

PRESENTASI

RINGKASAN HASIL COURSE

oleh: Rere Arga Dewanata

Data Analysis with Python (Coursera)

IBM.

Data Analysis with Python

This course is part of multiple programs. [Learn more](#)

 Instructor: [Joseph Santarcangelo](#)

Go To Course Already enrolled Financial aid available

360,678 already enrolled

Course
Gain insight into a topic and learn the fundamentals

4.7 ★ (17,081 reviews) |  94%

Beginner level
Recommended experience [\(i\)](#)

14 hours (approximately)
Learn at your own pace

Flexible schedule
Learn at your own pace

Progress towards a degree
[Learn more](#)

[View course modules](#)



Hal yang dipelajari

01

Mengumpulkan dan Mengimpor Data

Mempelajari berbagai cara untuk mengakses dan mengimpor data ke dalam lingkungan Python dari berbagai sumber

02

Pembersihan, Persiapan, dan Format Data

Mempelajari tahap persiapan data untuk analisis, termasuk agar sesuai dengan kebutuhan

03

Manipulasi Dataframe

Mempelajari mengenai memilih, menggabungkan, dan mengubah data dalam DataFrame.

04

Meringkas Data

Mempelajari statistik deskriptif dan teknik visualisasi data untuk menggambarkan informasi penting dalam data

05

Membuat Model Regresi

Mempelajari mengenai pembuatan model regresi yang sesuai dengan data yang ada dan bagaimana mengevaluasi kinerja model tersebut.

06

Penyempurnaan Model

Mempelajari mengenai teknik dalam menyempurnakan model yang telah dibuat dengan mengubah hiperparameter dan evaluasi kritis terhadap model yang telah dibangun.

07

Pembuatan Data Pipelines

Mempelajari mengenai penggabungan tahap-tahap berikutnya agar mampu membuat alur kerja yang efektif

Dataset yang Digunakan





Automobile

Donated on 5/18/1987

From 1985 Ward's Automotive Yearbook

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Other	Regression
Feature Type	# Instances	# Features
Categorical, Integer, Real	205	25

Dataset Information ^

Additional Information

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. ...

[SHOW MORE](#) ▾

Has Missing Values?

Yes

UCI Machine Learning Repository
Discover datasets around the world!
ics.uci.edu

Library Python yang Digunakan Komputasi Saintifik

Pandas



library yang Digunakan untuk memudahkan dalam mengimpor, membersihkan, menyaring, dan mengolah data tabular dengan adanya dataframe

Numpy



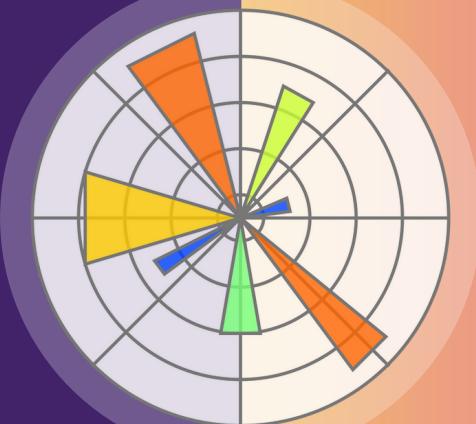
Digunakan untuk memudahkan dalam membuat matriks dan mengoperasikan matriks

Scipy



Ekstensi dari Numpy yang digunakan untuk menyelesaikan permasalahan integral, optimasi, persamaan diferensial

Library Python yang Digunakan Visualisasi



Matplotlib

Digunakan untuk mengubah data menjadi visualisasi grafik mulai dari grafik garis sederhana hingga grafik batang, scatter plot, heatmap, dan lainnya



Seaborn

Library yang dibangun diatas matplotlib. Digunakan untuk membuat visualisasi statistik yang menarik dan informatif yang mudah digunakan untuk menghasilkan plot seperti box plot, violin plot, dan pair plot

Library Python yang Digunakan Algoritma

Scikit-learn

Digunakan untuk menyediakan berbagai algoritma machine learning untuk berbagai tugas, seperti klasifikasi, regresi, dan klastering,

Statsmodels

Library yang dapat digunakan untuk menjelaskan hubungan antara variabel, menguji hipotesis, dan membuat prediksi berdasarkan data

Perbedaan Tipe Data pada Pandas dan Native Python

Pandas	Native Python
object	string
int64	int
float64	float
datetime64/timedelta [ns]	-

Import/Export Dataset menggunakan Pandas

Format Data	Import	Export
CSV	<code>pd.read_csv()</code>	<code>df.to_csv()</code>
JSON	<code>pd.read_json()</code>	<code>df.to_json()</code>
Excel	<code>pd.read_excel()</code>	<code>df.to_excel()</code>
SQL	<code>pd.read_sql()</code>	<code>df.to_sql()</code>

Hasil Import



```
df.head(10)
```

		symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke							
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68								
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68								
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47								
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40								
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40								
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19	3.40								
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi	3.19	3.40								
7	1	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136	mpfi	3.19	3.40								
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi	3.13	3.40								
9	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131	mpfi	3.13	3.40								

10 rows × 26 columns



Beberapa Method Dataframe (df)

Pandas untuk Melihat Kondisi Data

Method	Fungsi	Keterangan
df.dtypes	Mengecek tipe data	-
df.describe()	Melihat ringkasan statistik	Menambahkan parameter (include= "all") agar melihat lebih lengkap
df.info()	Melihat ringkasan lengkap (terdapat tipe data dan jumlah nilai null)	-

+ Mengatasi Nilai yang Hilang Drop +

```
df.dropna(subset=["price"], axis=0, inplace=True)
```

highway-mpg	price
20	23875
22	NaN
29	16430



axis = 0 -> Drop row
axis = 1 -> Drop column
inplace=True -> Menimpa data

highway-mpg	price
20	23875
29	16430

+ Mengatasi Nilai yang Hilang Mengganti +

```
mean = df[“normalized-losses”].mean()  
df[“normalized-losses”].replace(np.nan,mean)
```

normalized-losses	make
164	audi
NaN	audi
158	audi



normalized-losses	make
164	audi
162	audi
158	audi

Beberapa Cara Scaling Data

Agar data antar fitur terdapat pada rentang yang mirip



Simple

$$x_{new} = \frac{x_{old}}{x_{max}}$$

Min-Max

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$

Z-Score

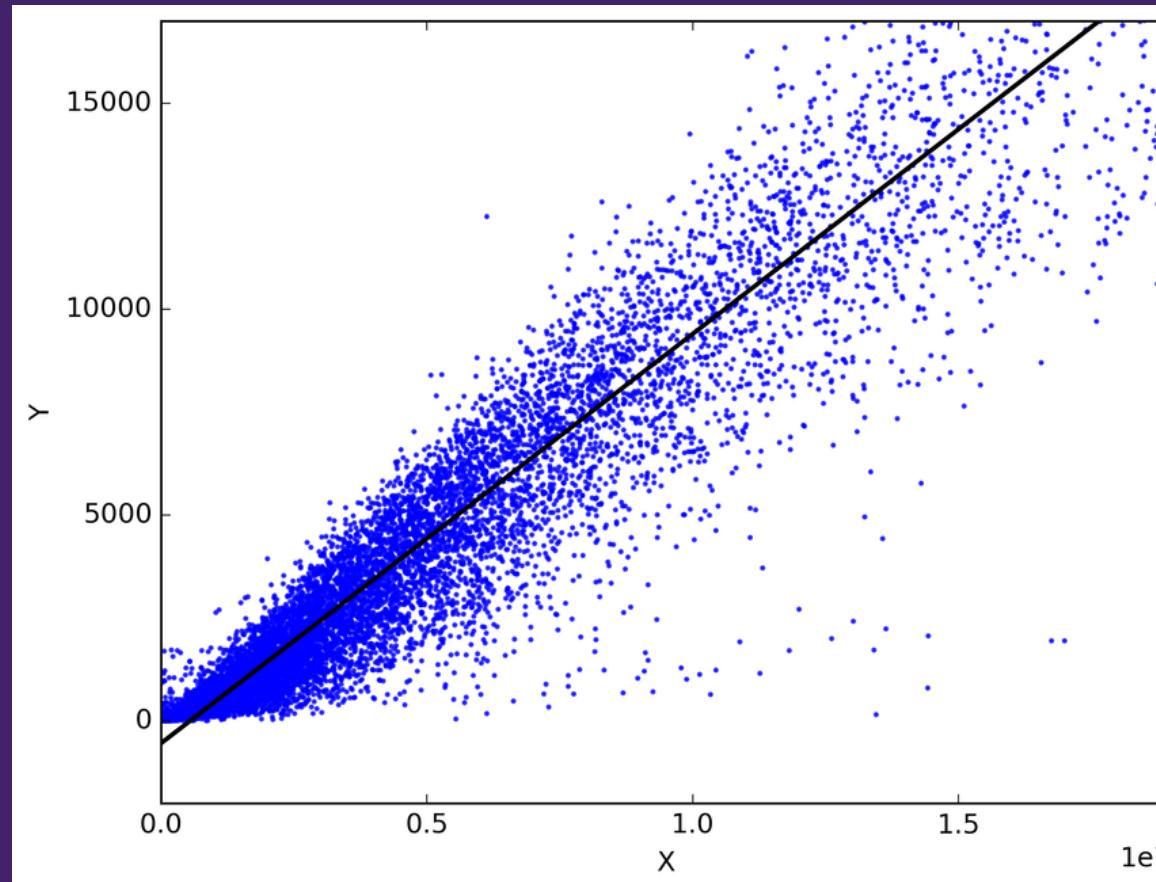
$$x_{new} = \frac{x_{old} - \mu}{\sigma}$$



