

Unsupervised Learning

Introduction +
Dimensionality Reduction



Hafizh Adi Prasetya

Education Background



2011-2015
Bachelor Degree
Informatics



2017-2019
Master Degree
Artificial Intelligence



Hafizh Adi Prasetya

<https://id.linkedin.com/in/hafizhadi>

Hafizh Adi Prasetya
Data Scientist

Unsupervised Learning

- **Sesi I: Intro + Dimensionality Reduction**
- Sesi II : Clustering

Objektif: Introduction + Dimensionality Reduction



Mendapatkan **pemahaman dasar mengenai unsupervised learning, jenis, posisinya** dalam machine learning, serta **kegunaannya** dalam kasus-kasus di dunia nyata.

Mendapatkan **pemahaman mengenai Dimensionality Reduction** dan kemampuan untuk mengimplementasikannya menggunakan Python.

Expected Output



1. Memahami konsep unsupervised learning dan perbedaannya dengan supervised learning
2. Memahami dua tipe unsupervised learning dan kegunaannya di dunia nyata
3. Memahami konsep dan alasan diperlukannya dimensionality reduction untuk data skala besar
4. Memahami Principal Component Analysis (PCA) secara umum dan contoh-contoh penggunaannya
5. Memahami cara menggunakan Python untuk melakukan PCA pada data
6. Memiliki pemahaman intuitif mengenai step-by-step algoritma PCA dan hubungannya dengan dimensionality reduction

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Apa itu
Unsupervised Learning?

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction



Definisi dan Jenis-jenis Unsupervised Learning

Contoh Kasus Unsupervised Learning

Dimensionality Reduction dan Penggunaannya

Intuisi dan Motivasi Principal Component Analysis (PCA)

PCA (Praktik)

Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

Intuisi dan Motivasi Clustering

Clustering dan Segmentasi dalam Bisnis

Intermezzo: Pengukuran Jarak

Algoritma Agglomerative Clustering dan Praktik

Algoritma K-means Clustering dan Praktik

Evaluasi Clustering

Berbagai Tipe Machine Learning

Supervised Learning

- ??

Unsupervised Learning

- ??

Lainnya

- Semi-supervised Learning
- Reinforcement Learning



Kunci membedakan berbagai jenis Machine Learning:

Jenis dan kuantitas data yang tersedia untuk dipelajari

Berbagai Tipe Machine Learning

Supervised Learning

- Pembelajaran dari data DENGAN LABEL
- Klasifikasi & regresi

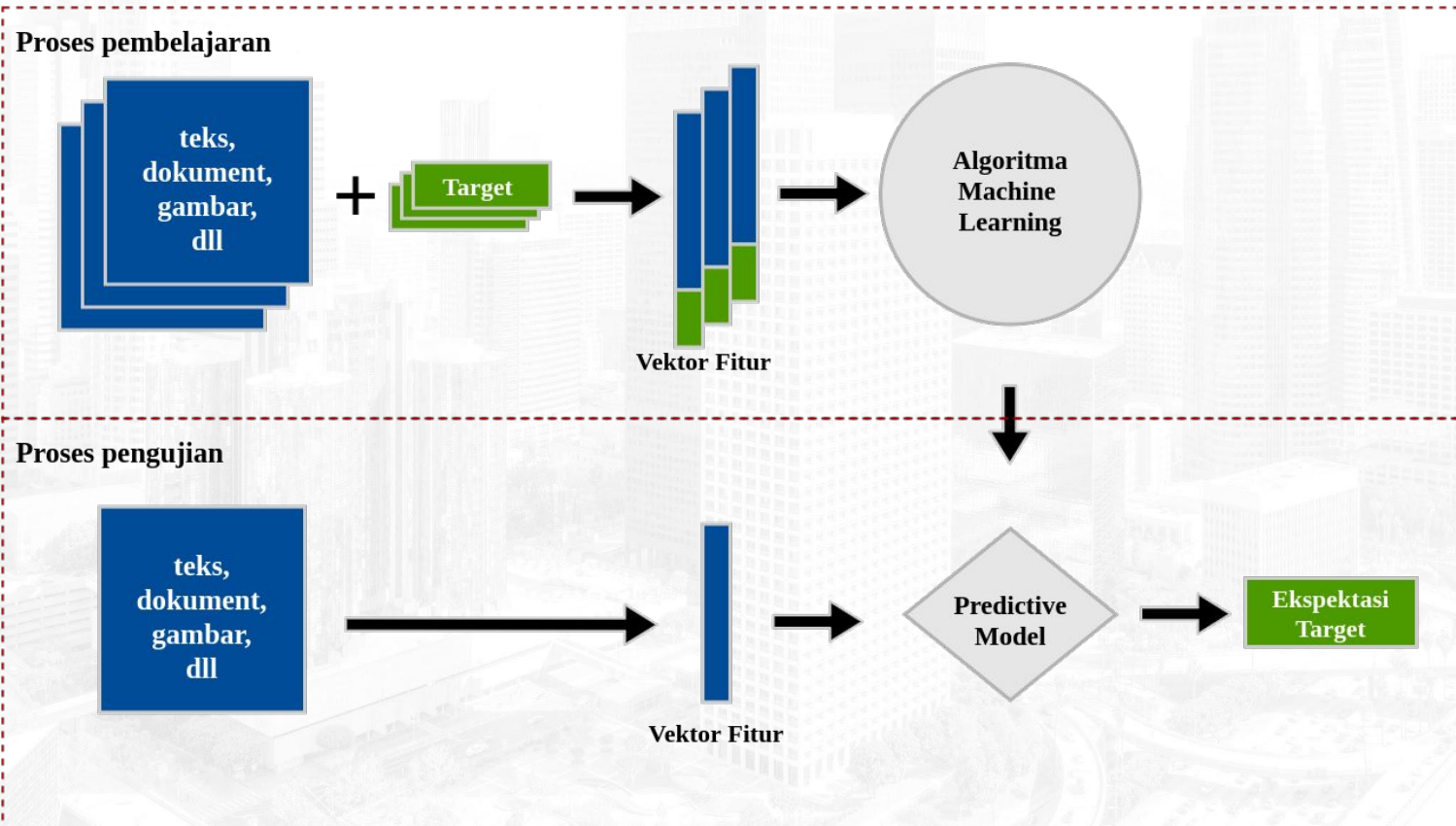
Unsupervised Learning

- Pembelajaran dari data TANPA LABEL
- Dimensionality reduction, clustering & representation learning

Lainnya

- Semi-supervised Learning -> BERLABEL SEBAGIAN
- Reinforcement Learning -> DATA DIHASILKAN DARI SIMULASI

Alur pemodelan Supervised Learning



SUPERVISED

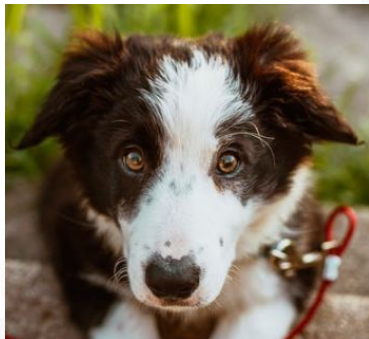
Feature + Label

is_diabetes	num_pregnant	glucose_concentration	blood_pressure	triceps_thickness	two_hour_insulin	bmi	pedigree_function	age
1	6	148	72	35	0	33.6	0.627	50
0	1	85	66	29	0	26.6	0.351	31
1	8	183	64	0	0	23.3	0.672	32
0	1	89	66	23	94	28.1	0.167	21
1	0	137	40	35	168	43.1	2.288	33
0	5	116	74	0	0	25.6	0.201	30
1	3	78	50	32	88	31	0.248	26
0	10	115	0	0	0	35.3	0.134	29
1	2	197	70	45	543	30.5	0.158	53
1	8	125	96	0	0	0	0.232	54
0	4	110	92	0	0	37.6	0.191	30
1	10	168	74	0	0	38	0.537	34
0	10	139	80	0	0	27.1	1.441	57
1	1	189	60	23	846	30.1	0.398	59

SUPERVISED



CAT



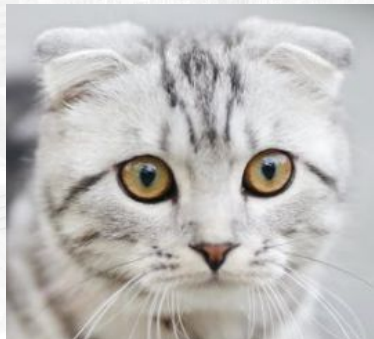
DOG



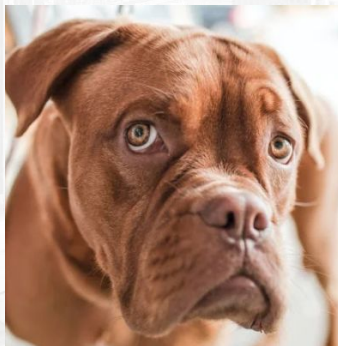
CAT



CAT



CAT



DOG



CAT



?

SUPERVISED



LABEL



LABEL



LABEL



LABEL



LABEL



LABEL



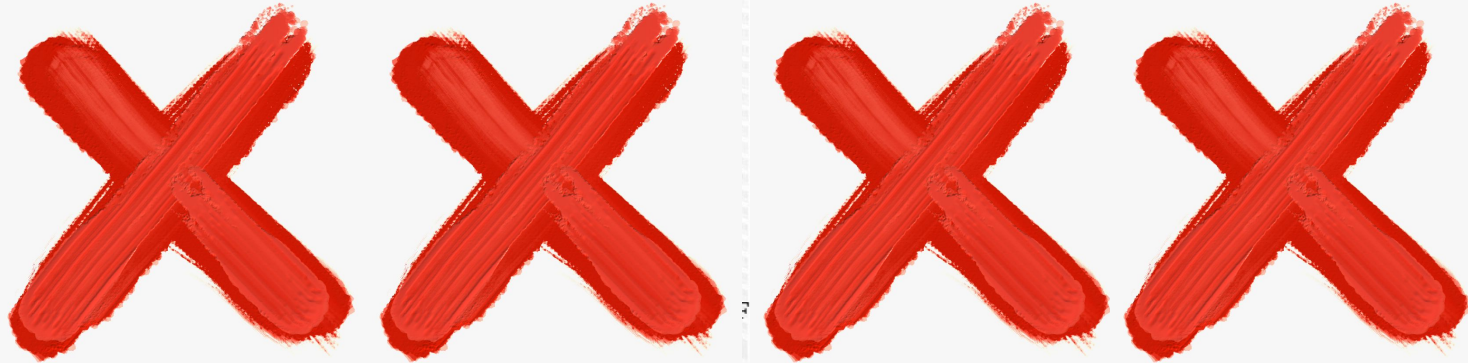
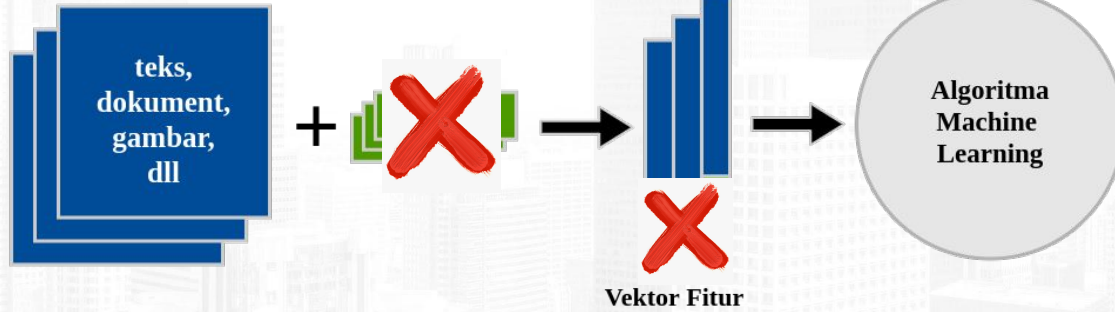
LABEL



**PREDIKSI
LABEL**

Unsupervised Learning?

