  
Analisis Hubungan Antar Variabel

## Sub CPMK

- Mahasiswa menjelajahi hubungan antara variabel yang berbeda.

## Pokok Bahasan

- Mencari nilai korelasi
- Menginterpretasikan hasil korelasi
- Memilih variabel signifikan
- Reduksi dimensi: CFA dan EFA

# Korelasi

- Analisis korelasi adalah alat yang berguna untuk mengukur hubungan antara variabel yang berbeda dari dataset.
- Metrik korelasi antara dua variabel numerik mencerminkan jika peningkatan nilai suatu variabel terjadi bersamaan dengan peningkatan nilai variabel lain.
- Misalnya, peningkatan konsumsi minuman manis dan peningkatan prevalensi diabetes.
- Untuk saat ini kita bahas koefisien korelasi Pearson.
- Metrik korelasi biasanya mengambil nilai antara -1 dan 1.
- Nilai nol menunjukkan tidak ada korelasi antara kedua variabel, sedangkan nilai positif atau negatif yang lebih besar menunjukkan bahwa variabel masing-masing berkorelasi positif atau negatif.

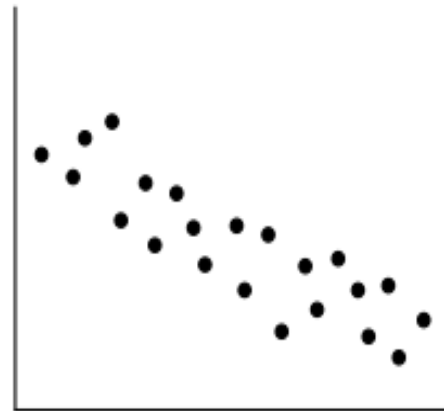
# Korelasi data & Scatter plot

- Scatter plot adalah metode yang berguna untuk memberikan pandangan pertama pada data bivariat untuk melihat kelompok dari titik-titik observasi dan outlier, atau untuk mengeksplorasi kemungkinan hubungan korelasi.
- Dua atribut,  $X$  dan  $Y$ , berkorelasi jika satu atribut berimplikasi pada atribut yang lain.

# Korelasi data & Scatter plot



(a)



(b)

- (a) korelasi positif
- (b) korelasi negatif antar atribut



Tiga kasus di mana tidak ada korelasi yang diamati antara dua atribut yang diplot di masing-masing set data

# Korelasi data numerik

- Identifikasi korelasi antar variabel dapat dilihat dari scatter plot dan koefisien korelasi ( $r$ ).
- Scatter plot dapat digunakan untuk mengetahui korelasi antar variabel secara visual dengan melihat persebaran data-points.
- Nilai koefisien korelasi ( $r$ ) mengindikasikan kekuatan dan arah dari hubungan yang linear antar dua variabel.

# Korelasi data numerik

- Koefisien korelasi ( $r$ ) bernilai antara -1 dan 1. Semakin mendekati -1 atau 1 maka hubungan semakin kuat.
- Nilai positif menunjukkan hubungan kedua variabel searah, sebaliknya, apabila nilainya negatif menunjukkan hubungan yang berlawanan.
- Jika, Koefisien korelasi ( $r$ ) = 0 maka dapat diartikan bahwa tidak ada hubungan antara dua variabel atau keduanya mempunyai hubungan yang lemah.

# Korelasi data numerik

Keterkaitan antara pola data-points pada scatter-plot dengan koefisien korelasi ( $r$ ) yaitu

- ( $r > 0$ ), jika  $x$  dan  $y$  membentuk pola rendah di sebelah kiri kemudian meningkat ke arah kanan
- ( $r < 0$ ), jika  $x$  dan  $y$  membentuk pola tinggi di sebelah kiri kemudian menurun ke arah kanan
- ( $r = +1$ ), jika  $x$  dan  $y$  membentuk pola garis lurus dengan arah yang positif
- ( $r = -1$ ), jika  $x$  dan  $y$  membentuk pola garis lurus dengan arah yang negatif
- ( $r \approx 0$ ) jika pola random/ acak



# Korelasi data numerik

Untuk data numerik, kita dapat mengevaluasi korelasi antara dua atribut, A dan B, dengan menghitung koefisien korelasi (juga dikenal sebagai koefisien momen produk Pearson, dinamai menurut penemunya, Karl Pearson).