Regresi

Simple Linear Regression

$$y = b_0 + b_1 x$$



**y → variabel target
x → variabel prediktor**

**b1 → gradien
b0 → intercept/bias**

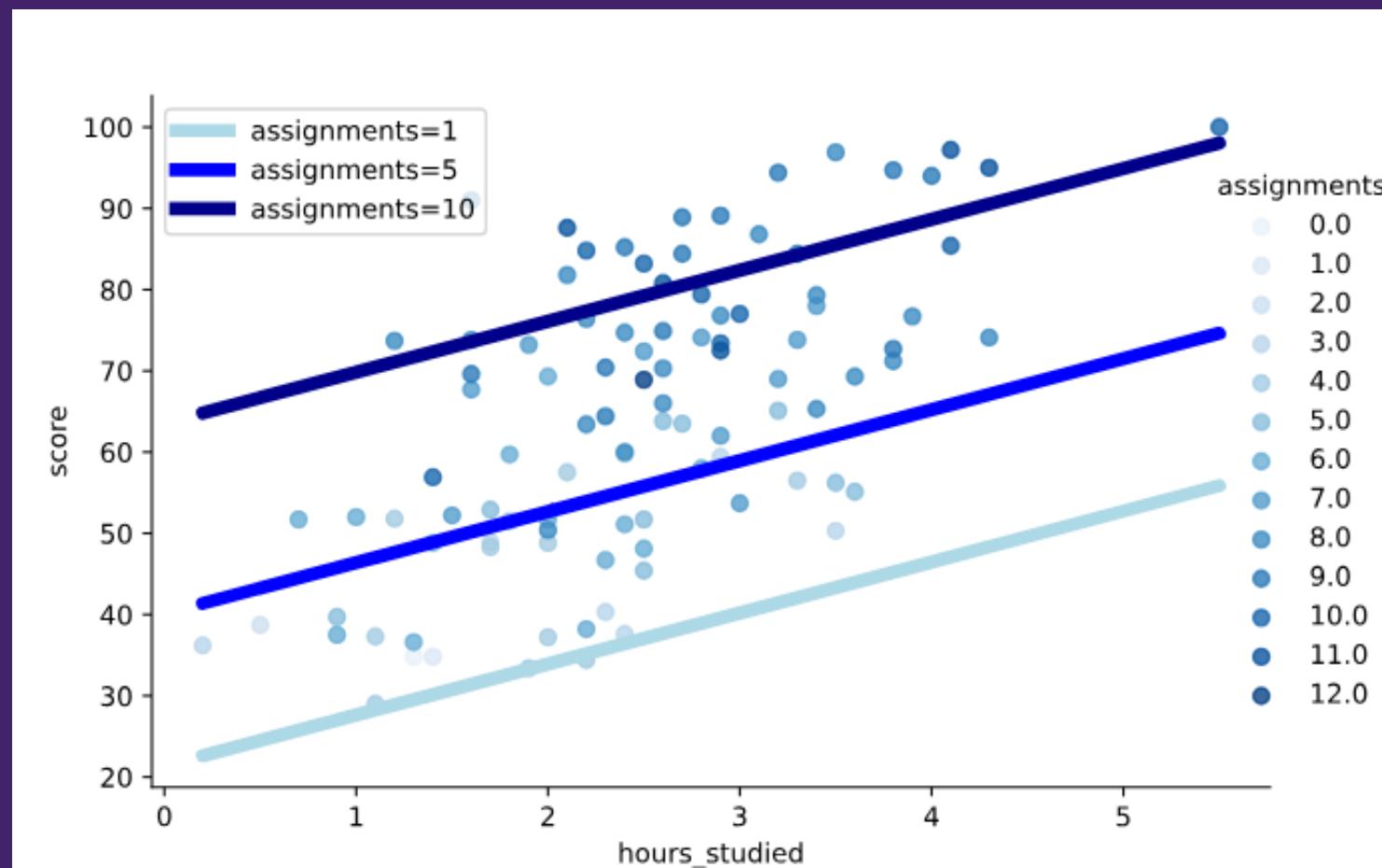


Regresi

Multiple Linear Regression



$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$



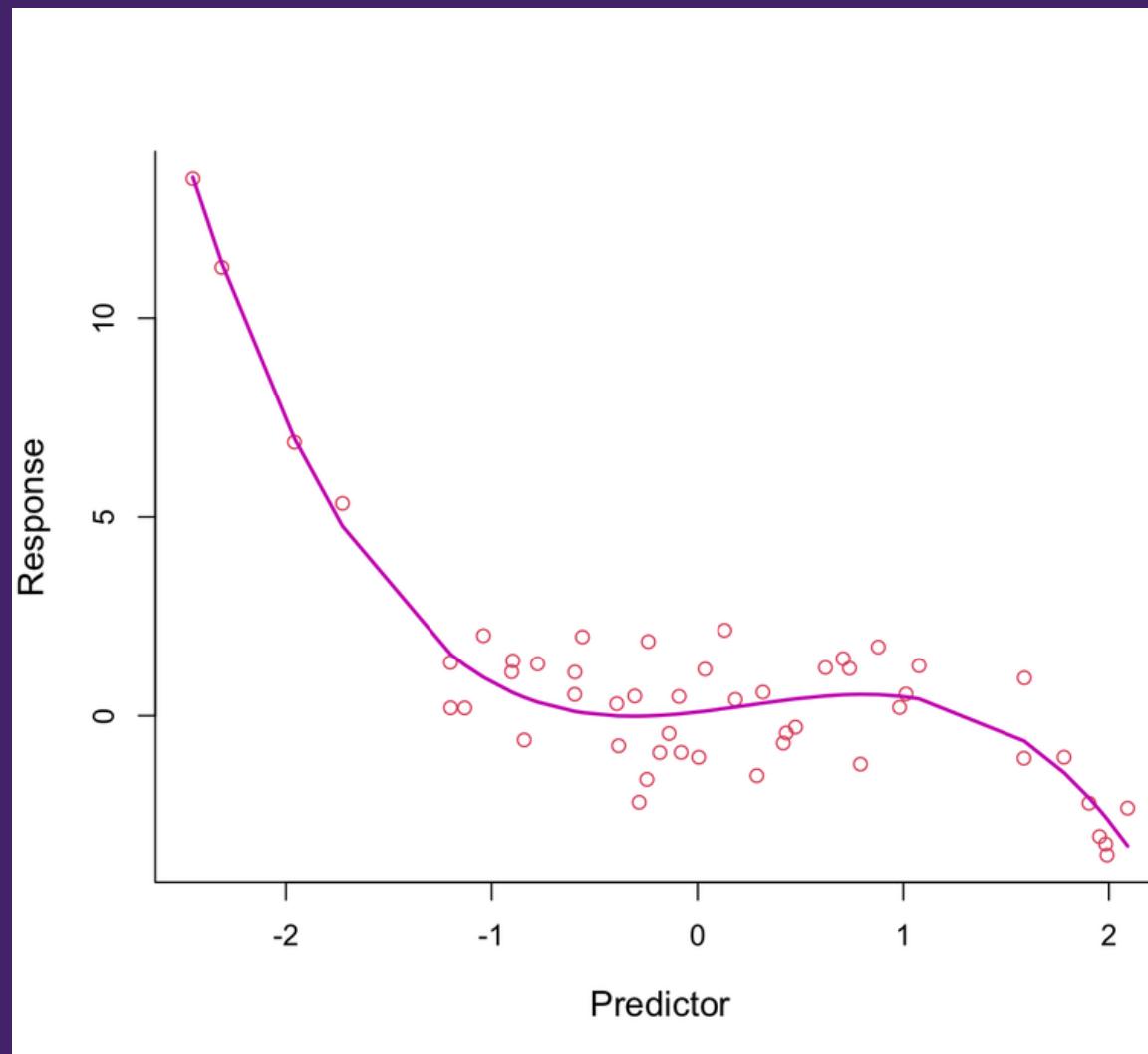
y → variabel target
x1 → variabel prediktor
x2 → variabel prediktor

