MODEL

Made Satria Wibawa, M.Eng.

2020

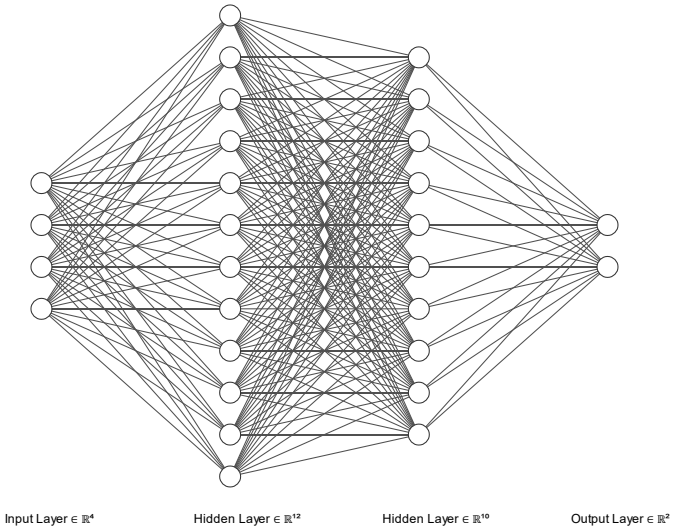
Outline

- *Model*
- *Tipe Algoritma Data Mining*
- *Seleksi dan Evaluasi Model*

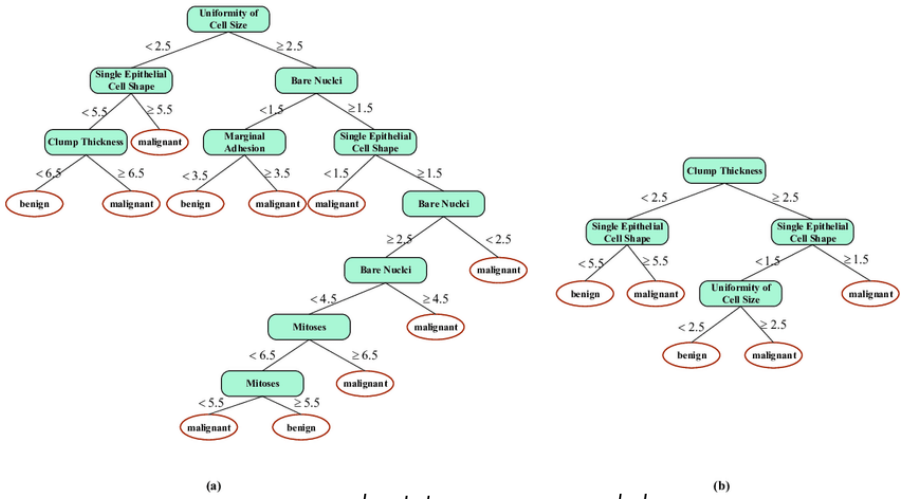
MODEL

Model

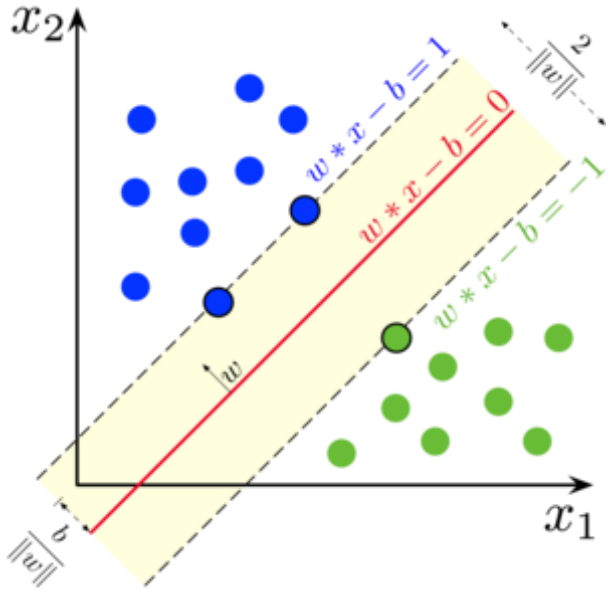
Model adalah kumpulan dari pola, statistik dari data yang bisa diaplikasikan pada data baru untuk melakukan prediksi atau inferensi.