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

$$-1 \leq r_{A,B} \leq +1$$

$n$  : jumlah tupel,

$a_i$  dan  $b_i$  : nilai dari A dan B dalam tupel  $i$

$\bar{A}$  dan  $\bar{B}$  : nilai rata-rata dari A dan B

$\sigma_A$  dan  $\sigma_B$  : standar deviasi dari A dan B

$\sum_{i=1}^n (a_i b_i)$  : jumlah dari produk silang AB (yaitu, untuk setiap tupel, nilai A dikalikan dengan nilai B dalam tupel itu).

# Korelasi data numerik

- Jika  $r_{A,B}$  lebih besar dari 0, maka A dan B berkorelasi positif, artinya nilai A meningkat seiring dengan peningkatan nilai B. Semakin tinggi nilainya, semakin kuat korelasinya.
- Jika nilai yang dihasilkan sama dengan 0, maka A dan B saling bebas dan tidak ada korelasi antara keduanya.
- Jika nilai yang dihasilkan kurang dari 0, maka A dan B berkorelasi negatif, di mana nilai satu atribut meningkat seiring dengan penurunan nilai atribut lainnya.
- Perhatikan bahwa korelasi tidak menyiratkan kausalitas. Artinya, jika A dan B berkorelasi, ini tidak berarti bahwa A menyebabkan B atau B menyebabkan A.

# Korelasi data kategorik

## Data Numerik

- Data yang berbentuk angka dengan skala data interval atau rasio
- Hubungan antar variabel:
  1. Scatter plot
  2. Koefisien korelasi ( $r$ )

## Data Kategorik

- Observasi yang diklasifikasikan menjadi kategori sehingga data berisi banyak frekuensi per kategori
- Hubungan antar variabel: Chi-Square  $\chi^2$  test

# Korelasi data kategorik

- Metode yang dapat digunakan yaitu Uji Chi-square ( $\chi^2$ ).
- Uji ini dapat dihitung menggunakan tabel kontingensi, yaitu tabel data jumlah frekuensi yang muncul dari klasifikasi pengamatan sampel menurut dua atau lebih karakteristik.
- Digunakan untuk mengetahui apakah dua variabel tersebut independent atau cenderung berasosiasi.
- Untuk data nominal, hubungan korelasi antara dua atribut, A dan B, dapat ditemukan dengan uji  $\chi^2$  (chi-square).

# Korelasi data kategorik

- Misalkan A memiliki  $c$  nilai yang berbeda, yaitu  $a_1, a_2, \dots, a_c$ . B memiliki  $r$  nilai yang berbeda, yaitu  $b_1, b_2, \dots, b_r$ .
- Tupel data yang dijelaskan oleh A dan B dapat ditampilkan sebagai tabel kontingensi, dengan nilai  $c$  dari A membentuk kolom dan nilai  $r$  dari B membentuk baris.
- Misalkan  $(A_i, B_j)$  menyatakan kejadian bersama bahwa atribut A mengambil nilai  $a_i$  dan atribut B mengambil nilai  $b_j$ , yaitu, di mana  $(A = a_i, B = b_j)$ . Setiap kejadian  $(A_i, B_j)$  bersama yang mungkin memiliki sel (atau slot) sendiri dalam tabel.