b1 → koefisien X1
b2 → koefisien x2
b0 → intercept/bias

Regresi

Polynomial Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \dots + b_n x_1^n$$



y -> variabel target
x1 -> variabel prediktor

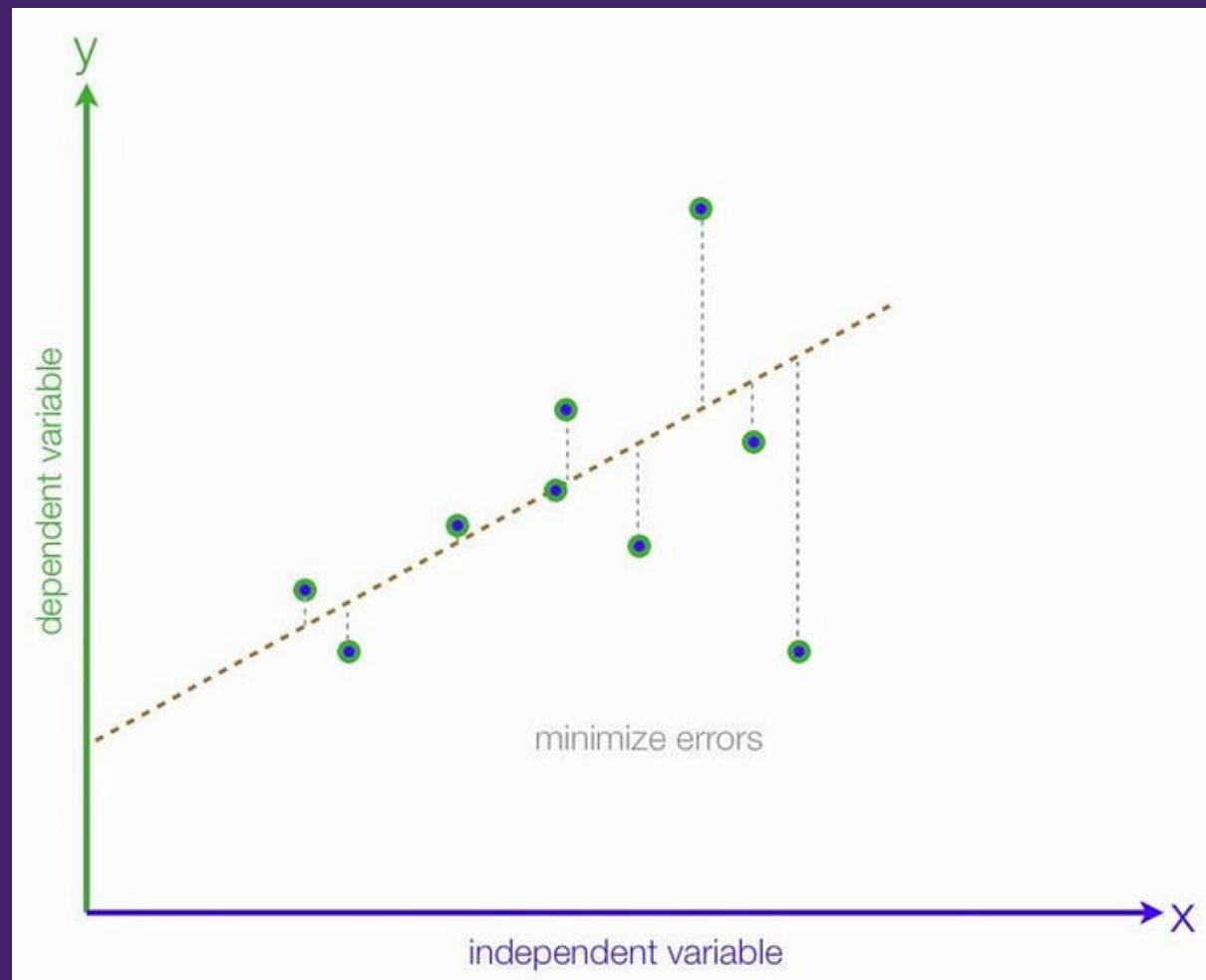
b1 -> koefisien X1
b2 -> koefisien x1²
b0 -> intercept/bias



Evaluasi Regresi

Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$



Mean

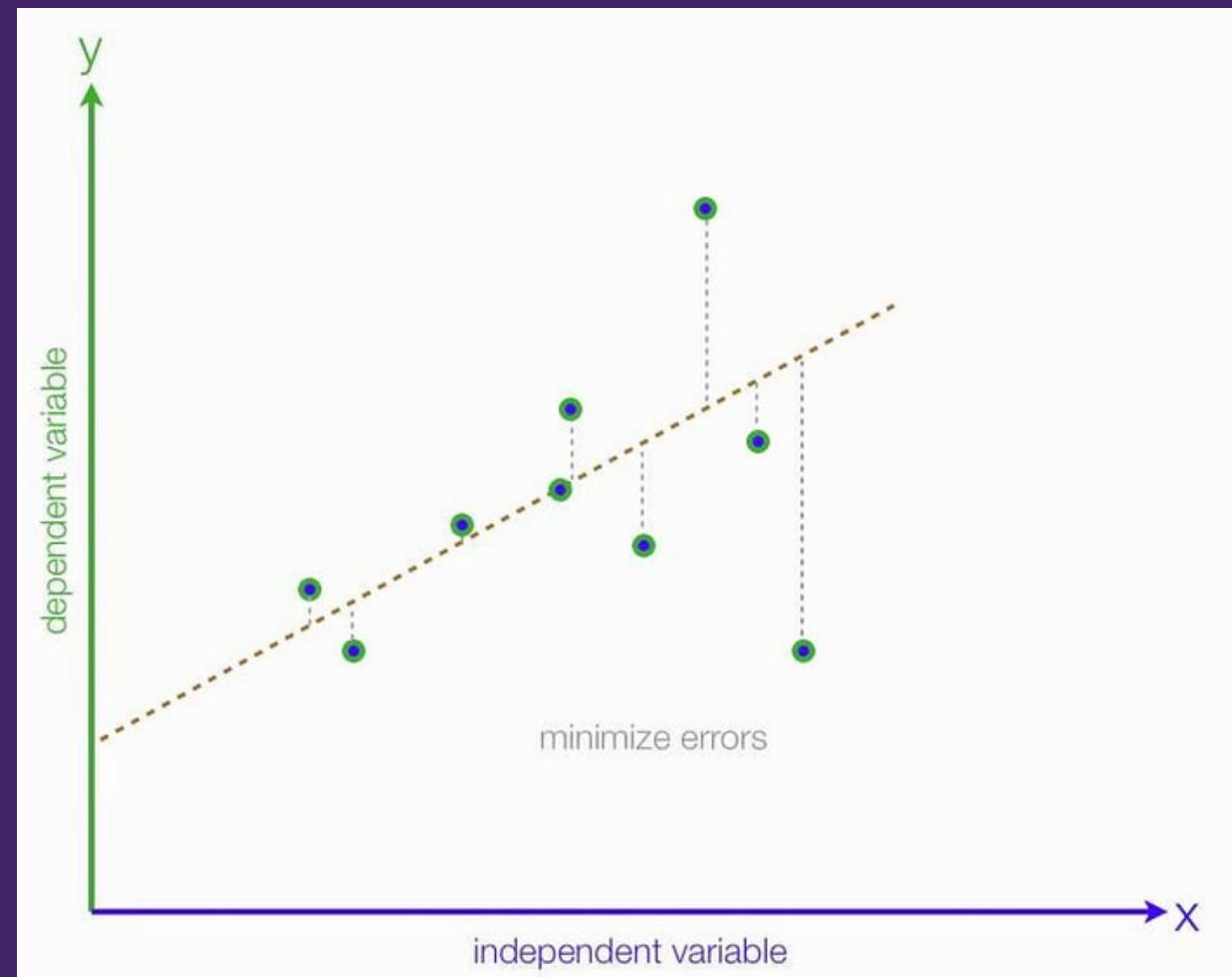
Error

Y_i-nilai aktual
Y_{hat i}→ Nilai prediksi

Evaluasi Regresi

R-Squared (R^2)