neural network model

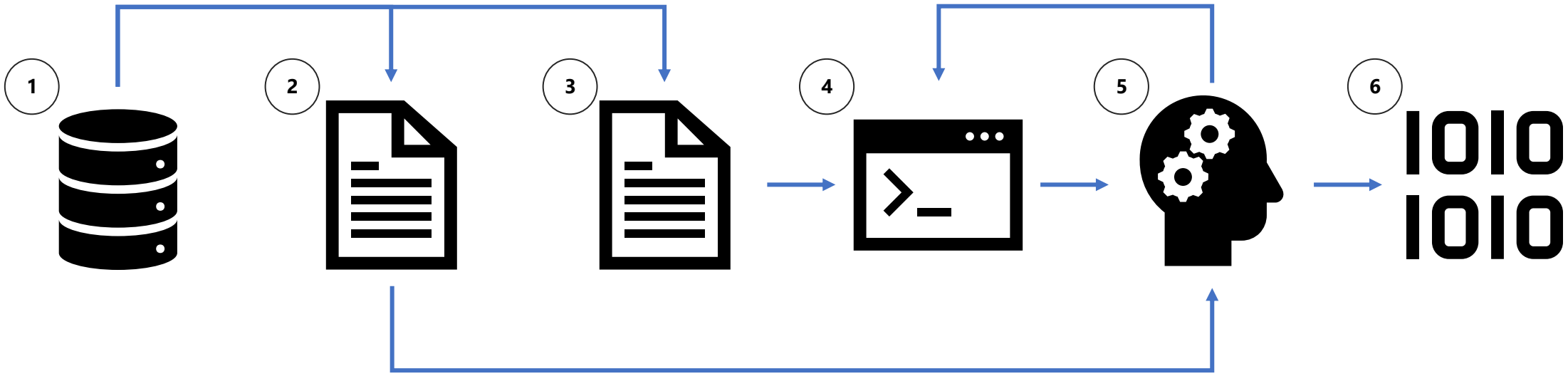


decision tree model



support vector machine model

Pembentukan Model Pengetahuan



1. Dataset
2. Testing data
3. Training data
4. Algoritma data mining
5. Model
6. Hasil

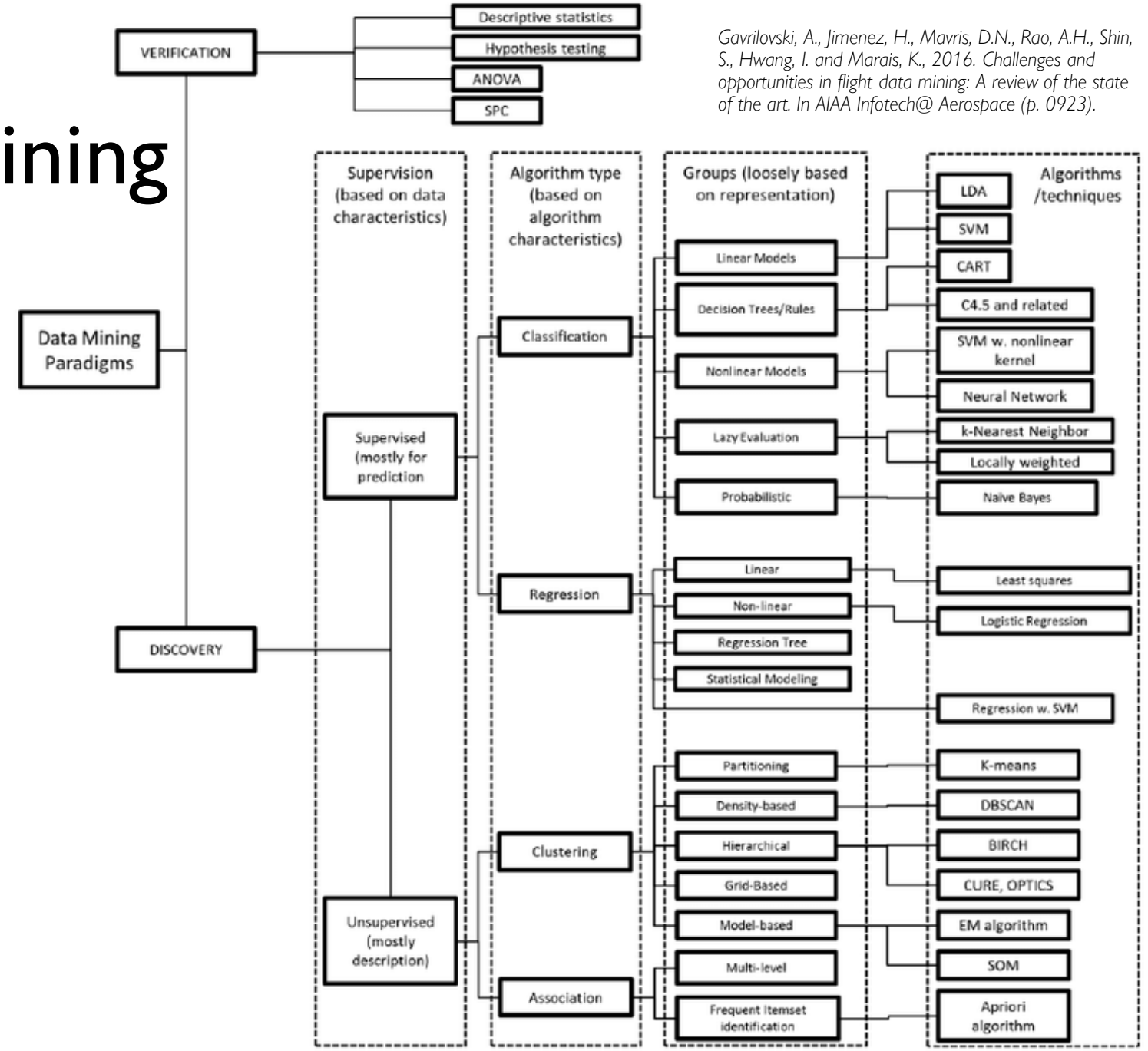
ALGORITMA DATA MINING

Algoritma Data Mining

Yang akan kita pelajari adalah *supervised* dan *unsupervised learning*.

Supervised learning atau pembelajaran terbimbing merupakan proses ekstraksi knowledge pada data yang memiliki label.

Unsupervised learning atau pembelajaran terbimbing merupakan proses ekstraksi knowledge pada data yang tidak memiliki label.



Gavrilovski, A., Jimenez, H., Mavris, D.N., Rao, A.H., Shin, S., Hwang, I. and Marais, K., 2016. Challenges and opportunities in flight data mining: A review of the state of the art. In AIAA Infotech@ Aerospace (p. 0923).