# Korelasi data kategorik

- Nilai  $\chi^2$  (juga dikenal sebagai statistik Pearson  $\chi^2$ ) dihitung sebagai

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

di mana  $o_{ij}$  adalah frekuensi yang diamati (yaitu, hitungan aktual) dari kejadian bersama  $(A_i, B_j)$  dan  $e_{ij}$  adalah frekuensi yang diharapkan dari  $(A_i, B_j)$ , yang dapat dihitung sebagai

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n}$$

$n$  adalah jumlah tupel data,  $\text{count}(A = a_i)$  adalah jumlah tupel yang memiliki nilai  $a_i$  untuk  $A$ , dan  $\text{count}(B = b_j)$  adalah jumlah tupel yang memiliki nilai  $b_j$  untuk  $B$

- Jumlah dalam persamaan  $\chi^2$  di atas, dihitung atas semua sel  $r \times c$ .
- Statistik  $\chi^2$  menguji hipotesis bahwa  $A$  dan  $B$  independen, yaitu tidak ada korelasi di antara keduanya.
- Pengujian didasarkan pada tingkat signifikansi, dengan  $(r - 1) \times (c - 1)$  derajat kebebasan.

# Korelasi data kategorik: Contoh

Misalkan sekelompok 1500 orang disurvei. Jenis kelamin setiap orang dicatat. Setiap orang disurvei apakah jenis bahan bacaan yang disukainya adalah fiksi atau nonfiksi. Jadi, kita memiliki dua atribut, jenis kelamin dan bacaan yang disukai.

Frekuensi yang diamati (atau jumlah) dari setiap kemungkinan kejadian bersama diringkas dalam tabel kontingensi yang ditunjukkan pada Tabel, di mana angka dalam tanda kurung adalah frekuensi yang diharapkan.

	Male	Female	Total
Fiction	250 (90)	200 (360)	450
Non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

# Korelasi data kategorik: Contoh

Frekuensi yang diharapkan dihitung berdasarkan distribusi data untuk kedua atribut menggunakan persamaan  $e_{ij}$ .

Menggunakan persamaan tersebut, kita dapat memverifikasi frekuensi yang diharapkan untuk setiap sel. Misalnya, frekuensi yang diharapkan untuk sel (laki-laki, fiksi) adalah

$$e_{11} = \frac{\text{count}(\text{male}) \times \text{count}(\text{fiction})}{n} = \frac{300 \times 450}{1500} = 90$$

Menggunakan tabel tersebut, apakah jenis kelamin dan preferensi bacaan berkorelasi?

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507,93$$

Untuk tabel  $2 \times 2$  ini, derajat kebebasannya adalah  $(2 - 1)(2 - 1) = 1$ . Untuk 1 derajat kebebasan, nilai  $\chi^2$  yang diperlukan untuk menolak hipotesis pada tingkat signifikansi 0,001 adalah 10,828 (diambil dari tabel distribusi  $\chi^2$ ).

Karena nilai yang dihitung di atas ini, kita dapat menolak hipotesis bahwa jenis kelamin dan bacaan yang disukai adalah independen dan menyimpulkan bahwa kedua atribut tersebut (sangat) berkorelasi untuk kelompok orang tertentu



# Pemilihan Variabel

- Tujuan dari eksplorasi data adalah untuk mengetahui karakteristik dari data yang ada
- Oleh karena itu, tahap ini lekat kaitannya dengan tahap *data preparation*, dimana hal ini berfokus pada kebersihan data dan penilaian kualitas data itu sendiri.
- Tujuan dari eksplorasi data adalah untuk memahami hubungan antar variabel untuk menginformasikan pemilihan variabel dan metode untuk memahami domain masalah
- Jumlah variabel yang banyak akan sangat merepotkan, oleh karena itu kita perlu memilih variabel terbaik yang benar-benar diperlukan.