Proses pembelajaran

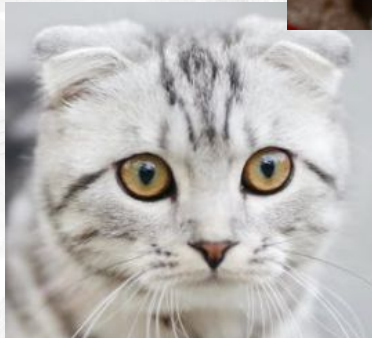


UNSUPERVISED

Hanya feature

is_diabetes	num_pregnant	glucose_concentration	blood_pressure	triceps_thickness	two_hour_insulin	bmi	pedigree_function	age
	6	148	72	35	0	33.6	0.627	50
	1	85	66	29	0	26.6	0.351	31
	8	183	64	0	0	23.3	0.672	32
	1	89	66	23	94	28.1	0.167	21
	0	137	40	35	168	43.1	2.288	33
	5	116	74	0	0	25.6	0.201	30
	3	78	50	32	88	31	0.248	26
	10	115	0	0	0	35.3	0.134	29
	2	197	70	45	543	30.5	0.158	53
	8	125	96	0	0	0	0.232	54
	4	110	92	0	0	37.6	0.191	30
	10	168	74	0	0	38	0.537	34
	10	139	80	0	0	27.1	1.441	57
	1	189	60	23	846	30.1	0.398	59

UNSUPERVISED



UNSUPERVISED

**Data
tanpa
label**



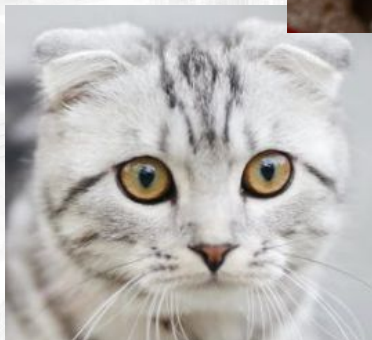
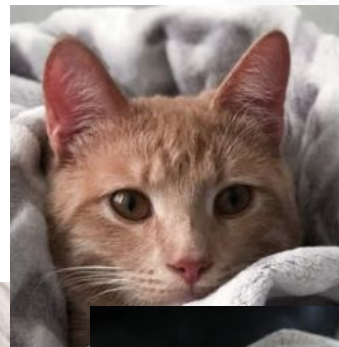


**Tanpa label, kegunaan apa yang
bisa kita dapatkan dari data?**

Unsupervised Learning #1: Mengambil 'intisari' dari dataset (atau subset dari dataset) yang kita miliki

Dari sebuah dataset yang memiliki N feature, kita ingin mendapatkan dataset baru dengan jumlah feature $< N$ namun dengan menghilangkan sesedikit mungkin 'informasi'

UNSUPERVISED



UNSUPERVISED



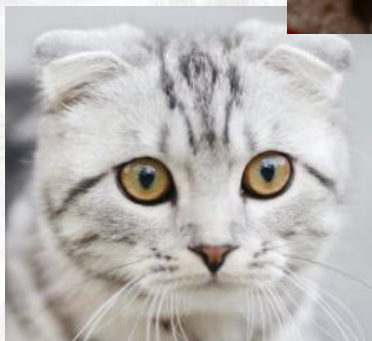
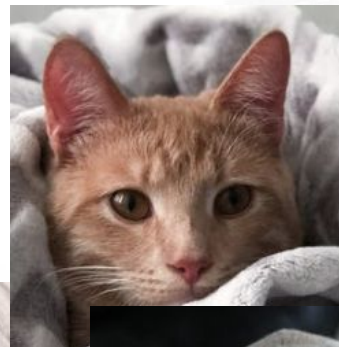
$\frac{1}{3}$ ukuran data tanpa menghilangkan makna

Unsupervised Learning #2: Kita tidak tahu apa-apa tentang 'kelas' dari setiap baris di data kita tapi kita ingin tahu apakah mereka dapat dipisahkan menjadi beberapa 'kelas'

Sebaik-baiknya hal yang dapat kita lakukan adalah:

- Kelompokkan data-data yang 'mirip'
- Kita lihat setiap kelompok yang dihasilkan secara kontekstual

UNSUPERVISED



UNSUPERVISED


KELOMPOK 1

KELOMPOK 2

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik



 Evaluasi Clustering

Contoh Kasus **Unsupervised Learning**

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis
Unsupervised Learning

  Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan
Penggunaannya

 Intuisi dan Motivasi Principal
Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi
dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative
Clustering dan Praktik

 Algoritma K-means Clustering dan
Praktik

 Evaluasi Clustering

Contoh Kasus #1

$N^{****}x$

Perusahaan streaming video N^{*tf*i*} memiliki data rating video dari setiap pengguna aplikasinya. Data ini memiliki dimensi [jumlah pengguna x jumlah film].

Bagaimana caranya kita bisa mendapatkan dataset yang lebih ringkas tapi masih berisi informasi tentang selera pengguna?

Meringkas Data dengan Dimensionality Reduction

User	KKN di Desa P	PLN di Desa P	Dilan 1990	...	Dilan 2077
A	5	5	NULL		1
B	NULL	NULL	4		2
C	NULL	NULL	NULL		6
D	3	NULL	NULL		2
...					
ZZZ	0	1	NULL		NULL

2 masalah utama data review video

- Ukuran data terlalu besar, mahal disimpan dan diproses
- Dataset bersifat sparse -> banyak nilai NULL pada dataset sehingga penyimpanan tidak efisien

Meringkas Data dengan Dimensionality Reduction

User	KKN di Desa P	PLN di Desa P	Dilan 1990	...	Dilan 2077
A	5	5	NULL		1
B	NULL	NULL	4		2
C	NULL	NULL	NULL		6
D	3	NULL	NULL		2
...					
ZZZ	0	1	NULL		NULL



User	f1	f2
A	8.43	1.12
B	1.23	6.32
C	0.34	4.43
D	2.45	1.23
...		
ZZ	1.12	0.76

Mengurangi dimensi dataset sedemikian rupa tanpa menghilangkan informasi penting

- f1: Menggambarkan selera terhadap film horor
- f2: Menggambarkan selera terhadap film romansa