Supervised Learning

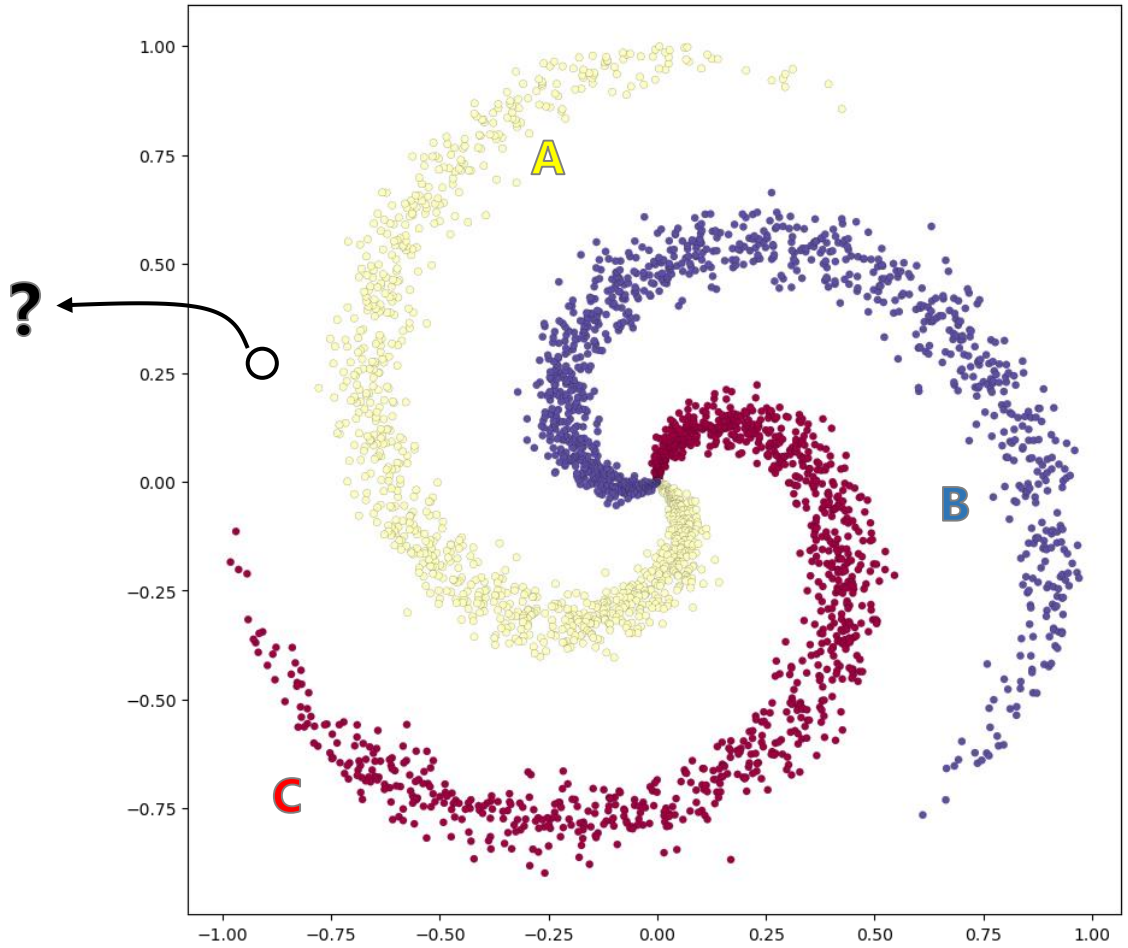
Klasifikasi

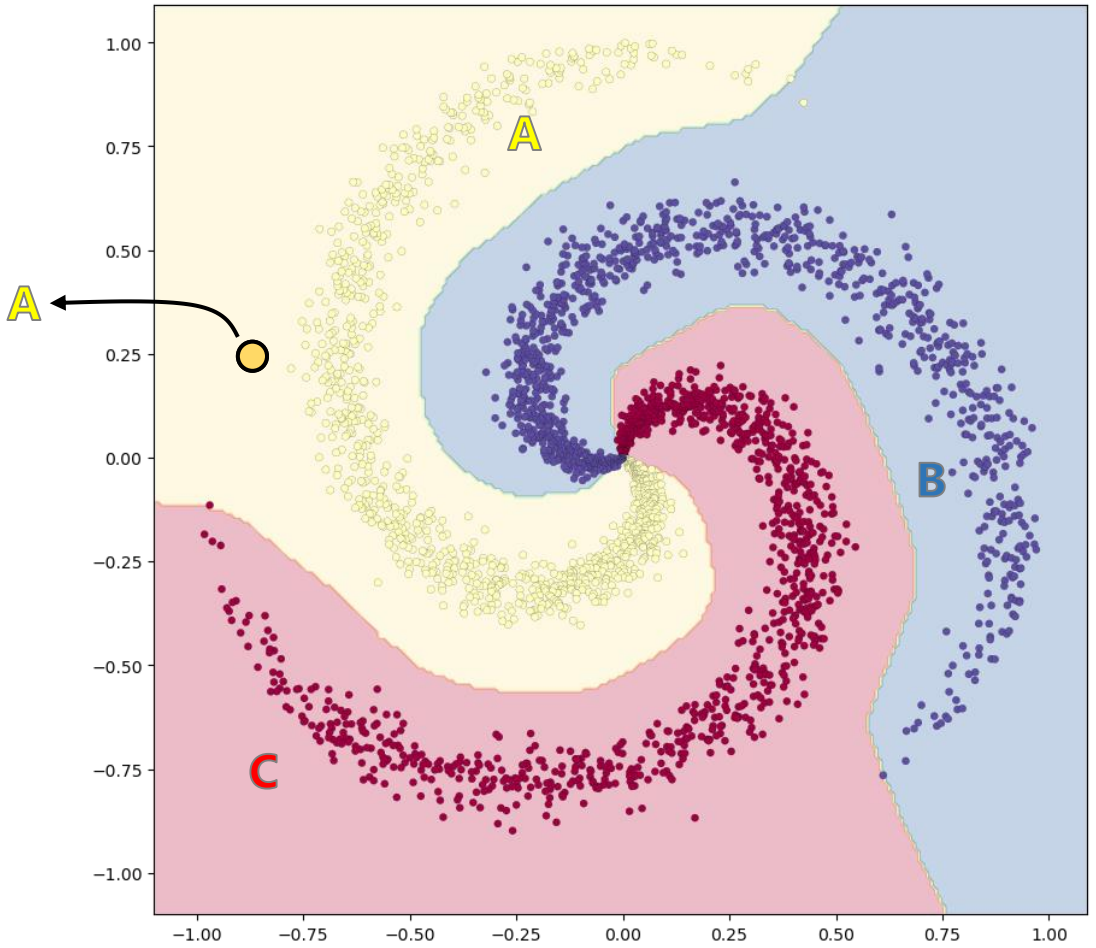
- Label target bertipe diskrit
- Algoritma:
  - Linear model
  - Decision trees/rule
  - Non-linear model
  - Lazy learning
  - Probabilistik

Regresi

- Label target bertipe kontinyu
- Algoritma
  - Linear model
  - Non-linear model
  - Regression Tree
  - Statistical modelling