$$R^2 = \left(1 - \frac{MSE_{regressionline}}{MSE_{averagedata}} \right)$$





Pembagian Dataset

Training dan Testing



Pada umumnya, dataset dibagi menjadi :

1. Training set (70%)



2. Testing set (30%)



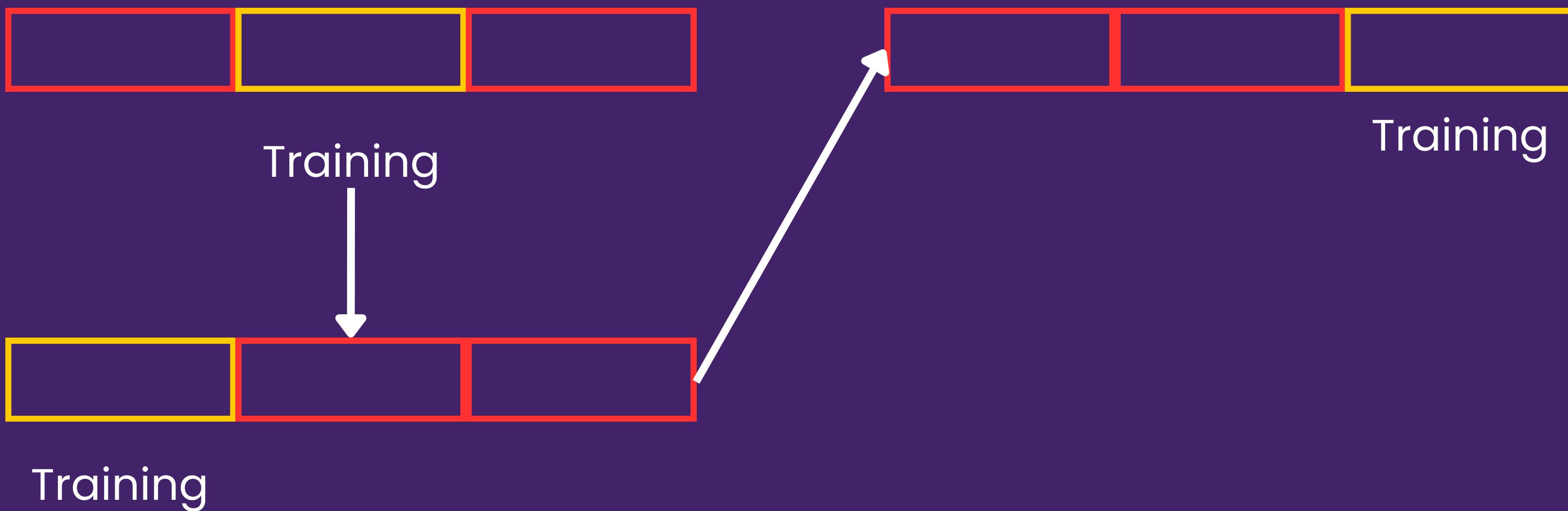


Pembagian Dataset

Cross Validation

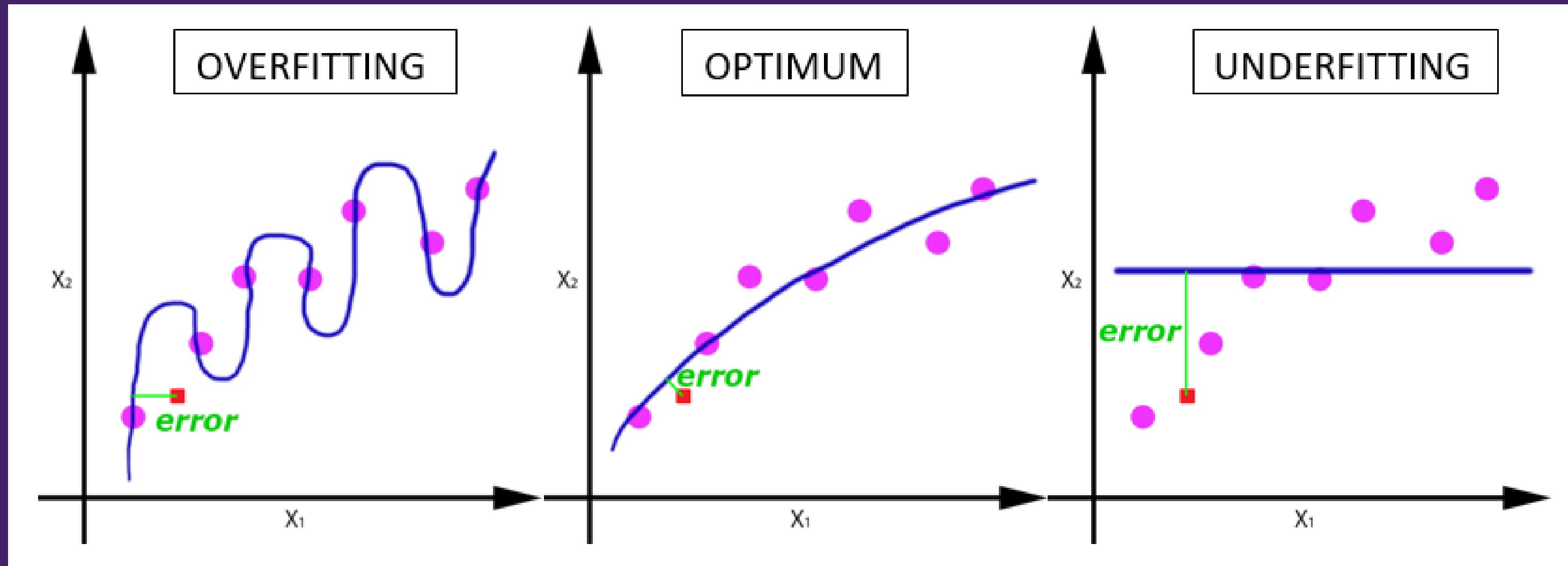


Dataset juga dapat dibagi menjadi beberapa bagian (folds)





Overfitting dan Underfitting



Intro to Machine Learning (Udacity)

All Programs ▶ School Of Artificial Intelligence ▶ Intro to Machine Learning

Free

Intro to Machine Learning

Course

This class will teach you the end-to-end process of investigating data through a machine learning lens, and...