# Reduksi dimensi

- Bayangkan Anda telah memilih data yang memiliki dimensi ratusan ribu pada kolom dan baris untuk dianalisis.
- Data ini sangatlah besar dan yang perlu diingat bahwa analisis data yang kompleks dan penambahan data dalam jumlah besar dapat memakan waktu lama, membuat analisis tersebut tidak praktis atau tidak layak.
- Strategi reduksi data (data reduction) termasuk pengurangan dimensi (dimensionality reduction), pengurangan jumlah (numerosity reduction), dan kompresi data (data compression).
- Pengurangan dimensi adalah proses mengurangi jumlah variabel acak atau atribut yang dipertimbangkan.
- Metode pengurangan dimensi termasuk transformasi wavelet dan analisis komponen utama, yang mengubah atau memproyeksikan data asli ke ruang yang lebih kecil.
- Pemilihan subset atribut adalah metode pengurangan dimensi di mana atribut atau dimensi yang tidak relevan, relevan lemah, atau redundan dideteksi dan dihilangkan.

# Reduksi dimensi: wavelet transforms

- Transformasi wavelet diskrit (*discrete wavelet transform*) adalah teknik pemrosesan sinyal linier yang ketika diterapkan pada vektor data  $X$ , mengubahnya menjadi vektor yang berbeda secara numerik,  $X'$ , dari koefisien wavelet.
- Kedua vektor tersebut memiliki panjang yang sama.
- Saat menerapkan teknik ini pada reduksi data, kita menganggap setiap tupel sebagai vektor data  $n$ -dimensi, yaitu,  $X = (x_1, x_2, \dots, x_n)$ , yang menggambarkan  $n$  pengukuran yang dilakukan pada tupel dari  $n$  atribut database.

# Reduksi dimensi: Principal Component Analysis (PCA)

- Analisis komponen utama (PCA; juga disebut metode Karhunen-Loeve, atau K-L) mencari vektor ortogonal  $k$   $n$ -dimensi yang paling baik digunakan untuk mewakili data, di mana  $k \leq n$ .
- Data asli dengan demikian diproyeksikan ke ruang yang jauh lebih kecil, menghasilkan pengurangan dimensi.
- Tidak seperti pemilihan subset atribut (*attribute subset selection*), yang mengurangi ukuran set atribut dengan mempertahankan subset dari set atribut awal, PCA “menggabungkan” esensi atribut dengan membuat alternatif, set variabel yang lebih kecil.

# Reduksi dimensi: Principal Component Analysis (PCA)

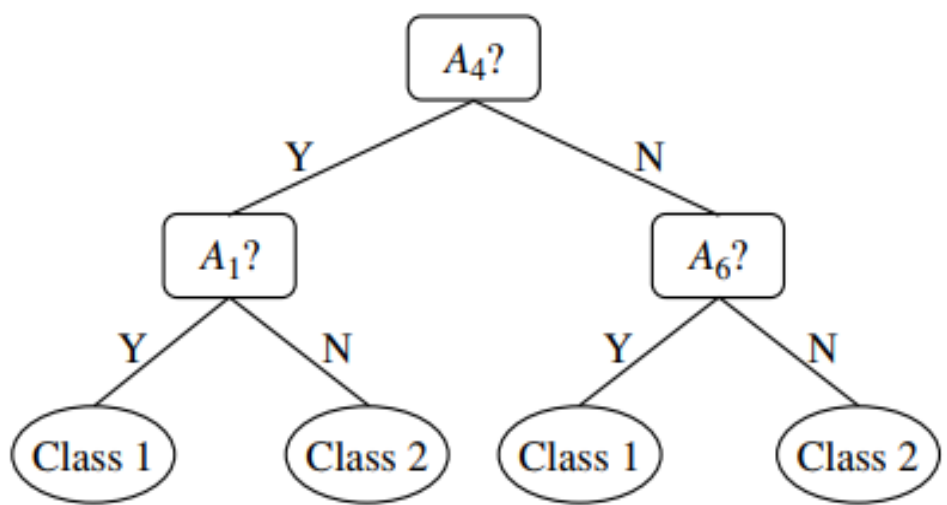
Prosedur dasar untuk PCA adalah sebagai berikut:

1. Data input dinormalisasi, sehingga setiap atribut berada dalam rentang yang sama..
2. PCA menghitung  $k$  vektor ortonormal yang menyediakan basis untuk data masukan yang dinormalisasi. Vektor ini adalah vektor satuan yang setiap titik dalam arah tegak lurus terhadap yang lain. Vektor-vektor ini disebut sebagai komponen utama. Data masukan merupakan kombinasi linier dari komponen utama.
3. Komponen utama diurutkan berdasarkan penurunan "signifikansi" atau kekuatannya.
4. Karena komponen diurutkan dalam urutan "signifikansi" yang menurun, ukuran data dapat dikurangi dengan menghilangkan komponen yang lebih lemah, yaitu komponen dengan varians rendah. Dengan menggunakan komponen utama yang paling kuat, seharusnya dimungkinkan untuk merekonstruksi perkiraan yang baik dari data asli.

# Reduksi dimensi: Attribute Subset Selection

- Pemilihan subset atribut mengurangi ukuran kumpulan data dengan menghapus atribut (atau dimensi) yang tidak relevan atau berlebihan.
- Tujuan pemilihan subset atribut adalah untuk menemukan himpunan atribut minimum sedemikian rupa sehingga distribusi probabilitas yang dihasilkan dari kelas data sedekat mungkin dengan distribusi asli yang diperoleh dengan menggunakan semua atribut

# Reduksi dimensi: Attribute Subset Selection

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p>Initial reduced set: <math>\{\}</math>  <math>\Rightarrow \{A_1\}</math>  <math>\Rightarrow \{A_1, A_4\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set: <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p><math>\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}</math>  <math>\Rightarrow \{A_1, A_4, A_5, A_6\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set: <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p>  <pre> graph TD     A4["A4?"] -- Y --&gt; A1["A1?"]     A4 -- N --&gt; A6["A6?"]     A1 -- Y --&gt; C1_1((Class 1))     A1 -- N --&gt; C2_1((Class 2))     A6 -- Y --&gt; C1_2((Class 1))     A6 -- N --&gt; C2_2((Class 2))     </pre> <p><math>\Rightarrow</math> Reduced attribute set: <math>\{A_1, A_4, A_6\}</math></p>

# Reduksi dimensi: Attribute Subset Selection

- **Stepwise forward selection:** Prosedur dimulai dengan set atribut kosong sebagai set yang direduksi. Atribut asli terbaik ditentukan dan ditambahkan ke himpunan tereduksi. Pada setiap iterasi atau langkah berikutnya, yang terbaik dari atribut asli tersisa ditambahkan ke set.
- **Stepwise backward elimination:** Prosedur dimulai dengan set atribut lengkap. Pada setiap langkah, atribut terburuk dihilangkan dari set.
- **Combination of forward selection and backward elimination:** Metode seleksi maju bertahap dan eliminasi mundur dapat digabungkan sehingga, pada setiap langkah, prosedur memilih atribut terbaik dan menghilangkan atribut terburuk di antara atribut yang tersisa.
- **Decision tree induction:** . Induksi pohon keputusan membangun struktur seperti diagram alur di mana setiap simpul internal / internal node (nonleaf) menunjukkan pengujian pada atribut, setiap cabang sesuai dengan hasil pengujian, dan setiap simpul eksternal (leaf) menunjukkan prediksi kelas. Semua atribut yang tidak muncul di pohon dianggap tidak relevan. Himpunan atribut yang muncul di pohon membentuk subset atribut yang direduksi.