Supervised Learning: Klasifikasi





Supervised Learning: Klasifikasi

age	income	student	credit rating	buy
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

age	income	student	credit rating	buy
>40	high	yes	excellent	…

Prediksi Pembelian Komputer

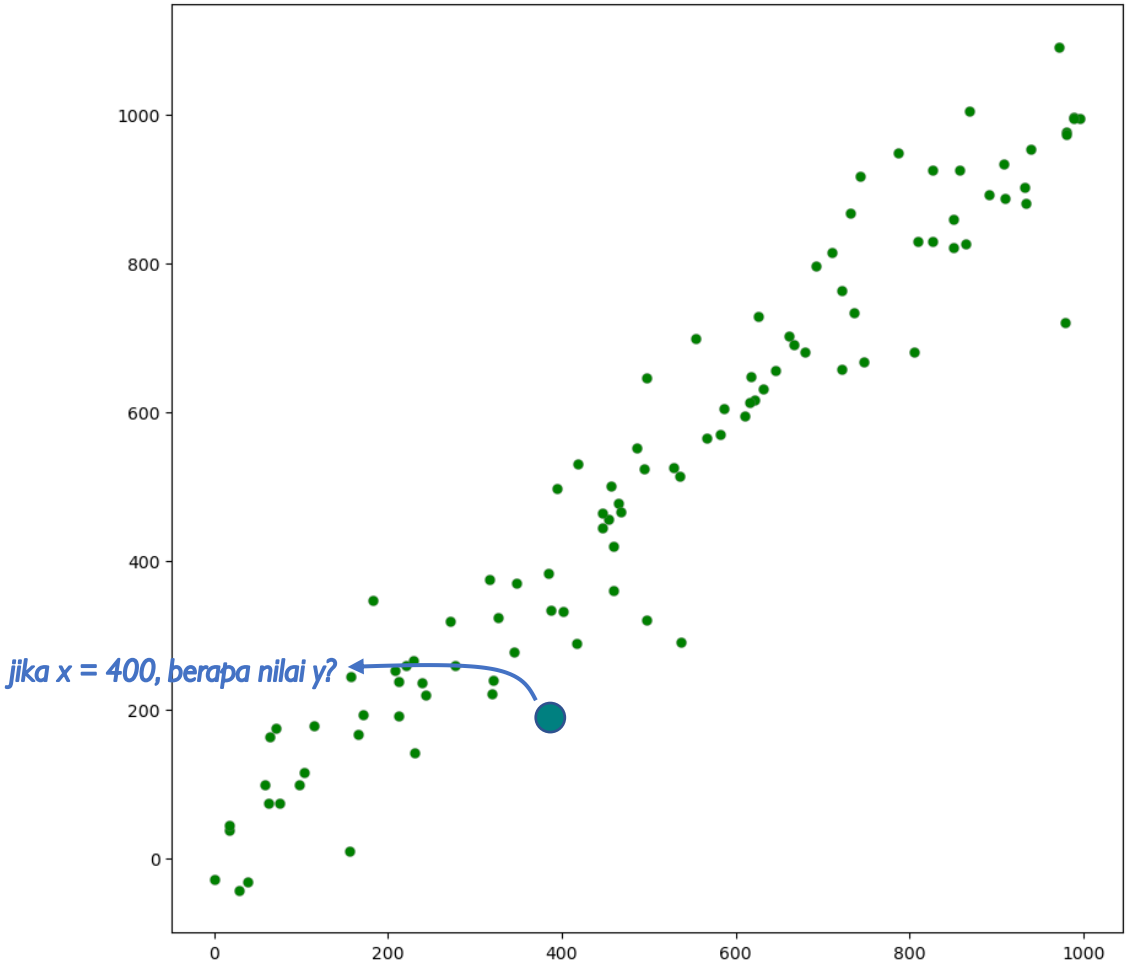
Pada tabel berwarna biru di samping, terdapat dataset dengan label buy dan nilai yes atau no.