[Read More](#)

Learn For Free ▶

Intermediate Last Updated March 7, 2022

Prerequisites: No experience required



Klasifikasi

Naive Bayes

$$P(Y|X) = \frac{P(X|Y) \bullet P(Y)}{P(X)}$$

Posterior

Prior



Contoh Klasifikasi

Naive Bayes



x1	x2	y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0

Apa label (Y) ketika nilai x= (0,2)?

$$P(Y=0) = 2/5$$

$$P(Y=1) = 3/5$$

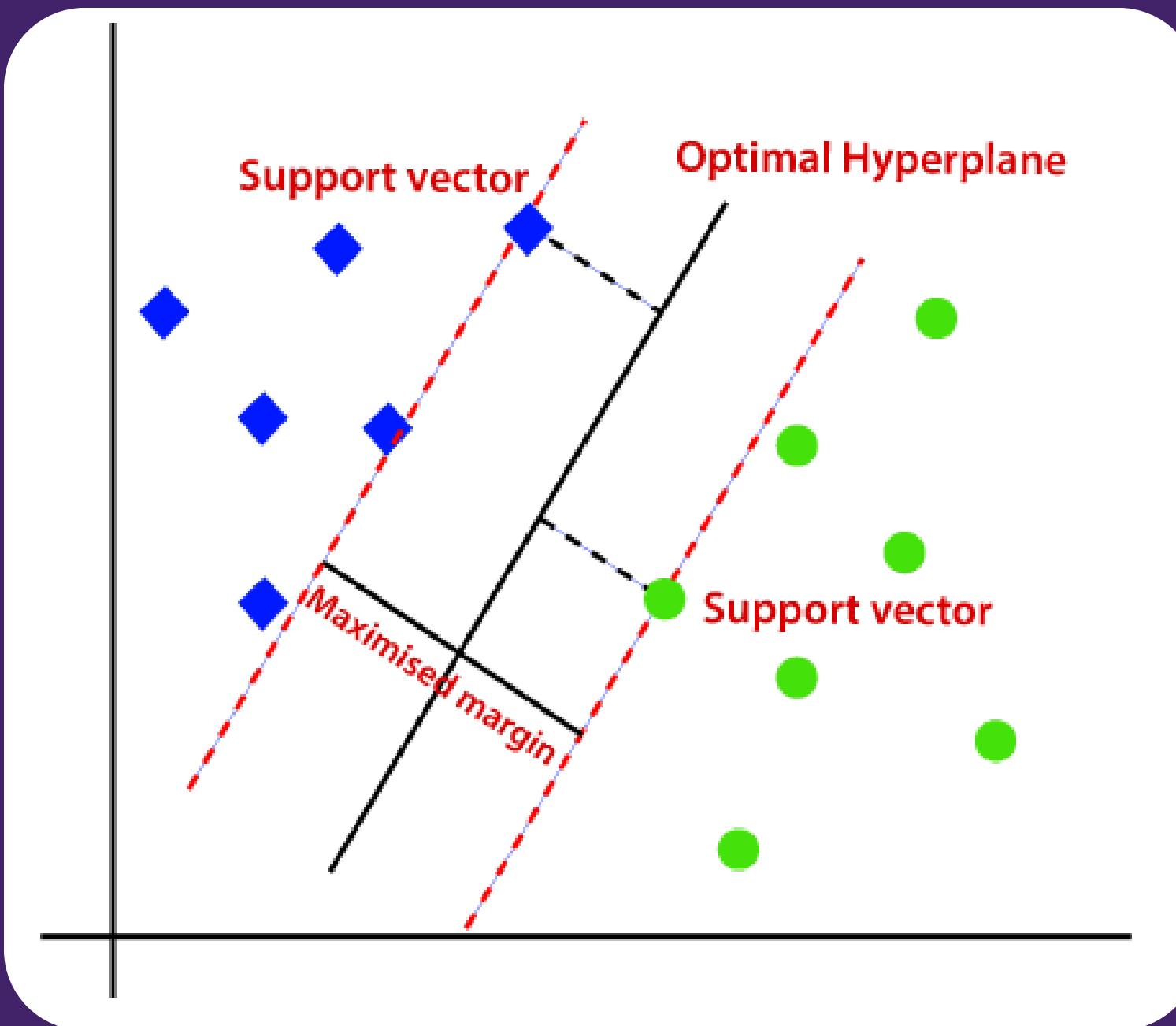
$$\begin{aligned} P(X=0,2|Y=1) &= P(X1=0|Y=1) \cdot P(X2=2|Y=1) \\ &= 2/3 \cdot 1/3 \end{aligned}$$

$$\begin{aligned} P(X=0,2|Y=0) &= P(X1=0|Y=0) \cdot P(X2=2|Y=0) \\ &= 1/2 \cdot 1/2 \end{aligned}$$

$$3/5 \cdot 2/3 \cdot 1/3 > 2/5 \cdot 1/2 \cdot 1/2$$

Jadi, nilai y = 1

Klasifikasi Support Vector Machine (SVM)



Support Vector → 2 Data terdekat yang berasal dari kelompok berbeda
Hyperplane → Garis pembatas antar support vector
(max)Margin → Jarak support vector ke hyper plane



Klasifikasi

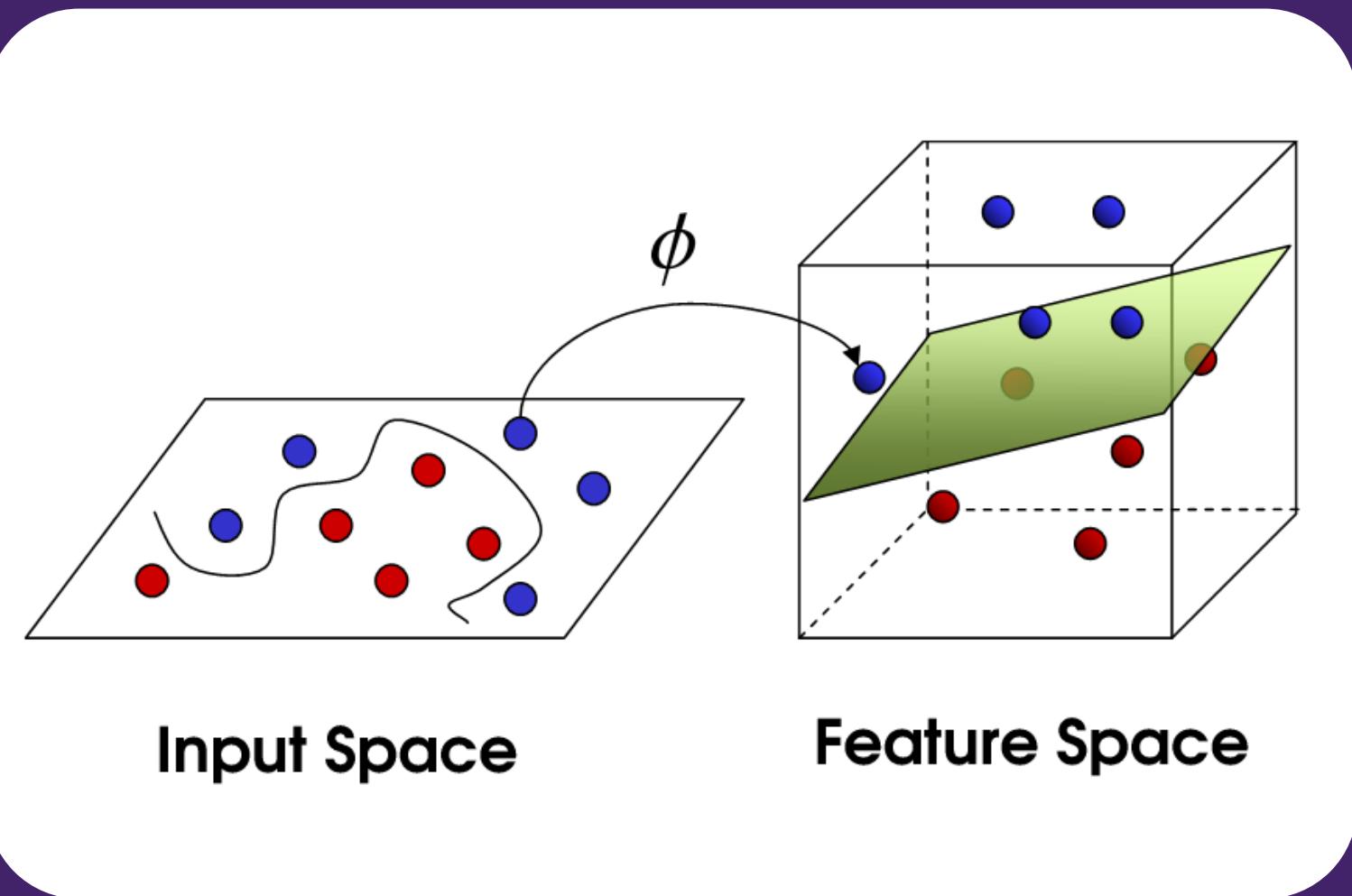
Support Vector Machine (SVM)