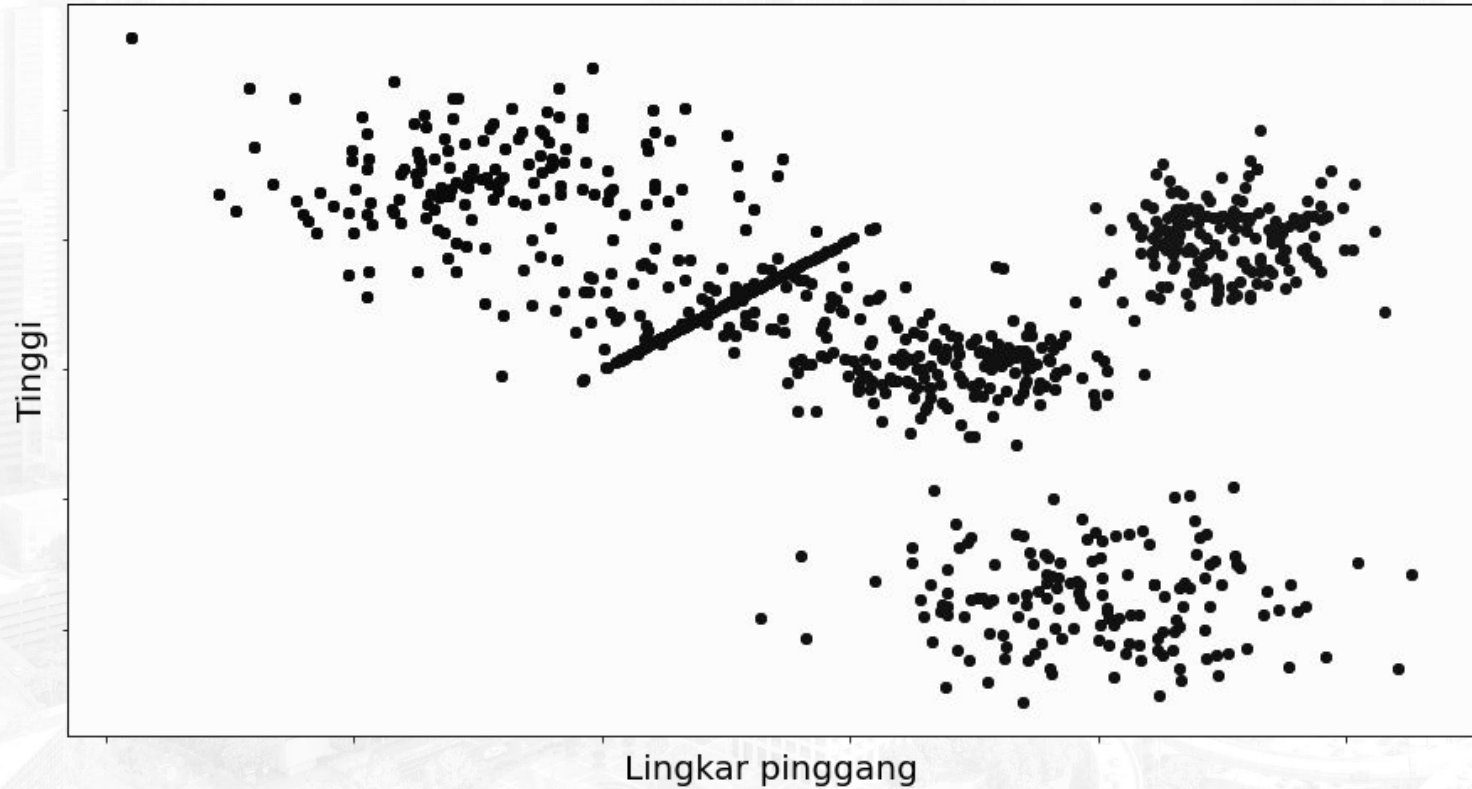
Contoh Kasus #2

U**q*o

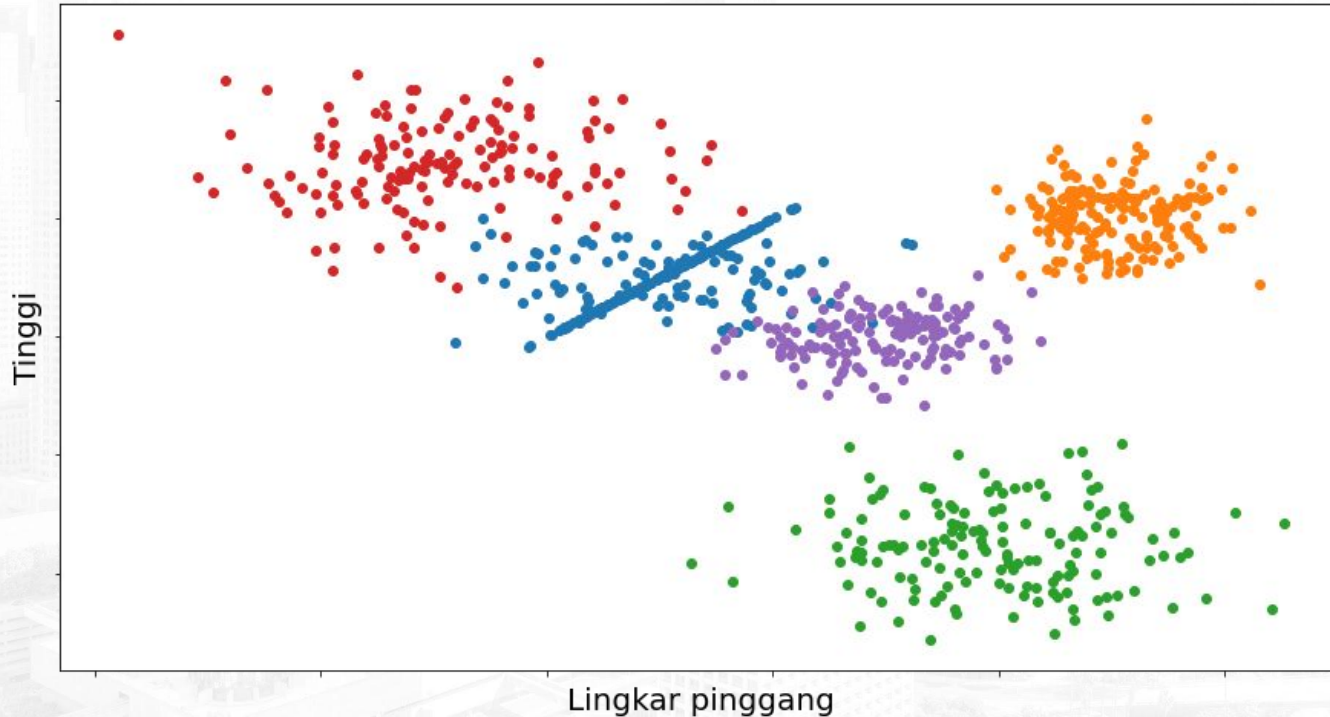
Perusahaan pakaian Uni*** akan mengeluarkan line pakaian baru untuk musim depan. Untuk mengurangi kerugian dari pakaian sisa, mereka ingin mengetahui berapa banyak mereka harus menyiapkan stok untuk setiap size dan variasi pakaian untuk line baru mereka.

Bagaimana kira-kira caranya?

Memahami Pelanggan dengan Unsupervised Learning



Memahami Pelanggan dengan Unsupervised Learning



Secara otomatis menemukan segmen!

Rencanakan desain dan stok per desain/ukuran berdasarkan segmen yang dihasilkan

Segmentasi dapat dilakukan secara manual namun **penggunaan teknik machine learning akan menghasilkan segmen yang lebih baik!**

Kelebihan Unsupervised Learning # 1

Kita dapat menggunakan lebih dari 2 feature untuk menghasilkan cluster secara otomatis dalam dimensi yang tidak dapat kita amati secara visual.

Contoh Kasus #3

Department Store Mataha*i

Department Store M***h*t* ingin meluncurkan sebuah campaign voucher diskon untuk semua pemilik kartu membernya.

Dengan harapan lebih banyak voucher diskon akan digunakan, Mat****i ingin membuat lebih dari satu jenis voucher diskon untuk memenuhi kebutuhan jenis member yang pola belanjanya berbeda.

Kira-kira bagaimana cara mulai mendesain jenis-jenis dan nominal voucher diskon untuk campaign ini?

Segmentasi Bisnis dengan Unsupervised Learning

RFM adalah salah satu metode segmentasi customer yang cukup sering digunakan di industri!

R

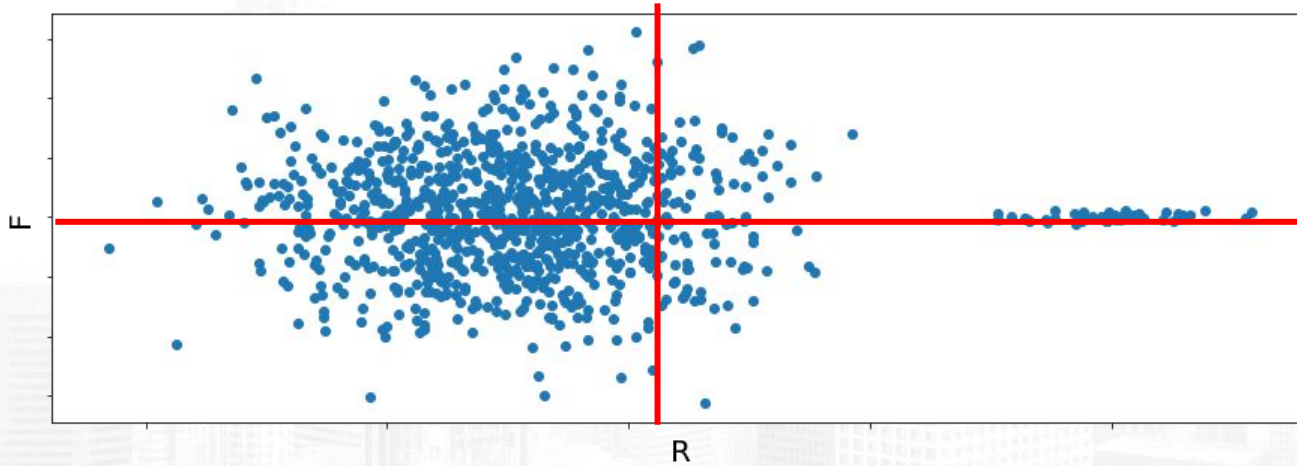
Recency - Kebaruan: Kapan terakhir kali sang pemegang kartu Matahati berbelanja?
Misal : user X terakhir transaksi 2 minggu yang lalu, maka nilai Recency adalah 14

F

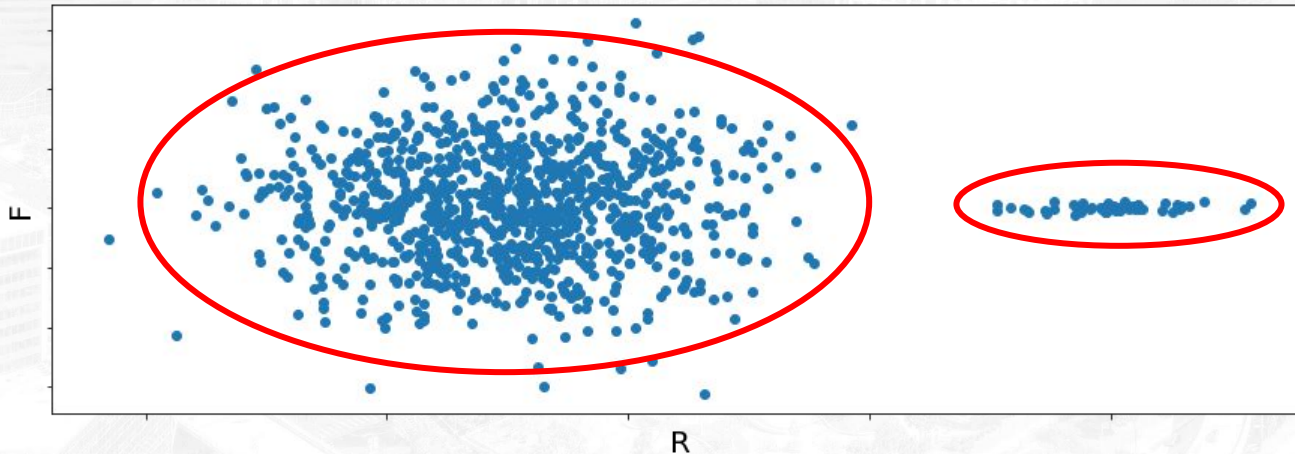
Frequency - Frekuensi: Berapa kali sang pemegang kartu Matahati belanja sebulan terakhir?
Misal : user X melakukan 5x transaksi pada bulan terakhir, maka nilai Frequency adalah 5

M

Monetary - Nilai Moneter: Berapa banyak uang yang sudah dihabiskan sang pemegang kartu Matahati?
Misal : user X menghabiskan Rp 1 juta untuk membeli produk di Matahati, nilai Monetary adalah 1 juta



Menentukan
segmen
dengan
batas
manual
(average).



Menentukan
segmen
dengan
unsupervised
learning

Segmentasi dapat dilakukan secara manual namun **penggunaan teknik machine learning akan menghasilkan segmen yang lebih baik!**

Kelebihan Unsupervised Learning # 2

Segmen yang dihasilkan akan relatif lebih homogen apabila dibandingkan dengan segmen manual. Alasannya adalah karena batasan segmen bisa bersifat tidak linear.

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Apa itu **Dimensionality
Reduction?**

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction

 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

  Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Kenapa kita ingin mengurangi jumlah fitur?

Interpretability

Data lebih mudah divisualisasikan/model lebih mudah dimengerti

Curse of Dimensionality

Menghindari kutukan dimensi

Cost-efficiency

Model lebih mudah dan murah untuk dilatih

Remove noise

Mengurangi informasi yang sebenarnya tidak penting/mengganggu

Tapi feature lebih sedikit = informasi lebih sedikit kan?

Dimensionality Reduction

Teknik pengurangan dimensi dengan jaminan bahwa informasi berguna yang terbuang akan minimal.

Dimensionality Reduction != Feature Selection

Feature selection serta merta mengurangi jumlah dimensi dengan membuang fitur mentah-mentah.

Dimensionality reduction melakukan pemrosesan pada fitur untuk menghasilkan kumpulan fitur baru dengan dalam dimensi lebih kecil.