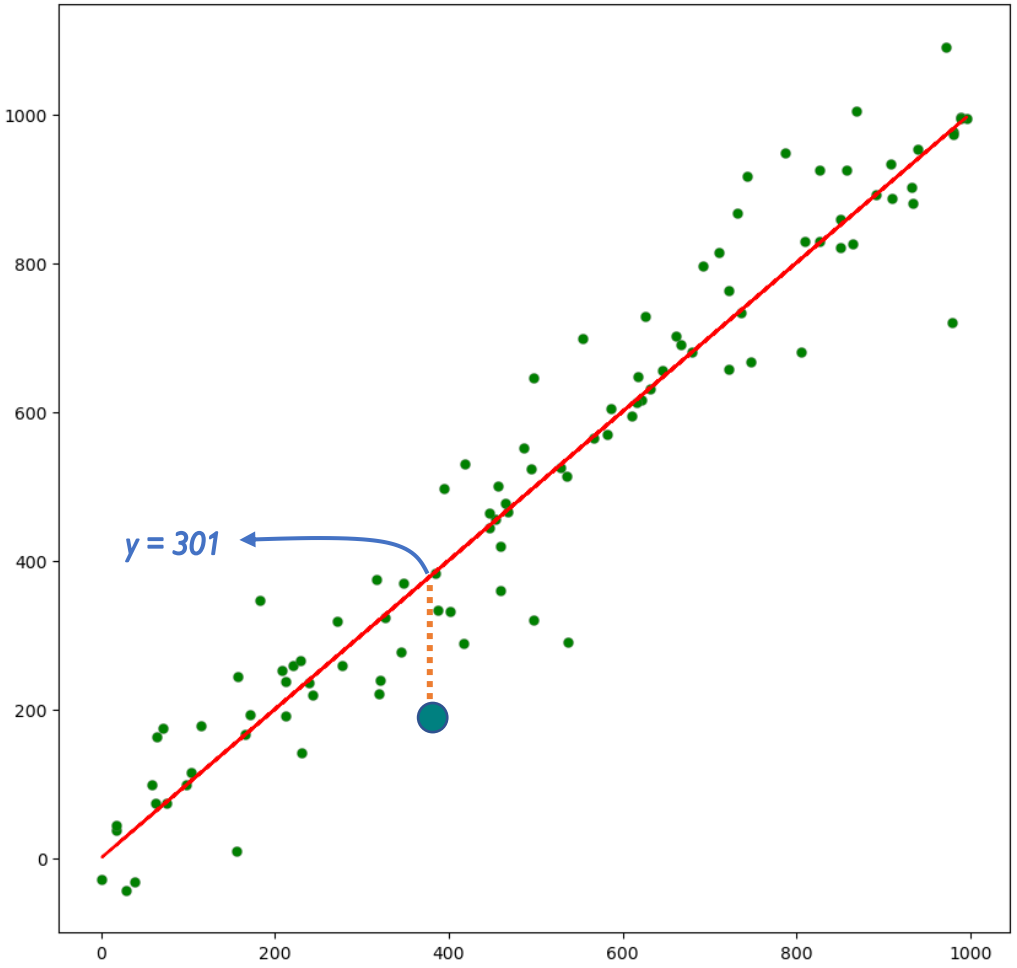
Klasifikasi akan memprediksikan kelas/label dari data yang baru berdasarkan nilai dari atribut age, income, student dan credit rating-nya.

Metode Klasifikasi

- Linear model
  - Linear Discriminant Analysis
  - Support Vector Machine
- Decision trees/rule
  - ID3
  - C4.5
- Non-linear model
  - Artificial Neural Network
  - Support Vector Machine with non-linear kernel
- Lazy learning
  - K-nearest neighbor
  - Locally weighted
- Probabilistik
  - Naive bayes

Supervised Learning: Regresi





Supervised Learning: Regresi

cylinders	displacement	horsepower	weight	mpg
8	340	160	3609	14
8	400	150	3761	15
8	455	225	3086	14
4	113	95	2372	24
6	198	95	2833	22
6	199	97	2774	18
6	200	85	2587	21
4	97	88	2130	27
4	97	46	1835	26
4	110	87	2672	25
4	107	90	2430	24
4	104	95	2375	25
4	121	113	2234	26
6	199	90	2648	21

cylinders	displacement	horsepower	weight	mpg
6	162	80	1800	…

Prediksi Efisiensi Kendaraan

Pada kasus regresi, nilai yang diprediksikan adalah nilai numerik.

Contohnya, berdasarkan atribut cylinders, displacement, horsepower dan weight kita dapat memprediksikan efisiensi konsumsi bahan bakar kendaraan (dalam satuan mile per gallon)

Metode Regresi

- Linear model
  - Least square
- Non-linear model
  - Logistic regression
  - Neural network
  - Support vector regressor
- Regression Tree
- Statistical modelling