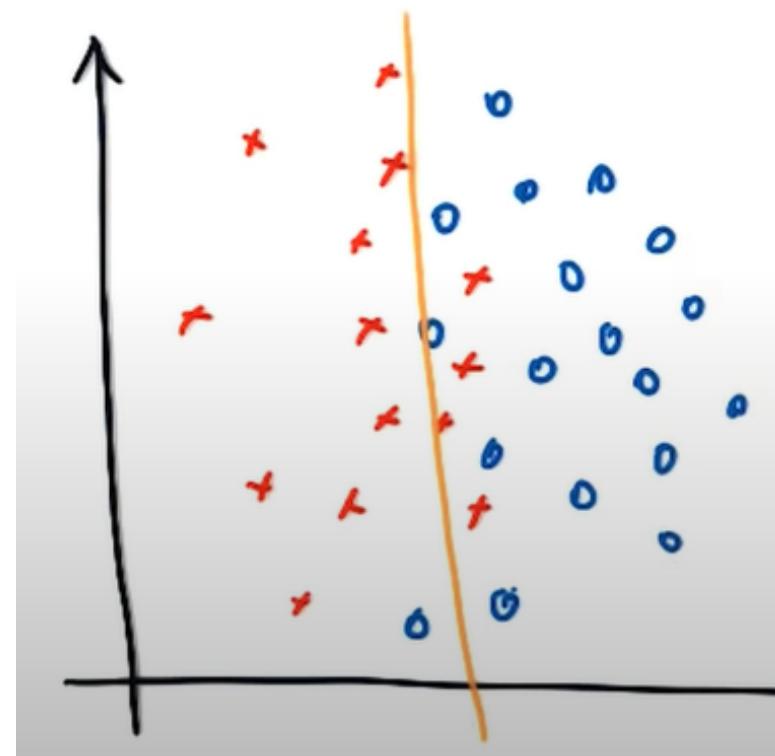
Kernel Trick

mengubah ruang fitur data ke ruang dimensi yang lebih tinggi. Tiga jenis yang sering dipakai:

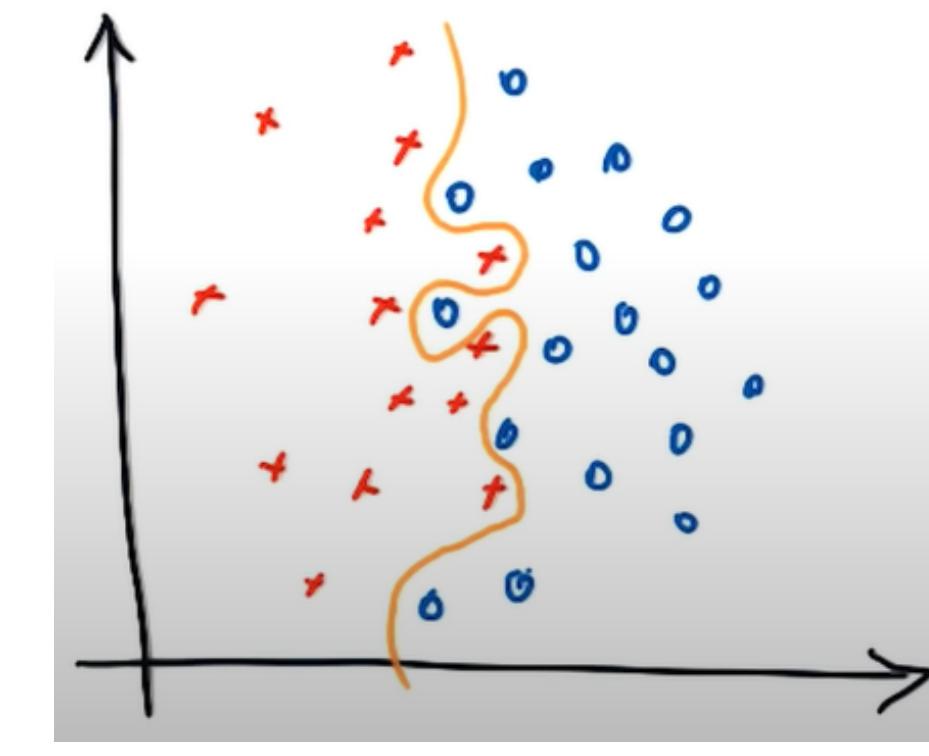
1. Linear
2. Polinomial
3. Radial Basis Function



Klasifikasi Support Vector Machine (SVM)



C Rendah

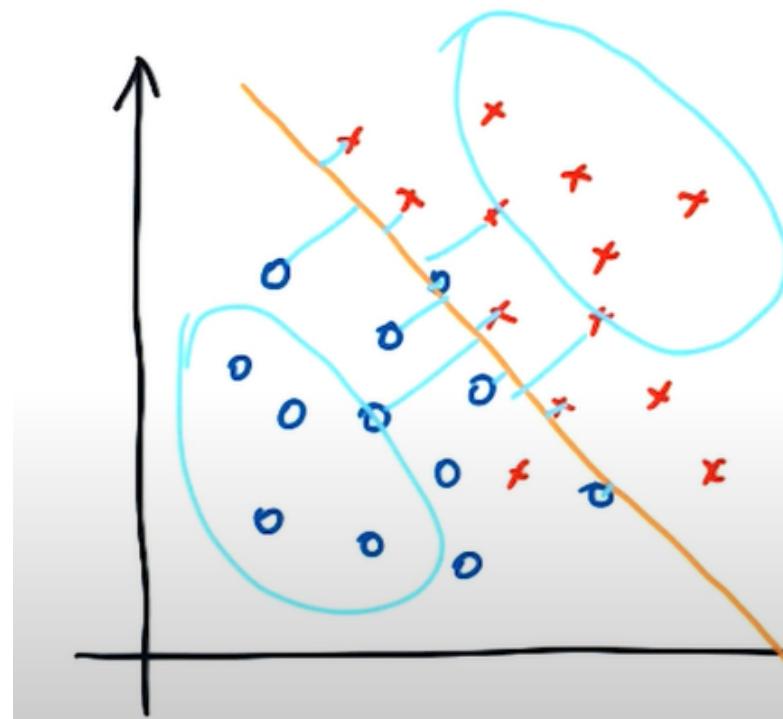


C Tinggi

Nilai C

Parameter yang mengontrol trade-off antara mencapai margin yang lebih besar dan mengurangi kesalahan klasifikasi

Klasifikasi Support Vector Machine (SVM)



Gamma Tinggi



Gamma Rendah

Nilai Gamma

Parameter yang mengontrol pengaruh dari sebuah data.