Feature Selection

User	KKN di Desa P	PLN di Desa P	Dilan 1990	...	Dilan 2077
A	5	5	NULL		1
B	NULL	NULL	4		2
C	NULL	NULL	NULL		6
D	3	NULL	NULL		2
...					
ZZZ	0	1	NULL		NULL



User	KKN di Desa P	Dilan 1990
A	5	NULL
B	NULL	4
C	NULL	NULL
D	3	NULL
...		
ZZZ	0	NULL

Feature selection: buang semua feature kecuali KKN di Desa P dan Dilan 1990 sebagai film perwakilan

Dimensionality Reduction

User	KKN di Desa P	PLN di Desa P	Dilan 1990	...	Dilan 2077
A	5	5	NULL		1
B	NULL	NULL	4		2
C	NULL	NULL	NULL		6
D	3	NULL	NULL		2
...					
ZZZ	0	1	NULL		NULL



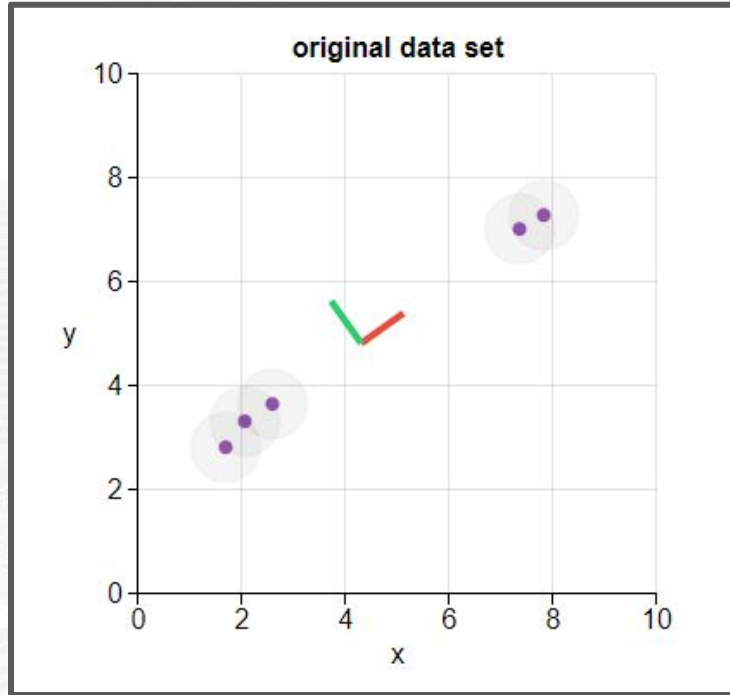
User	f1	f2
A	8.43	1.12
B	1.23	6.32
C	0.34	4.43
D	2.45	1.23
...		
ZZ	1.12	0.76

Dimensionality reduction: proses seluruh fitur untuk menghasilkan 2 fitur baru yang merepresentasikan seluruh fitur lama



**N dimensi -> N/3 dimensi tanpa
menghilangkan 'informasi'**

2 dimensi -> 1 dimensi tanpa menghilangkan 'informasi'



Informasi tersisa bahkan di ruang 1 dimensi:

- Terdapat 2 kelompok data
- 1 beranggotakan 3 titik dan 1 lagi 2 titik
- Jarak antara 2 kelompok tersebut kurang lebih sama

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik


 Evaluasi Clustering

Apa itu **Principal Component Analysis?**

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction

 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

  Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Berbagai Teknik Dimensionality Reduction

Principal Component Analysis (PCA)

Non-negative Matrix Factorization (NMF)

Lainnya

- Latent Discriminant Analysis (LDA)
- Autoencoders

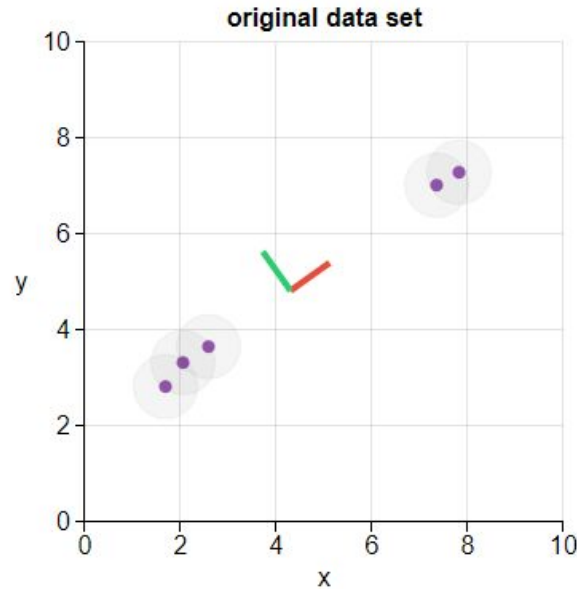
Teknik berbeda:

- Memiliki kompleksitas berbeda
- Mengawetkan 'informasi' yang berbeda

Principal Component Analysis

- Dari dataset dengan N fitur/dimensi/sumbu
 - Hasilkan dataset baru dengan N fitur/dimensi/sumbu
 - Dengan kandungan 'informasi' yang berurut
- Sehingga membuat fitur/dimensi/sumbu paling ujung menjamin informasi yang hilang akan minimal

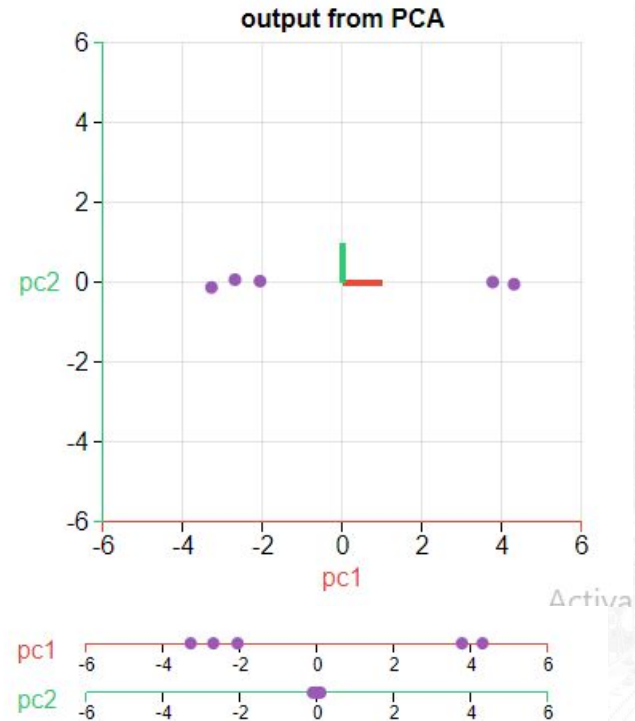
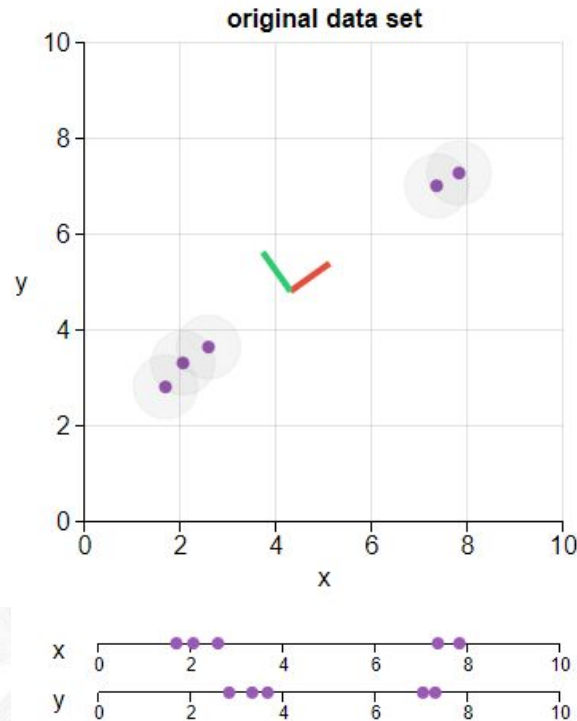
PCA: Menemukan sumbu baru dengan level informasi yang berurut



Data awal:

- 2 fitur/dimensi/sumbu: X dan Y
- Proyeksi data di sumbu X dan Y sama-sama memiliki sebaran yang signifikan

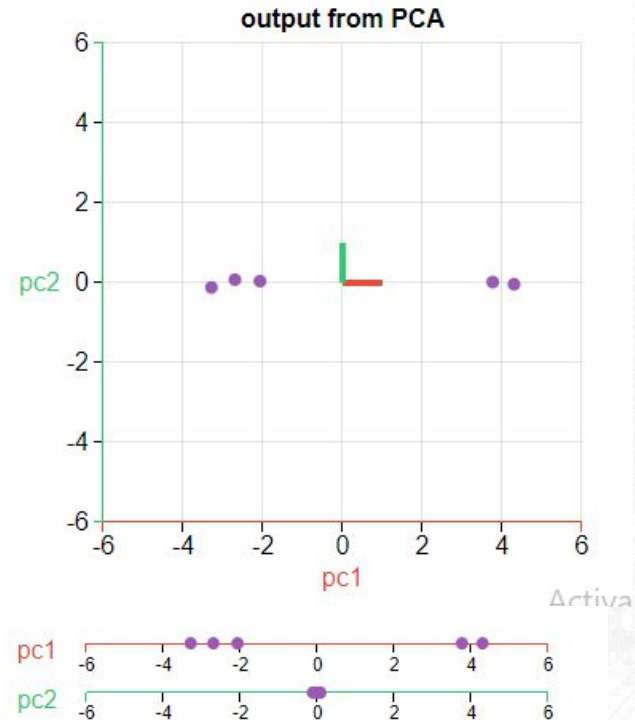
PCA: Menemukan sumbu baru dengan level informasi yang berurut



PCA: Menemukan sumbu baru dengan level informasi yang berurut

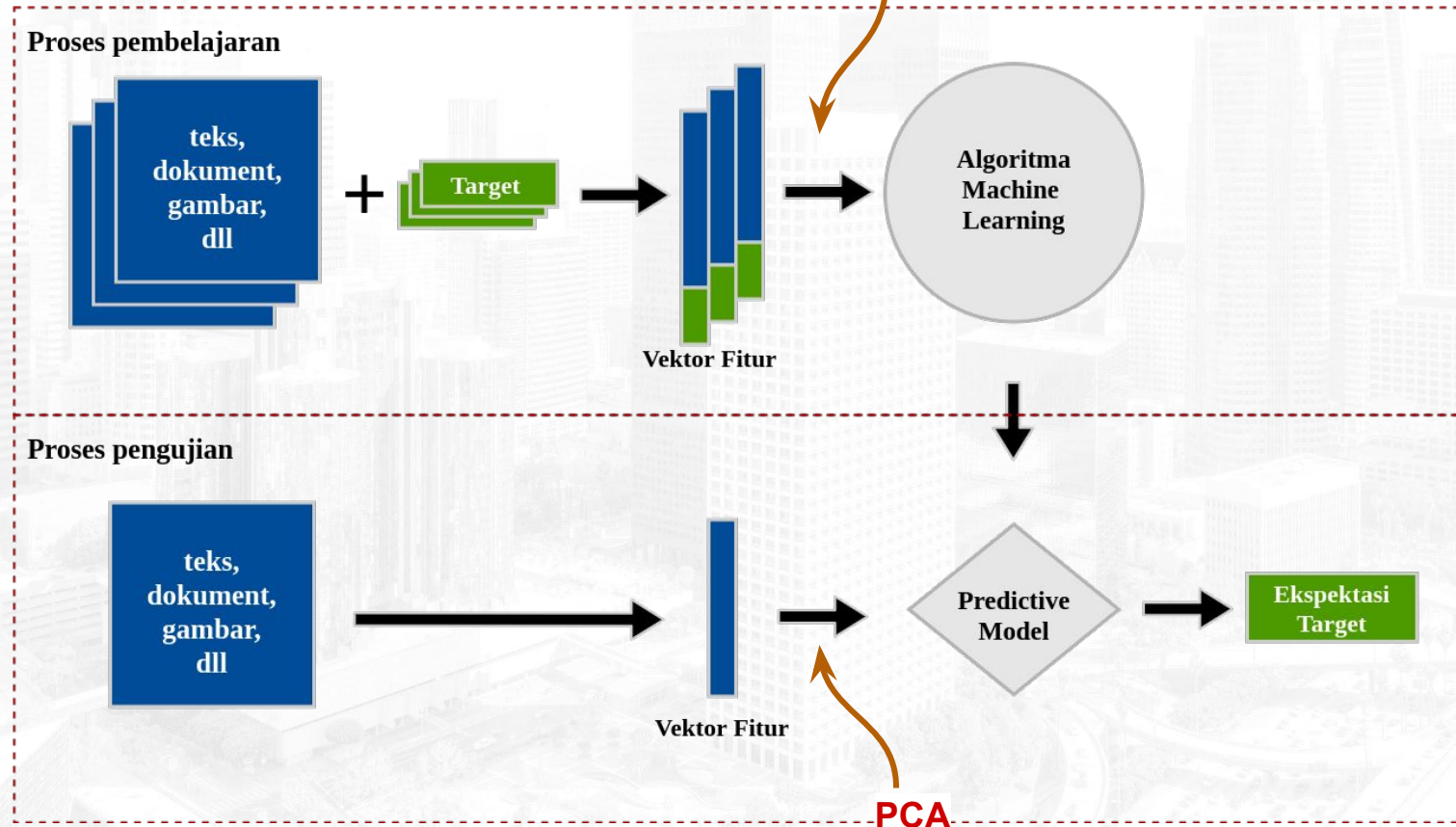
Data setelah PCA:

- 2 fitur/dimensi/sumbu: pc1 dan pc2
- Proyeksi data di sumbu pc1 memiliki sebaran
- Proyeksi data di sumbu pc2 hampir tidak memiliki sebaran



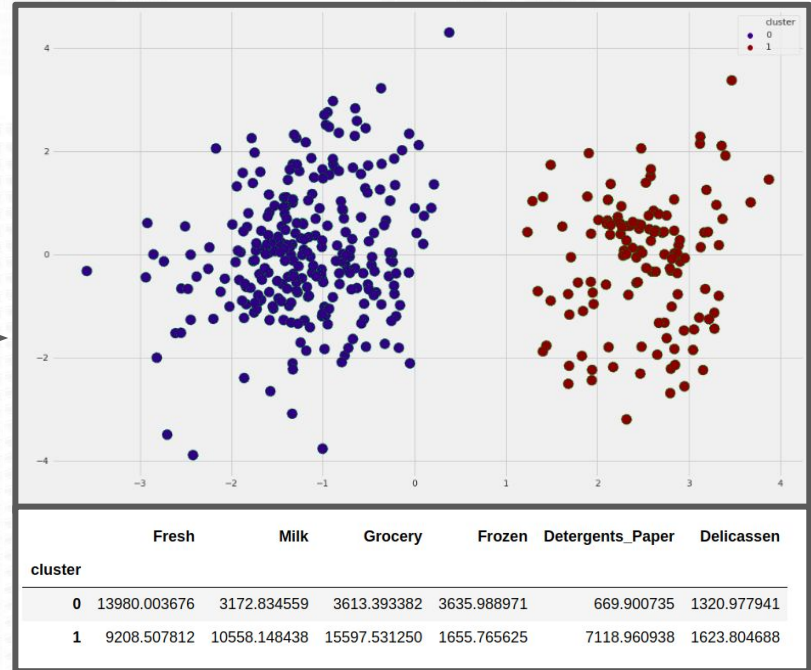
Kapan Principal Component Analysis **digunakan?**

1: PCA + Supervised



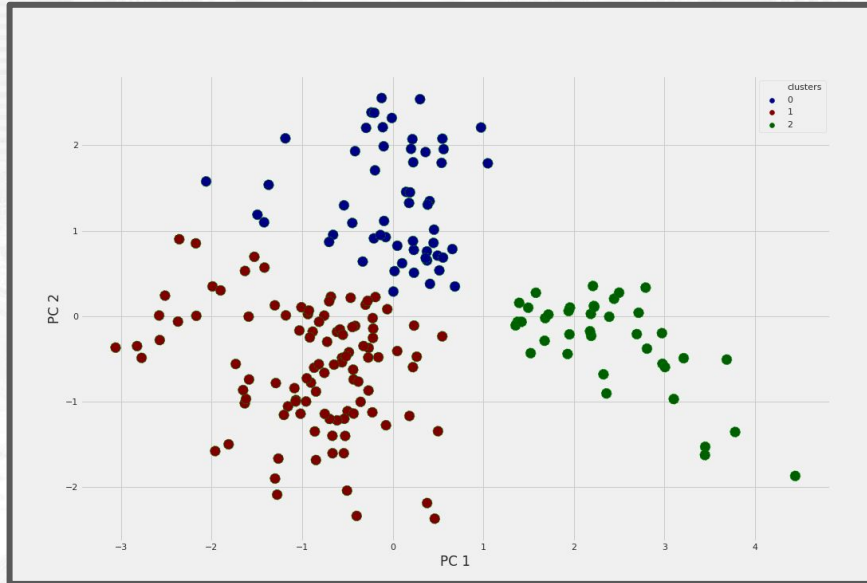
2: Clustering -> PCA -> Visualize

	ID_Customers	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	0	12669	9656	7561	214	2674	1338
1	1	7057	9810	9568	1762	3293	1776
2	2	6353	8808	7684	2405	3516	7844
3	3	13265	1196	4221	6404	507	1788
4	4	22615	5410	7198	3915	1777	5185
5	5	9413	8259	5126	666	1795	1451
6	6	12126	3199	6975	480	3140	545
7	7	7579	4956	9426	1669	3321	2566
8	8	5963	3648	6192	425	1716	750
9	9	6006	11093	18881	1159	7425	2098



3: PCA -> Clustering -> Visualize

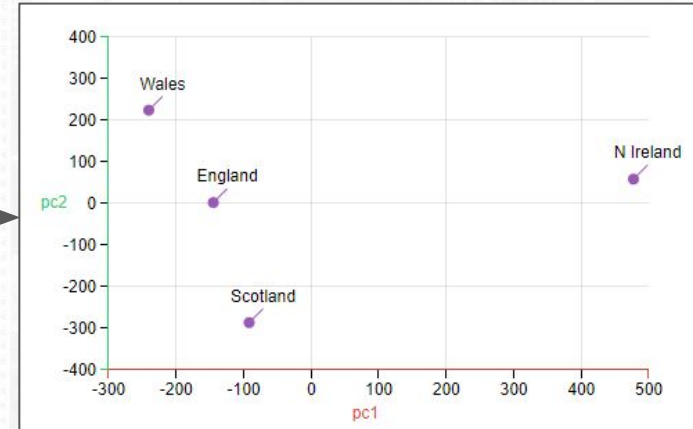
	CustomerID	Age	Annual Spending (k\$)	Spending Score (1-100)	Transaction
0	1	19	15	39	585
1	2	21	15	81	1215
2	3	20	16	6	96
3	4	23	16	77	1232
4	5	31	17	40	680



	PC 1	PC 2
0	-1.185732	2.080852
1	-0.123255	2.552804
2	-2.060763	1.576050
3	-0.235643	2.384080
4	-1.370530	1.536436

4: PCA for Data Understanding

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering


Implementasi PCA

menggunakan Python (contoh kode)

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction

 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

  PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

PCA di SKLearn



SKLearn memiliki implementasi PCA siap pakai:

- Di bawah modul `pca.decomposition`
- Cara pemakaian sangat mirip dengan standarisasi/normalisasi
- Beberapa alternatif dimensionality reduction di modul yang sama

Dataset

Iris

- **Deskripsi:**

Memprediksi spesies bunga Iris berdasarkan pengukuran kelopak.

- **Data:**

Setiap baris mewakili satu sampel, setiap kolom mewakili salah satu ukuran yang dimiliki sampel.

- **Link Kaggle:** <https://www.kaggle.com/uciml/iris>

Load Data (langsung dari URL)

```
[7] 1 url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
    2 df = pd.read_csv(url, names=['sepal length', 'sepal width', 'petal length', 'petal width', 'target'])
    3 df.sample(5)
```

	sepal length	sepal width	petal length	petal width	target
68	6.2	2.2	4.5	1.5	Iris-versicolor
65	6.7	3.1	4.4	1.4	Iris-versicolor
15	5.7	4.4	1.5	0.4	Iris-setosa
29	4.7	3.2	1.6	0.2	Iris-setosa
30	4.8	3.1	1.6	0.2	Iris-setosa

Kita sebenarnya bisa membaca data langsung melalui URL apabila data tersedia online!

Standardisasi Feature

```
1 feats = ['sepal length', 'sepal width', 'petal length', 'petal width']
2 X = df[feats].values
3 y = df['target'].values
```

```
1 from sklearn.preprocessing import StandardScaler
2 X_std = StandardScaler().fit_transform(X)
3 new_df = pd.DataFrame(data = X_std, columns = feats).head()
4 new_df.describe()
```

Untuk menggunakan PCA, kita wajib menstandarkan feature!

	sepal length	sepal width	petal length	petal width
count	1.500000e+02	1.500000e+02	1.500000e+02	1.500000e+02
mean	-4.736952e-16	-6.631732e-16	3.315866e-16	-2.842171e-16
std	1.003350e+00	1.003350e+00	1.003350e+00	1.003350e+00

Lakukan PCA!

```
1 from sklearn.decomposition import PCA
2 pcs = PCA(n_components=4).fit_transform(X_std)
3 pdf = pd.DataFrame(data = pcs, columns = ['pc1', 'pc2', 'pc3', 'pc4'])
4 pdf['target'] = y
5 pdf.describe()
```

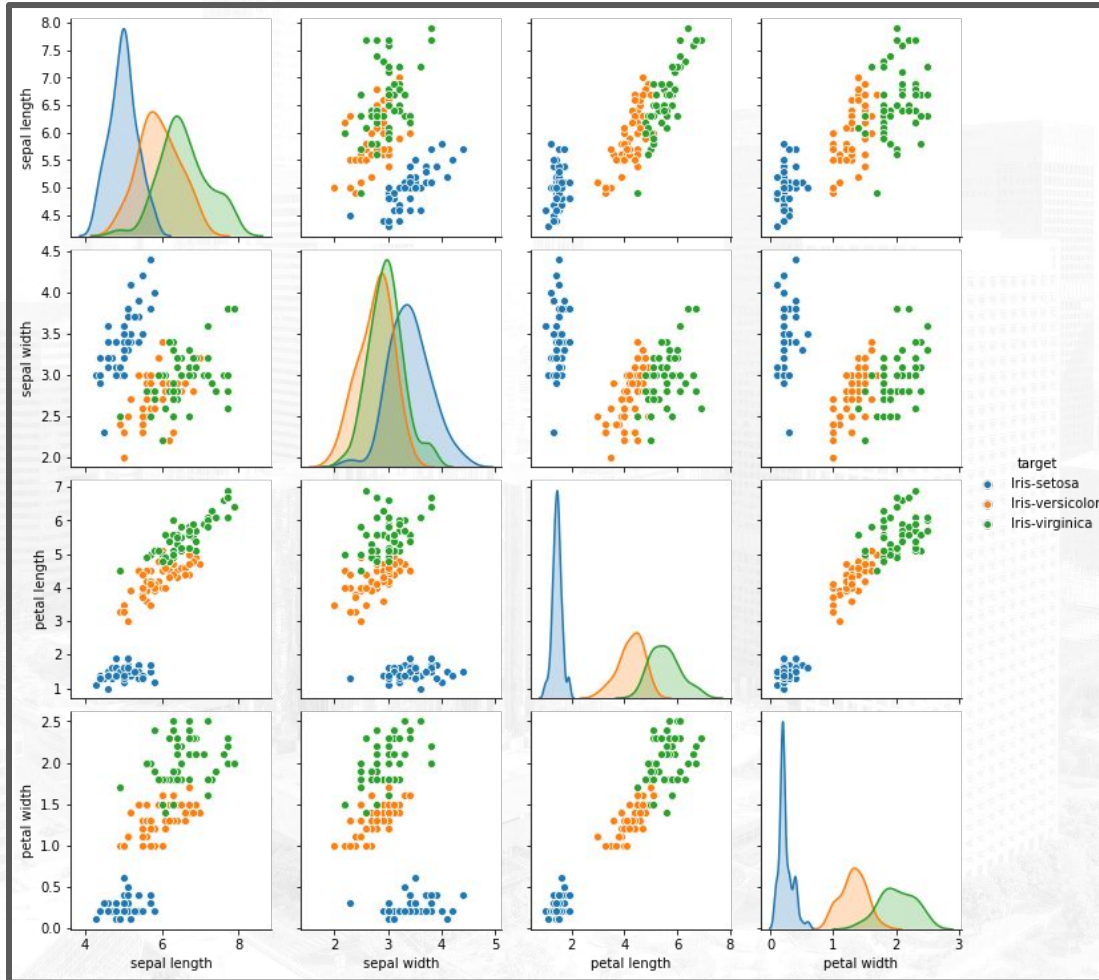
Library scikit-learn menyediakan PCA di modul sklearn.decomposition. Penggunaannya mirip dengan Scaler.