Unsupervised Learning

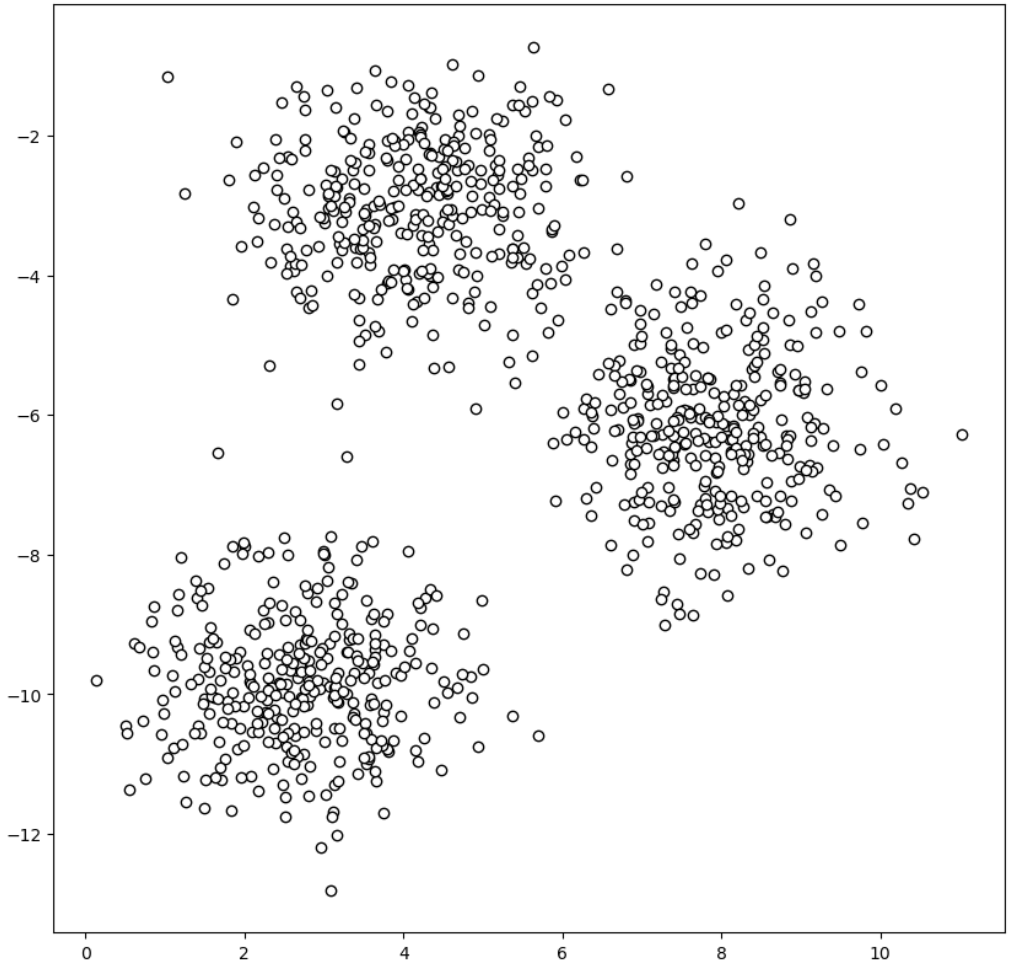
Clustering

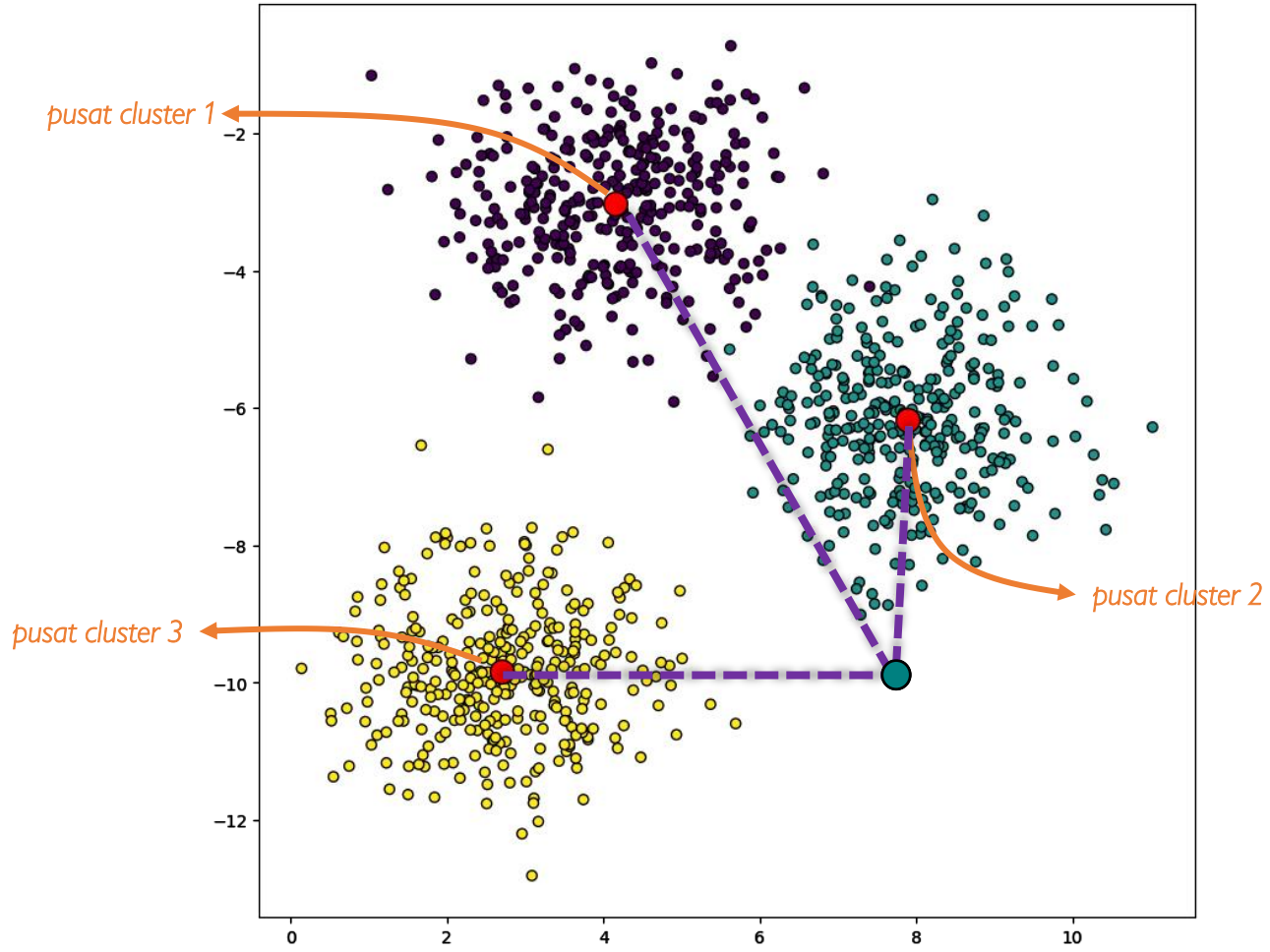
- Pengelompokan data menjadi kelompok, dimana satu kelompok memiliki karakteristik yang sama
- Algoritma:
  - Partitioning
  - Density-based
  - Hierarchical
  - Grid-based
  - Model-based