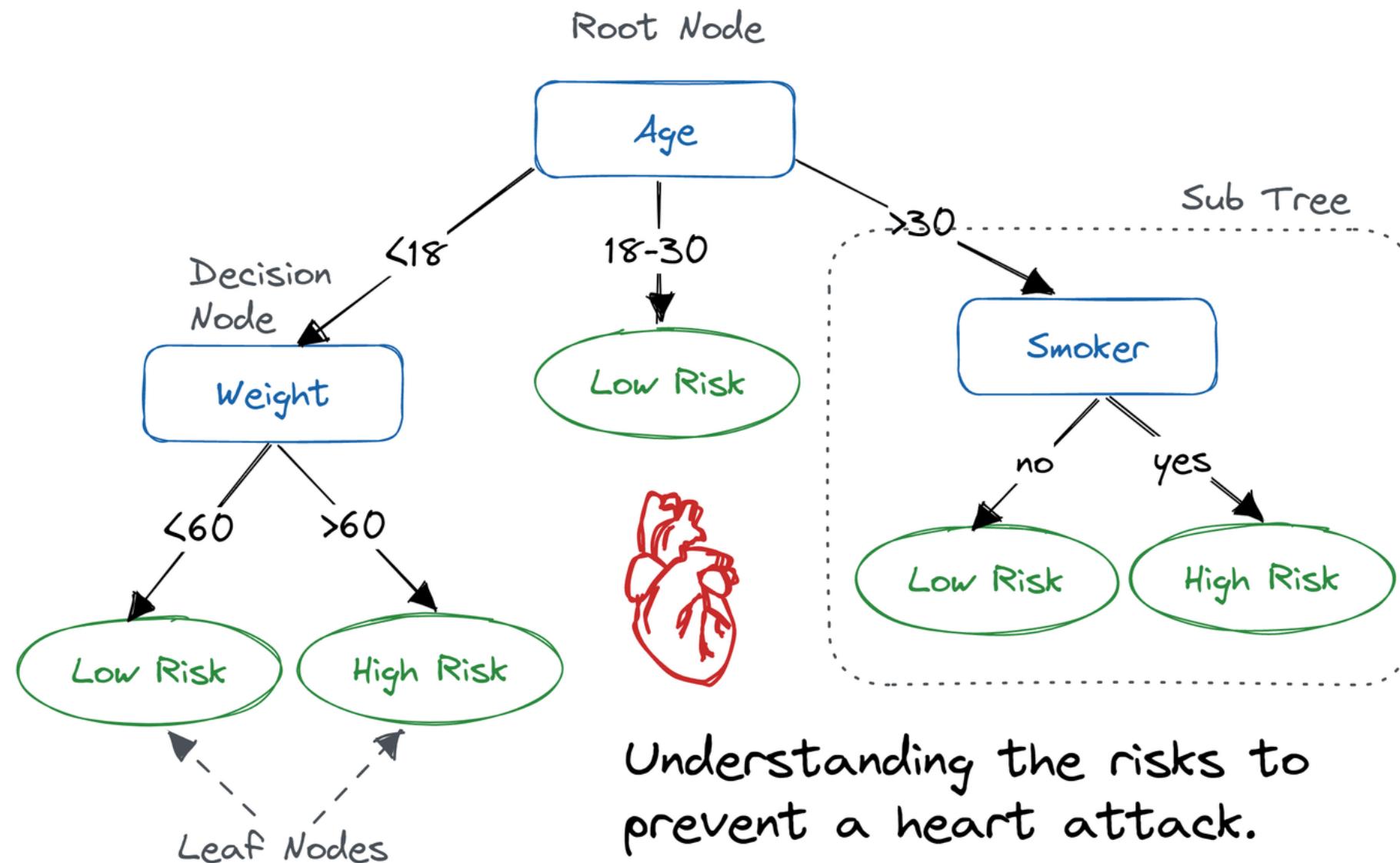
Nilai rendah -> data jauh

Nilai tinggi -> data dekat



Klasifikasi

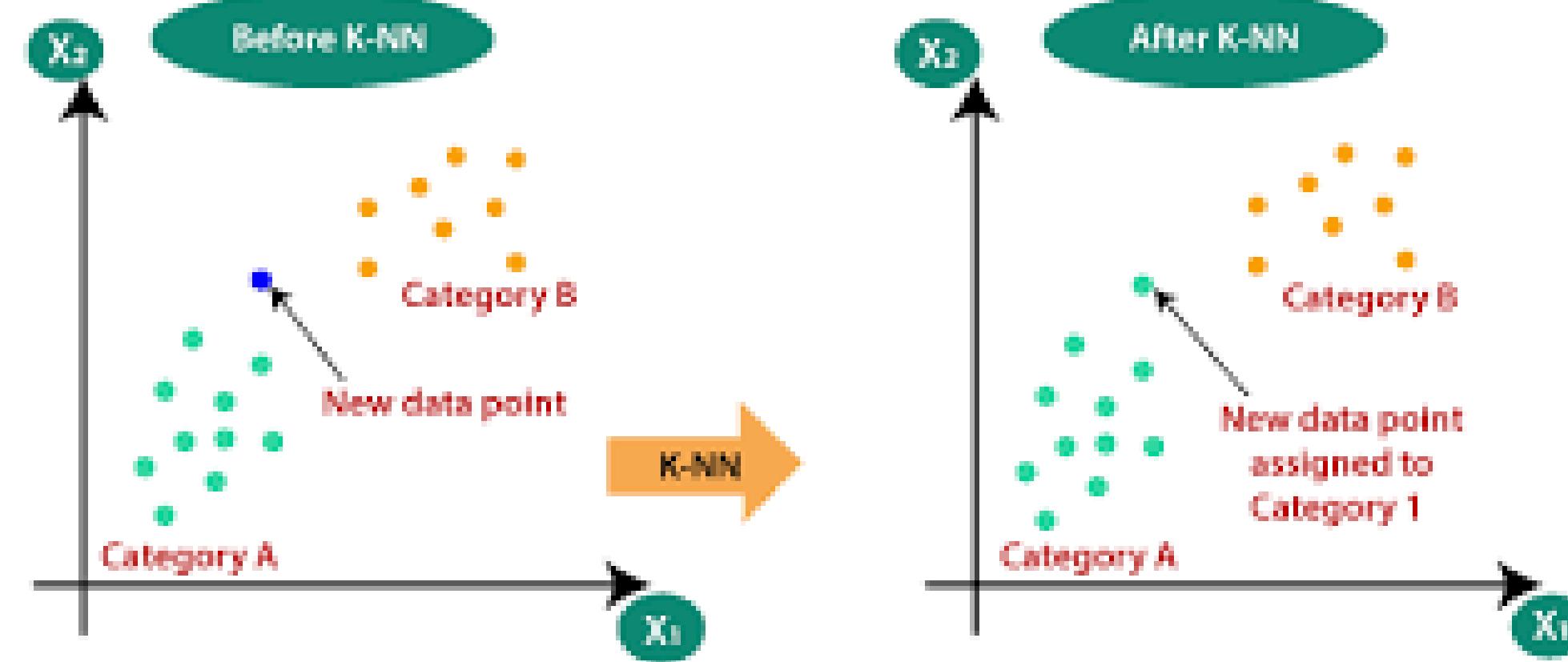
Decision Tree





Klasifikasi

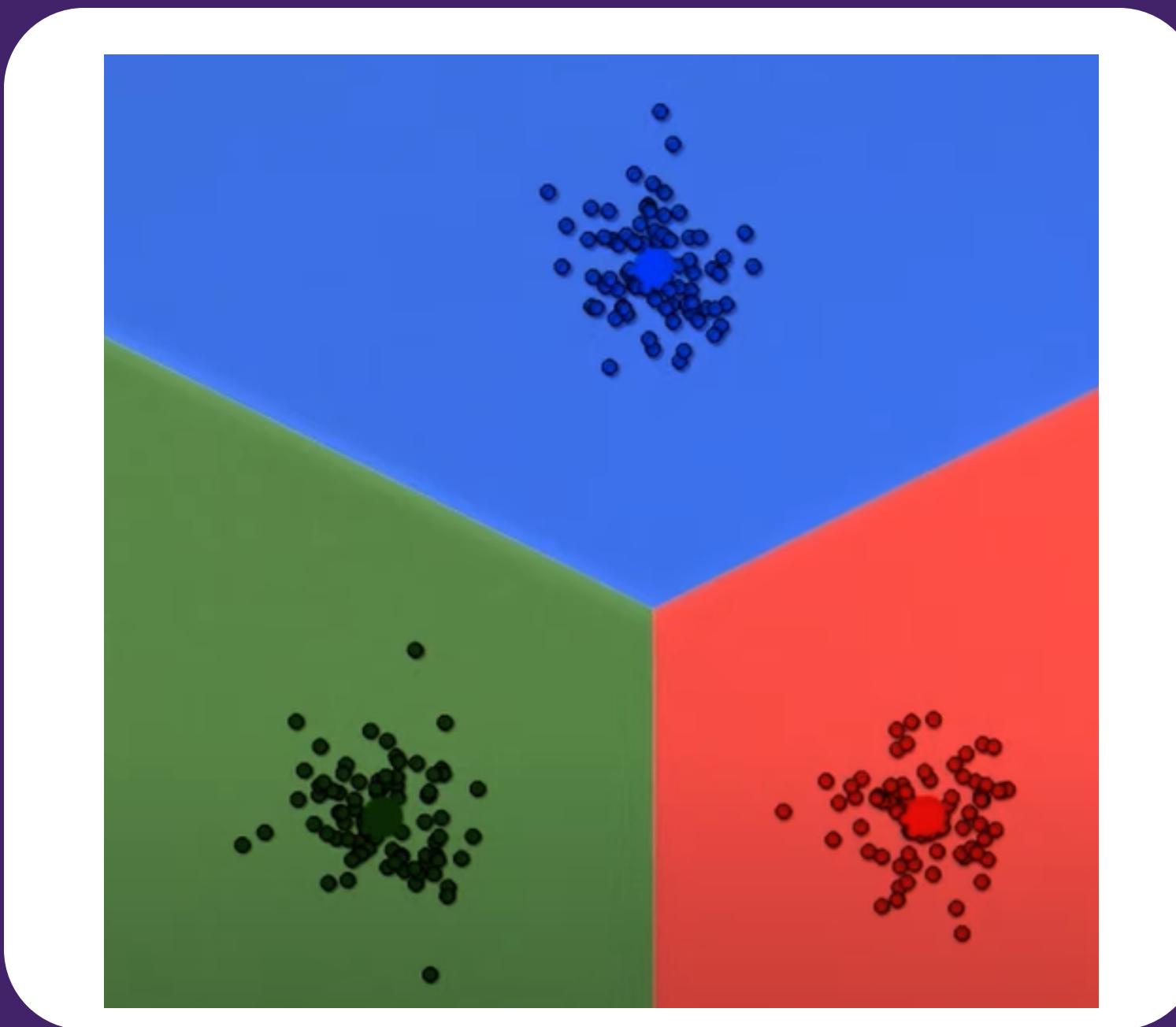
K-Nearest Neighbors (KNN)



K-NN bekerja berdasarkan hasil dari data yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat.