Parameter `n_components` digunakan untuk memilih banyak *principal component* yang ingin diambil.

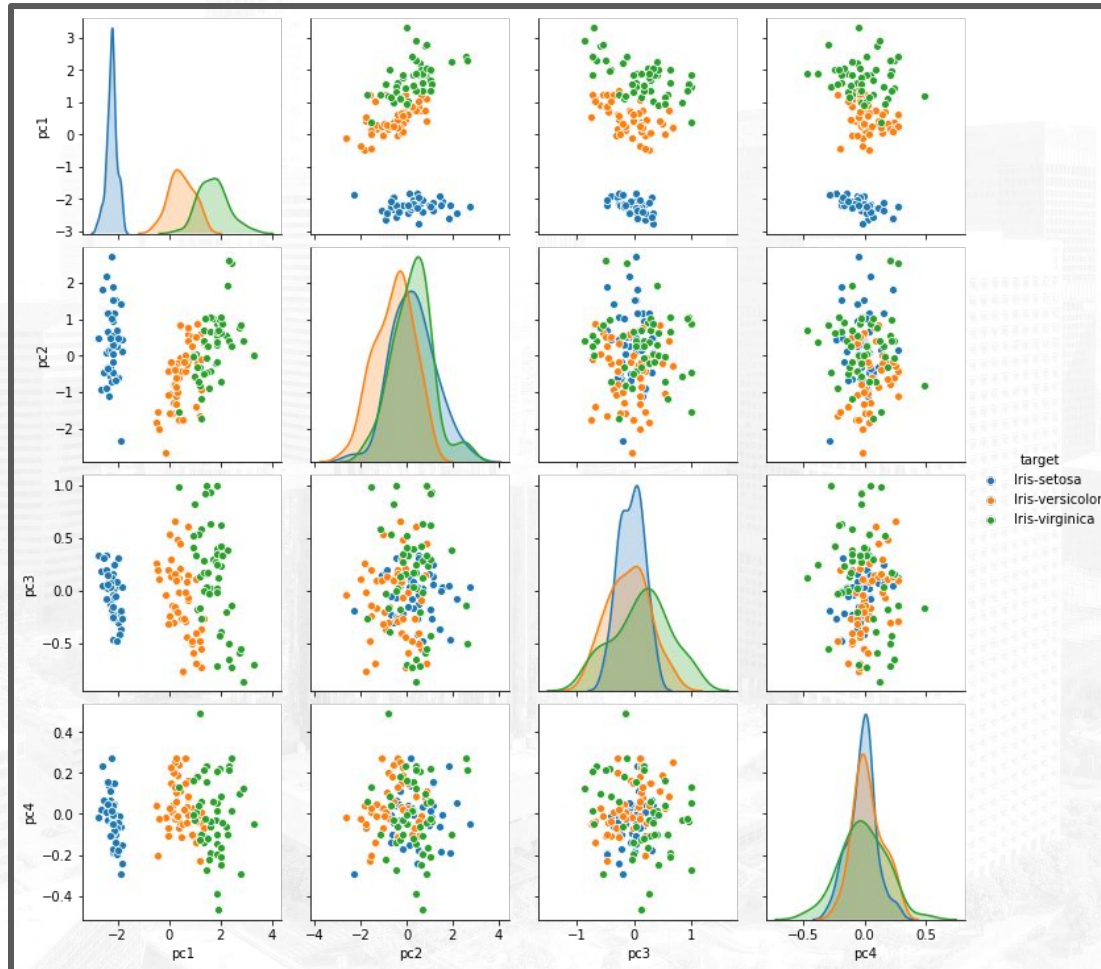
	pc1	pc2	pc3	pc4
count	1.500000e+02	1.500000e+02	1.500000e+02	1.500000e+02
mean	3.049413e-16	7.126244e-17	3.700743e-17	-7.105427e-17
std	1.711828e+00	9.630180e-01	3.851522e-01	1.440348e-01

Membandingkan hasil! Pairplot (sebelum PCA)

- Perhatikan bahwa `petal_width` dan `petal_length` memiliki informasi yang cukup mirip



Membandingkan hasil! Pairplot (setelah PCA)



- Perhatikan bahwa hampir semua 'informasi' yang diperlukan tersimpan di *principal component* pertama
- Target menjadi tidak dapat dibedakan mulai dari *principal component* kedua dan seterusnya

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Implementasi PCA menggunakan Python (live code)

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction

 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

  **PCA (Praktik)**

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Algoritma PCA **step-by-step**

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction

 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

  **Algoritma PCA Langkah-demi-langkah**

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

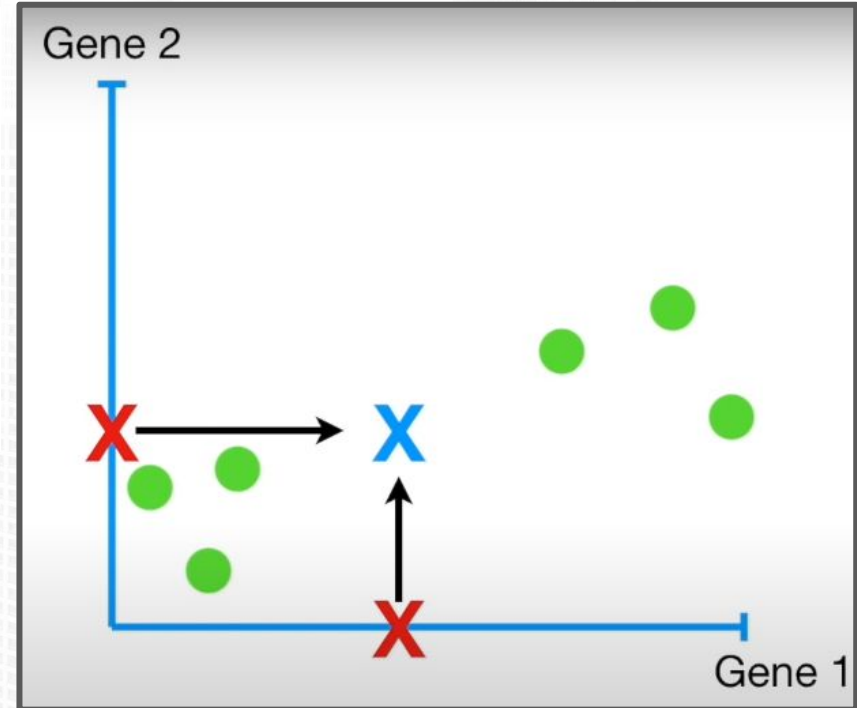
PCA Step-by-Step: Kondisi awal

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

Contoh prosedur PCA untuk sebuah dataset 2 dimensi mengenai gene tikus.

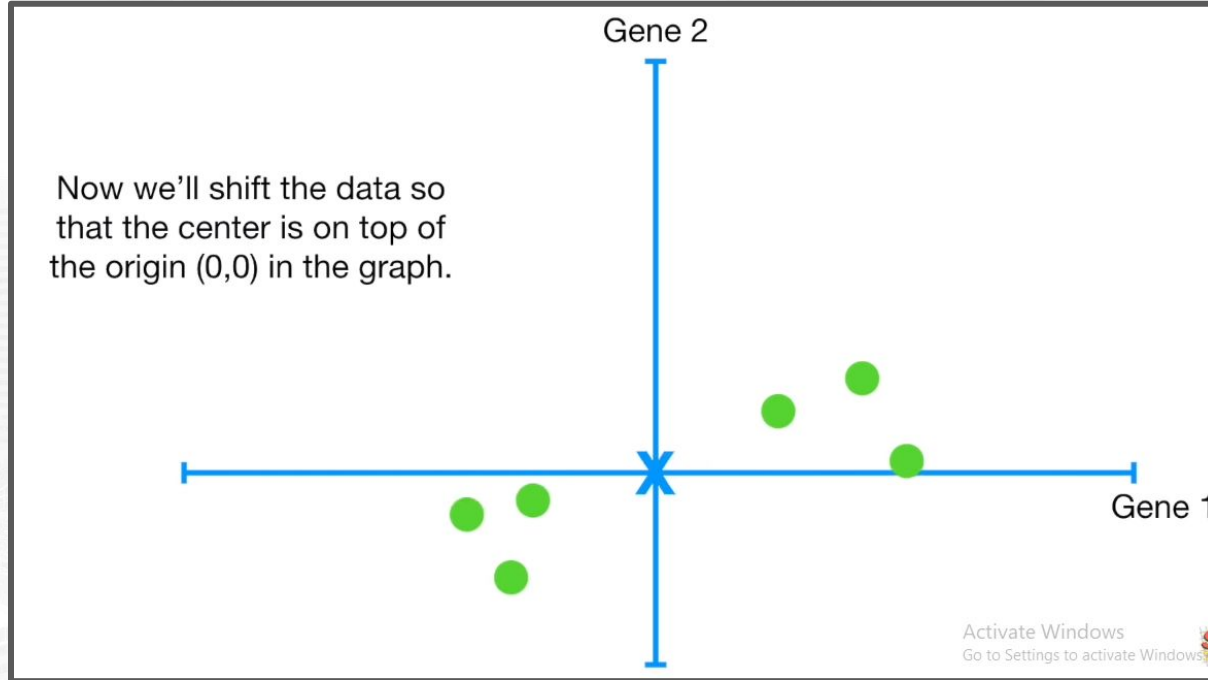
- 1 baris = informasi 1 tikus
- Kolom: Informasi gen tertentu tikus tsb

Step 1: Cari rata-rata untuk setiap sumbu dari dataset.