# Reduksi dimensi: Regression & Loglinear Models: Parametric Data Reduction

- Model regresi dan log-linier dapat digunakan untuk meng-*approximate* data yang diberikan.
- Dalam regresi linier (sederhana), data dimodelkan agar sesuai dengan garis lurus.
- Misalnya, variabel acak,  $y$  (disebut variabel respons), dapat dimodelkan sebagai fungsi linier dari variabel acak lain,  $x$  (disebut variabel prediktor), dengan persamaan:

$$y = \beta_0 + \beta_1 x$$

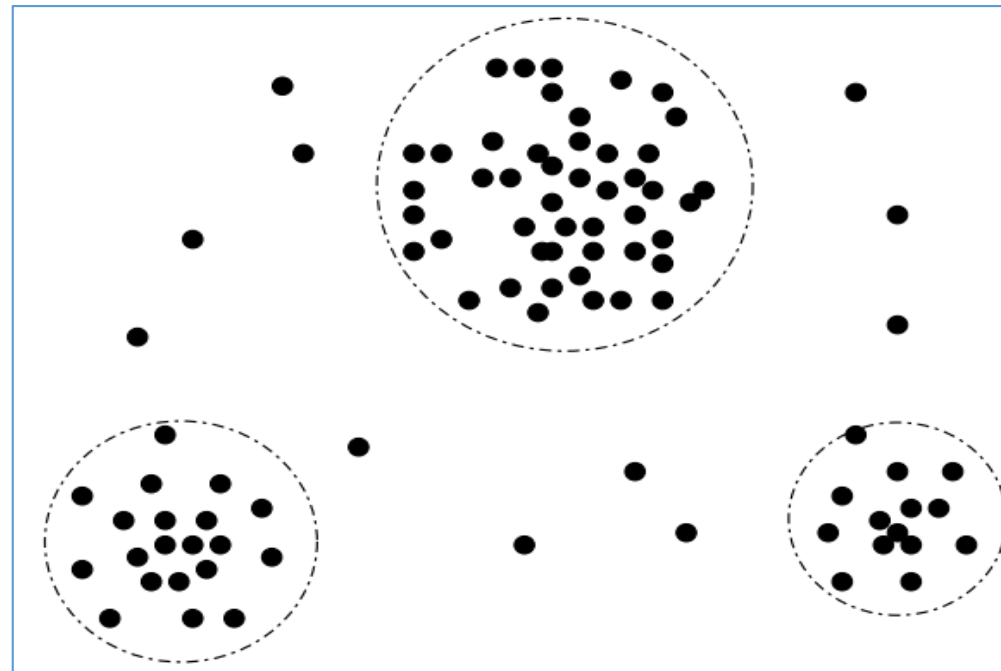
- Variabel yang signifikan adalah variabel terbaik yang dapat dimasukkan kedalam model regresi ini.

# Reduksi dimensi: Regression & Loglinear Models: Parametric Data Reduction

- Model regresi dan log-linear keduanya dapat digunakan pada data yang jarang (*sparse data*), meskipun aplikasinya mungkin terbatas.
- Kedua metode dapat menangani data miring (*skewed data*), dimana regresi bekerja dengan sangat baik.
- Regresi dapat menjadi komputasi intensif bila diterapkan pada data dimensi tinggi (*high dimensionality data*), sedangkan model log-linear menunjukkan skalabilitas yang baik hingga 10 atau lebih dimensi.

# Reduksi dimensi: Clustering

- Dalam reduksi data, representasi *cluster* dari data digunakan untuk menggantikan data yang sebenarnya. Efektivitas teknik ini tergantung pada sifat data.
- Teknik ini jauh lebih efektif untuk data yang dapat diatur ke dalam kelompok yang berbeda (*distinct clusters*) daripada untuk *smearred data*.



Plot data pelanggan 2-D sehubungan dengan lokasi pelanggan di sebuah kota, menunjukkan tiga klaster data

# Tugas

1. Mencari hubungan antar variabel/atribut pada dataset yang telah diunduh dari Kaggle atau UCI Machine Learning Repository. (Jumlah atribut pada data minimal 10 atribut)
2. Lakukan reduksi dimensi menggunakan salah satu teknik yang sudah dibahas dalam materi.
  - Tugas dikerjakan berkelompok
  - Tugas dikumpulkan paling lambat pukul 23.59 WIB di LMS.
  - Beri nama file tugas: Tugas 07\_Kelompok XX (Contoh: Tugas 07\_Kelompok 01)



# Terima Kasih