Klastering

K-Means



Pengertian

K- Means adalah algoritma klastering yang mengelompokkan data kedalam K jumlah klaster.

Klastering

K-Means

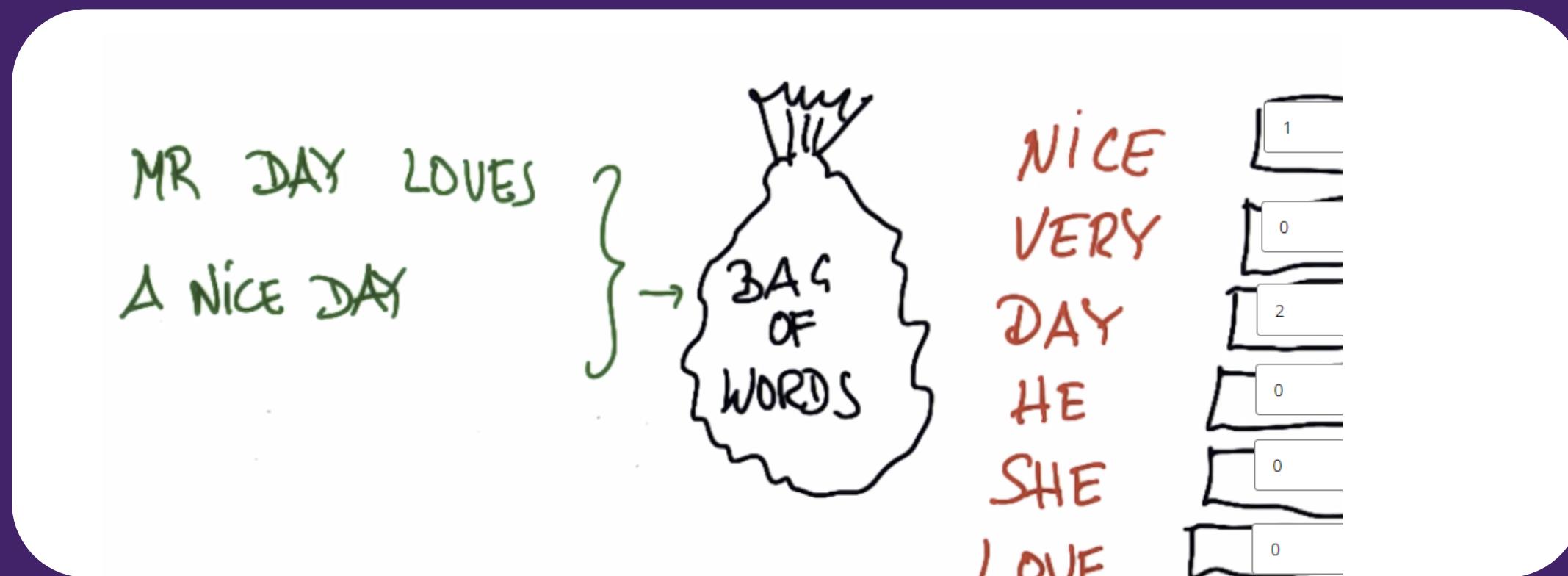


Elbow Method

Teknik yang digunakan untuk menentukan jumlah kluster yang optimal (nilai K) dalam algoritma clustering dengan melihat nilai Sum Squared Eror (SSE) dari setiap jumlah kluster

Text Learning

Bag of Words



Pengertian

Teknik yang digunakan untuk memecah sebuah kalimat menjadi kata dan akan digunakan untuk menghitung frekuensi kemunculan pada kalimat baru

Text Learning

Stopwords

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

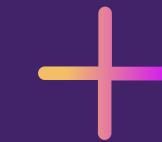
Pengertian

Kata yang tidak memiliki makna khusus atau tidak memiliki dampak yang signifikan

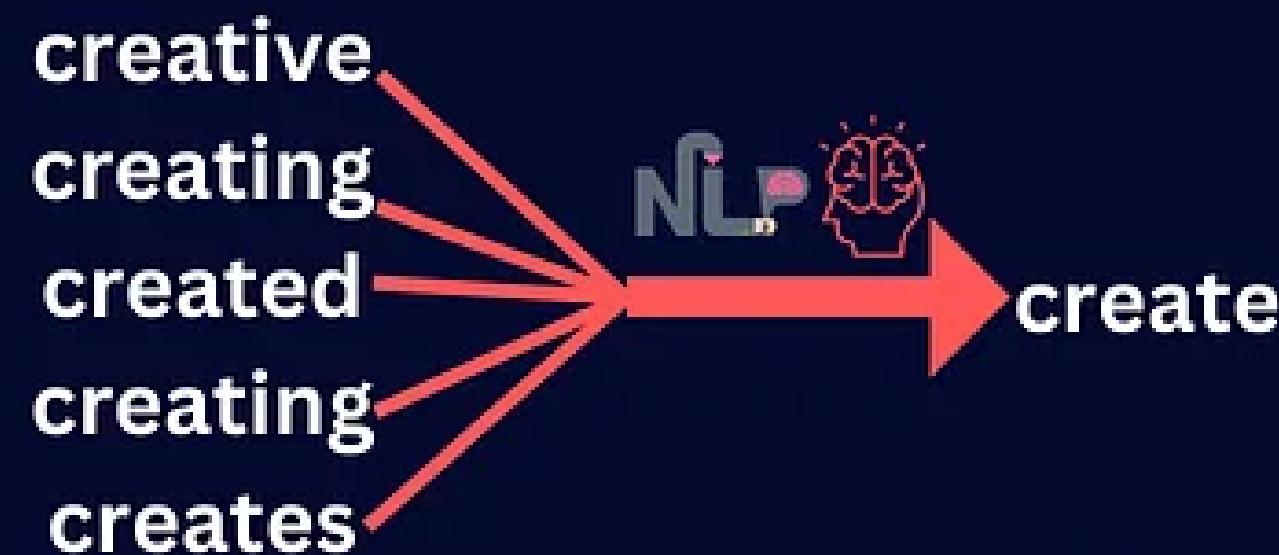


Text Learning

Stemming



Stemming in NLP



Pengertian

Mengubah kata menjadi bentuk dasar dari kata tersebut



Principle Componen Analysis (PCA)



Pengertian

Teknik dalam analisis data yang digunakan untuk mereduksi dimensi data dengan tujuan mengidentifikasi pola yang signifikan dalam data. Dengan mereduksi dimensi data, PCA dapat membantu dalam visualisasi data, mengurangi noise, dan memudahkan analisis data yang lebih efisien.

Evaluation Metrics

		Predicted	
		Positive (+)	Negative (-)
Actual	Positive (+)	True Positive (TP)	False Negative (FN)
	Negative (-)	False Positive (FP)	True Negative (TN)

Pengertian

Tabel atau matriks yang digunakan dalam analisis klasifikasi dan pengujian model untuk mengukur kinerja model dalam memprediksi kelas-kelas target.

Matriks Evaluasi

Akurasi

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Presisi

$$\frac{TP}{TP + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

F1-Score

$$\frac{2 \bullet Presisi \bullet Recall}{Presisi + Recall}$$



**TERIMA
KASIH**