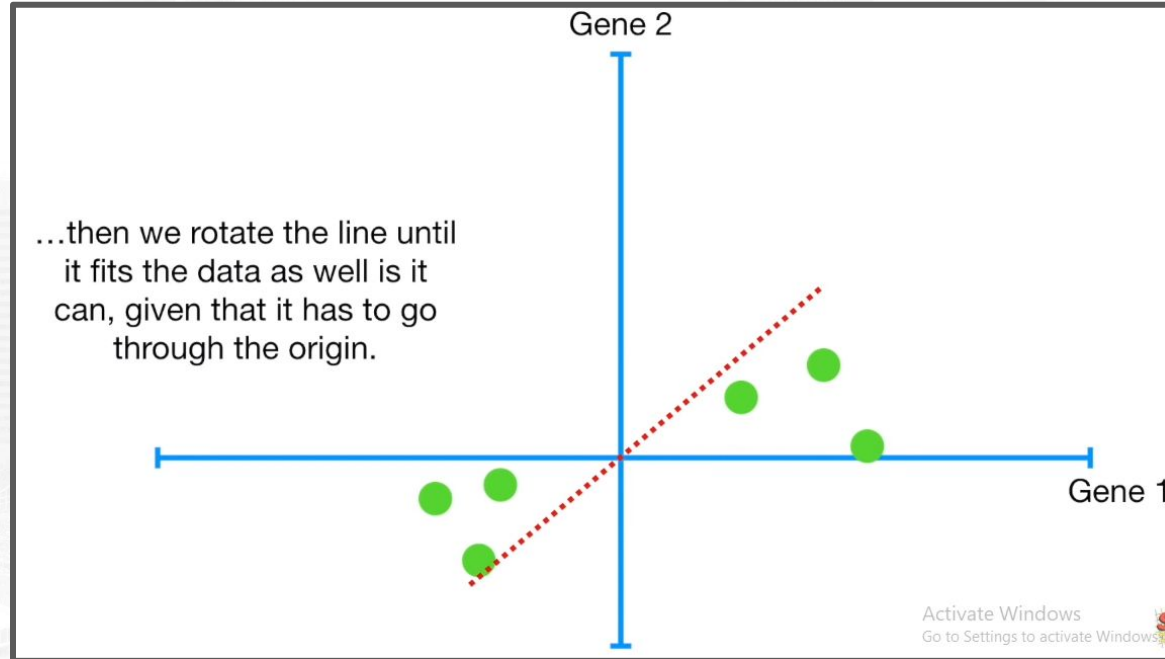


PCA Step-by-Step: Memposisikan data

Step 2: 'Geser' data sehingga posisi rata-rata ada di titik asal (0,0)



PCA Step-by-Step: Mencari PC1 - Line fitting



Step 3: Kita cari sebuah garis yang melewati titik asal (0, 0) dan paling pas dengan data (semacam regresi)

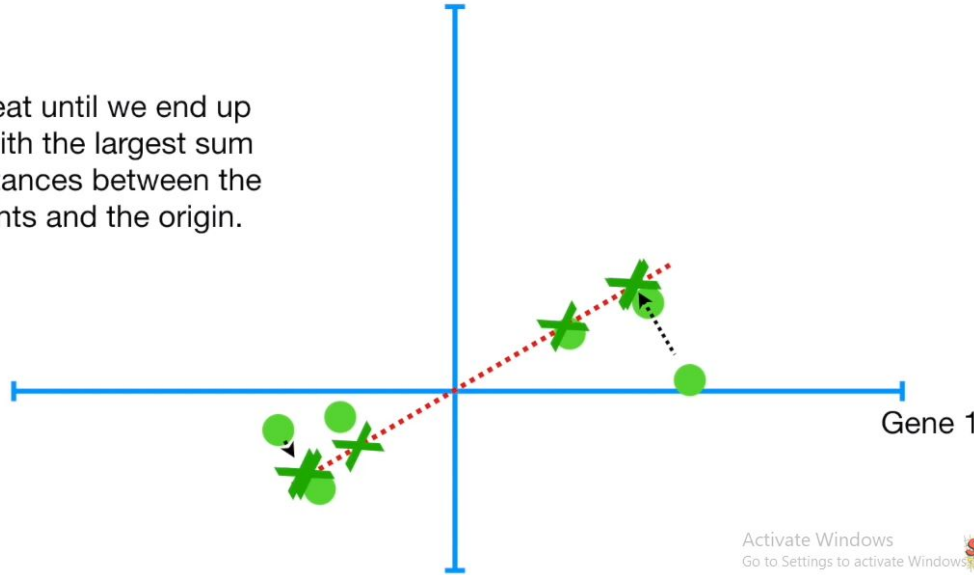
Untuk apa sih kita mencari garis yang paling pas?

Ingat bahwa kita pada dasarnya **mencari sumbu yang paling menjelaskan data.**

PCA Step-by-Step: Mencari PC1 - Fit terbaik

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(\text{distances})$$

...and we repeat until we end up with the line with the largest sum of squared distances between the projected points and the origin.

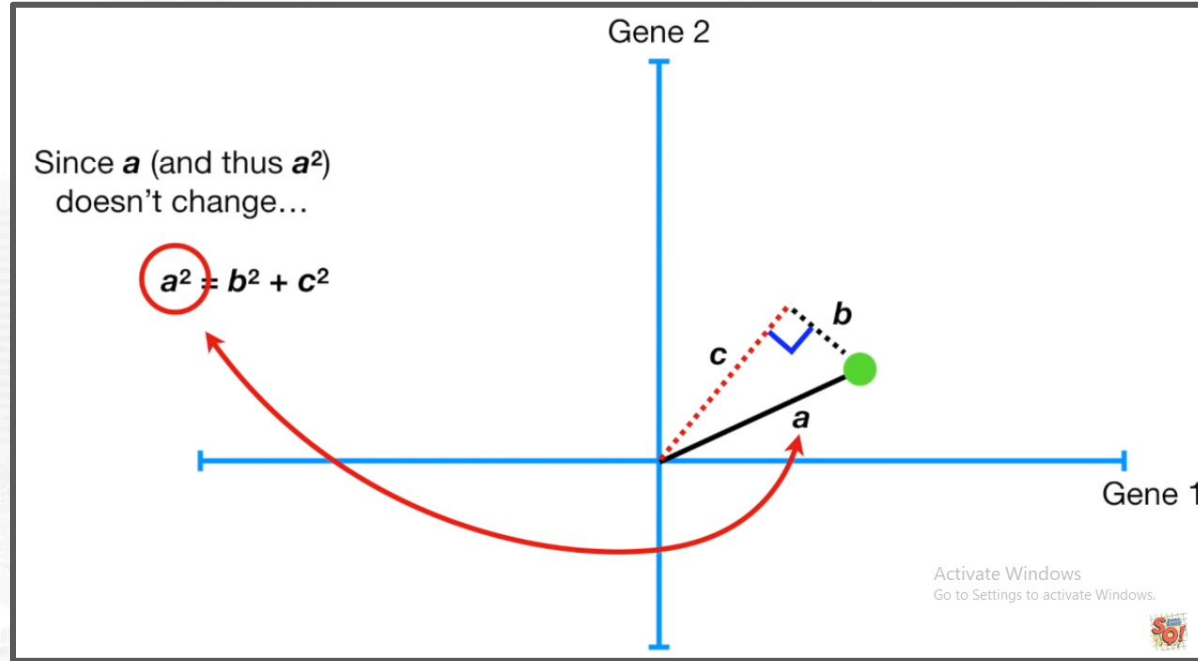


Garis mana yang paling pas dengan data? Yang sum of squared distancenya paling besar.

Proyeksikan setiap titik ke kandidat garis kita dan hitung sum of squared distancenya.

Sum of squared distance adalah jumlah dari kuadrat setiap titik ke (0, 0)

PCA Step-by-Step: Mencari PC1 - Fit terbaik (2)

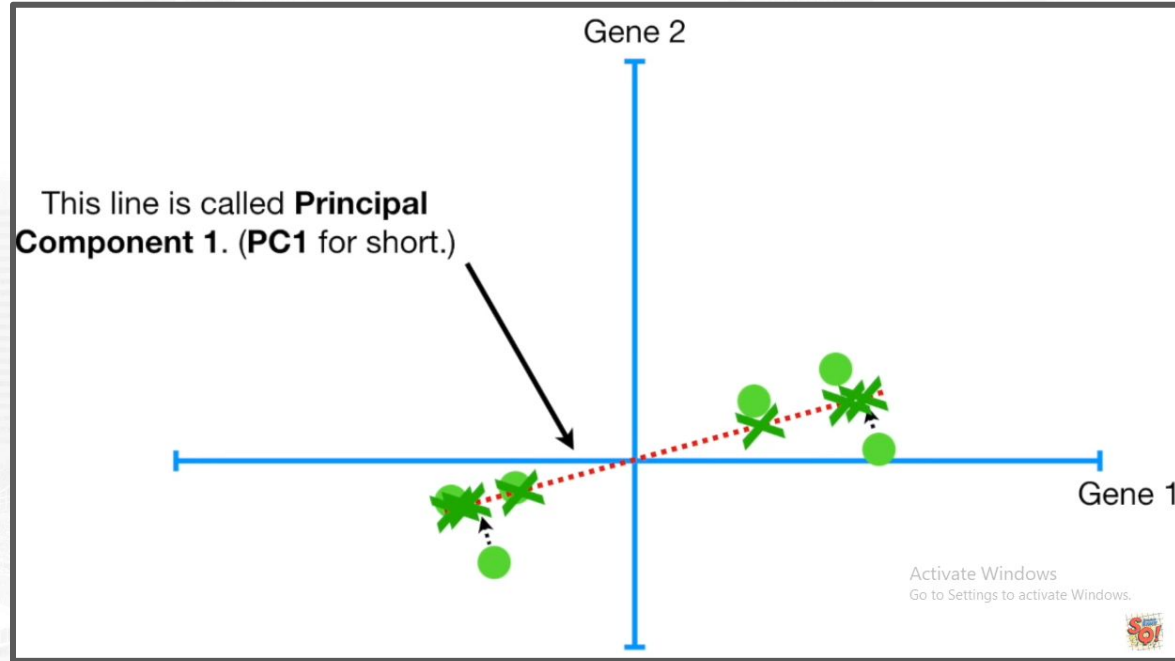


Mengapa yang paling pas adalah yang SSnya paling besar?

Lihat segitiga yang dihasilkan proyeksi titik ke kandidat garis (garis titik-titik merah)

Garis paling pas adalah ketika b paling kecil. b paling kecil adalah ketika c paling besar.

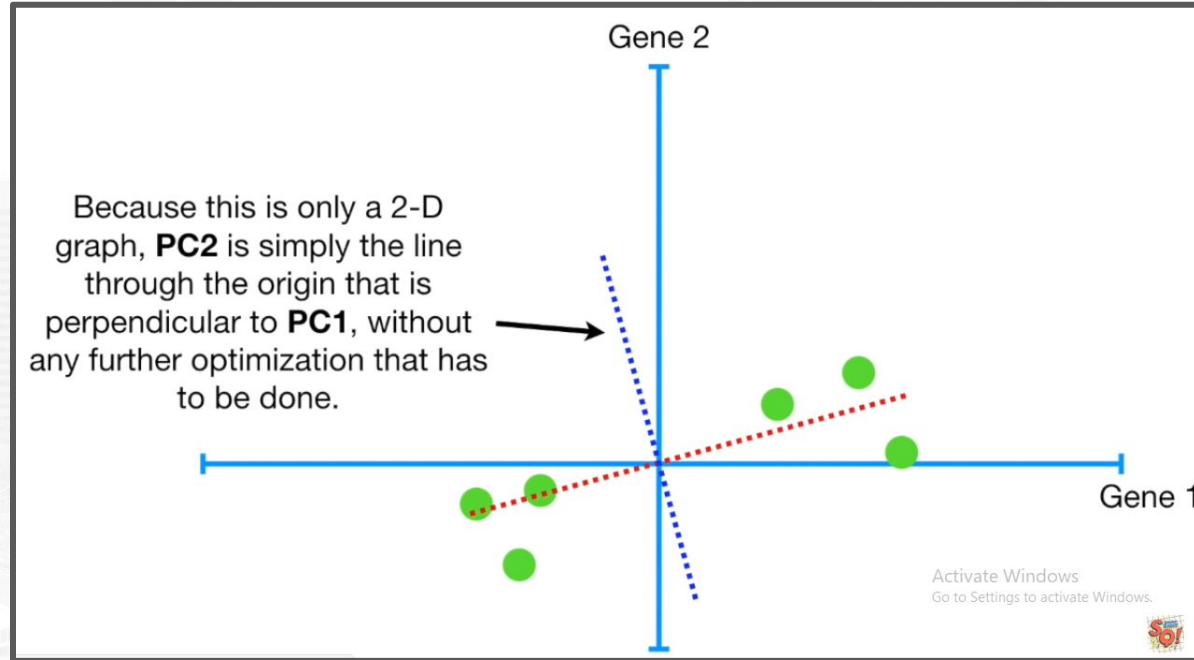
PCA Step-by-Step: PC1 GET!



Ketika kita mendapat garis paling pas, itulah PC1!

PC 1 adalah sumbu yang paling mendeskripsikan data.

PCA Step-by-Step: PC 2?

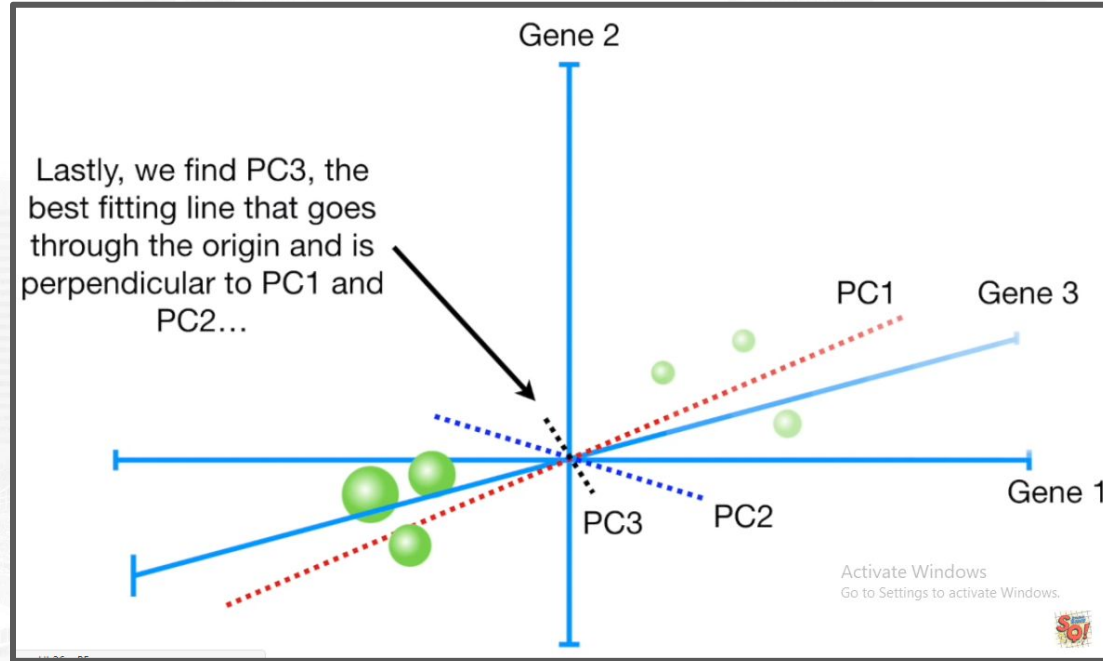


Sumbu selanjutnya (PC 2) adalah garis yang tegak lurus dengan PC 1

Dalam kasus 2 dimensi, hanya ada satu kemungkinan garis tegak lurus dengan PC 1.

Dalam kasus +2 dimensi, kita mencari lagi dengan memaksimalkan SS seperti tadi.

PCA Step-by-Step: PC 3?



Sumbu selanjutnya (PC 3) adalah garis yang tegak lurus dengan PC 1 DAN PC 2

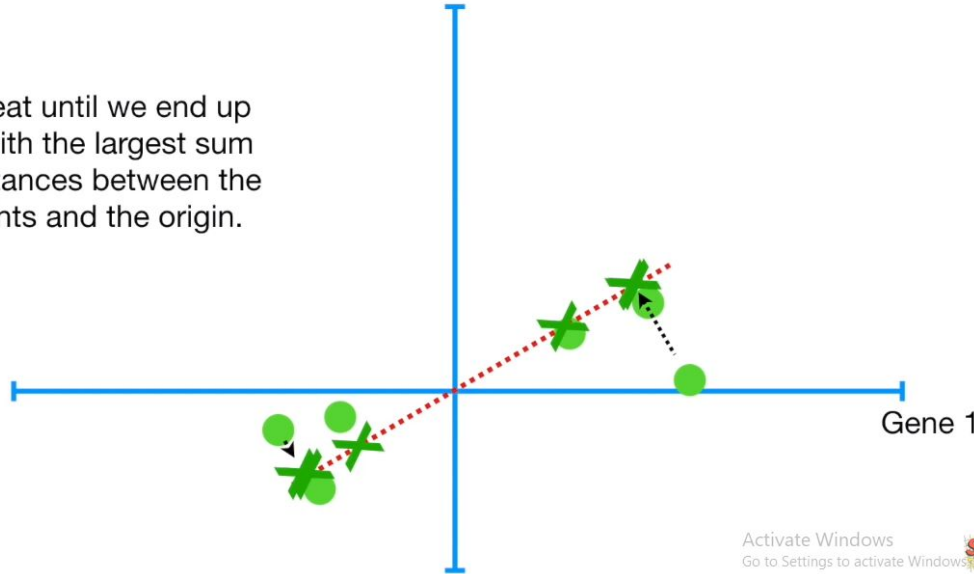
Dan seterusnya, dan seterusnya, dan seterusnya hingga PC ke-N

PCA Step-by-Step: Menghitung faedah masing-masing PC

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(\text{distances})$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

...and we repeat until we end up with the line with the largest sum of squared distances between the projected points and the origin.

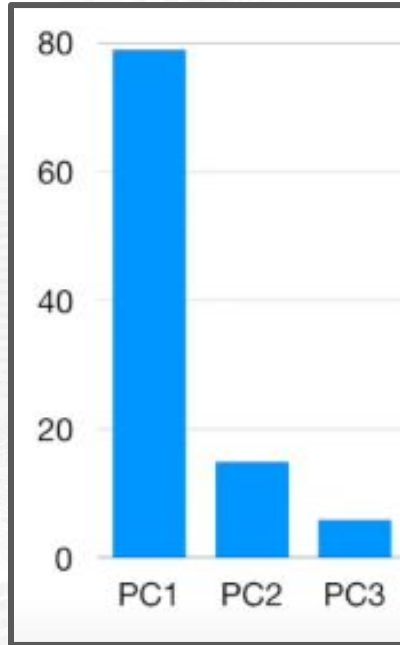


Faedah/jumlah informasi yang tersimpan dalam tiap PC adalah SS.

Kalau rumus SS terlihat familiar, itu karena memang dia sama dengan rumus untuk *variance* (bagian pembilang).

Intinya dia mendeskripsikan variasi data yang tertangkap oleh PC tertentu.

PCA Step-by-Step: Scree plot



Scree plot (kiri) menggambarkan kontribusi setiap PC dalam menjelaskan variasi dalam data.

Scree plot bisa berupa persentase, bisa berupa SSnya langsung.

Menghitung persentasenya gampang, hitung SS untuk PC itu / total SS untuk semua PC.

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering

Tips/Trick + Review

Preprocessing Wajib untuk PCA

DATA CLEANING

Data tidak boleh **bolong** ataupun memiliki nilai NULL. Hal ini karena kita harus menghitung jarak untuk setiap baris data dan apabila ada data yang kosong jarak tidak bisa dihitung.

STANDARDISASI DATA

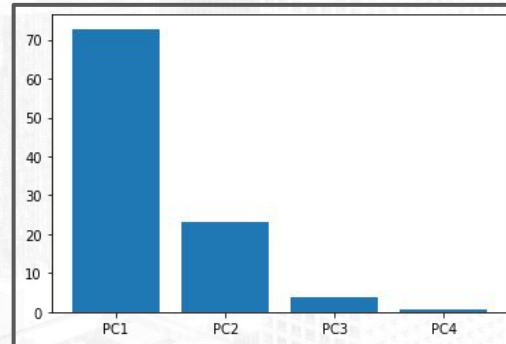
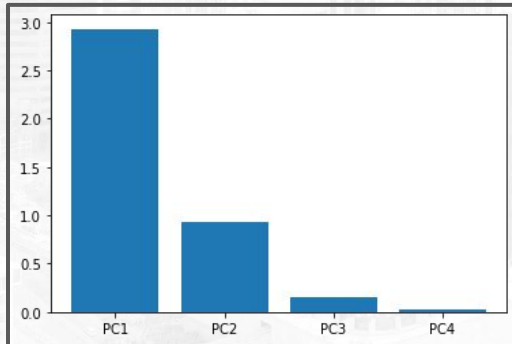
data harus distandarisasi, jika data tidak distandarisasi maka bisa ada yang mendominasi perhitungan PCA

Berapa principal component yang harus dipakai?

```
1 print('Explained variance:', pca.explained_variance_)
2 print('Explained variance ratio:', pca.explained_variance_ratio_)

1 plt.bar(['PC1', 'PC2', 'PC3', 'PC4'], pca.explained_variance_)

1 plt.bar(['PC1', 'PC2', 'PC3', 'PC4'], pca.explained_variance_ratio_ * 100)
```



Gunakan Scree Plot untuk menentukan berapa PC yang harus diambil

Ambil PC hingga sekitar 90% variasi tercover.

Atau ambil PC yang SSnya masih di atas 1.

Kapan pakai PCA dan kapan tidak pakai PCA?

PAKAI PCA KALAU

- Perlu lebih sedikit *feature* dari yang dimiliki sekarang karena satu dan lain alasan namun tidak bisa membuang langsung kolom
- Fitur-fitur dari data memiliki korelasi satu sama lain, kira-kira di atas 0.3

JANGAN PAKAI PCA KALAU

- Tidak ingin data jadi lebih sulit untuk dimengerti/artikan
- Fitur-fitur dari data tidak memiliki korelasi satu sama lain, di bawah 0.3. Kalau dipakai, PCA hanya akan mengeluarkan komponen yang sama persis dengan sumbu awal.
- Dalam kasus klasifikasi, variasi data di dalam kelas lebih besar daripada variasi data antar kelas - PCA bisa malah membuang variasi antar kelas

Topik Unsupervised Learning

Sesi I: Intro + Dimensionality Reduction


 Definisi dan Jenis-jenis Unsupervised Learning

 Contoh Kasus Unsupervised Learning

 Dimensionality Reduction dan Penggunaannya

 Intuisi dan Motivasi Principal Component Analysis (PCA)

 PCA (Praktik)

 Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

 Intuisi dan Motivasi Clustering

 Clustering dan Segmentasi dalam Bisnis

 Intermezzo: Pengukuran Jarak

 Algoritma Agglomerative Clustering dan Praktik

 Algoritma K-means Clustering dan Praktik

 Evaluasi Clustering



Terima Kasih!



Hafizh Adi Prasetya

<https://id.linkedin.com/in/hafizhadi>

Hafizh Adi Prasetya

Data Scientist

Bukalapak