Asosiasi

- Metode ini mencari hubungan antara data
- Algoritma
  - Multi-level
  - Frequent itemset identification

Unsupervised Learning: Clustering





Metode Clustering

- Partitioning

Metode ini membentuk data ke dalam partisi, dimana setiap partisi merepresentasikan cluster.

  - K-means
  - K-medoid
  - Fuzzy c-means

- Density-based

Metode ini membentuk cluster dengan mempertimbangkan kerapatan (jumlah data) dalam area terdekat.

  - DBSCAN

- Hierarchical

Metode ini membentuk cluster dalam bentuk dekomposisi hirarki

  - BIRCH
  - CURE, OPTICS

- Grid-based

Metode ini membentuk cluster ke dalam bentuk struktur jaringan

- Model-based

Metode cluster ini memperkenalkan probability cluster, tidak seperti k-means yang membentuk hard-cluster

  - SOM
  - EM algorithm

MODEL SELECTION & EVALUATION

Evaluasi Model

- Accuracy
  Kemampuan model untuk memprediksi data yang benar
- Speed
  Computational cost (memori, waktu, biaya) yang diperlukan untuk membangun dan menggunakan model
- Robustness
  Kemampuan model untuk melakukan prediksi yang benar dari real data (noise, missing value)
- Scalability
  Kemampuan model untuk dibuat dan dijalankan secara efisien pada data dengan jumlah yang sangat besar
- Interpretability
  Karakteristik model untuk dapat dipahami oleh manusia
- Domain Dependet Indicators
  Indikator khusus dimana model tersebut diterapkan

Training & Testing

Untuk dapat mengevaluasi dan memilih model pada data, kita harus membagi data menjadi dua subset.

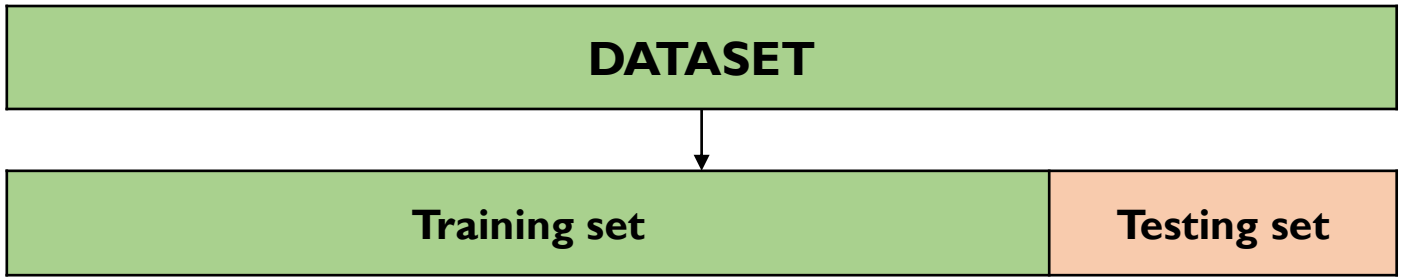
Subset pertama adalah training set, yang berfungsi untuk membentuk model.

Subset kedua adalah testing set, berfungsi untuk menguji model yang terbentuk dari training set.

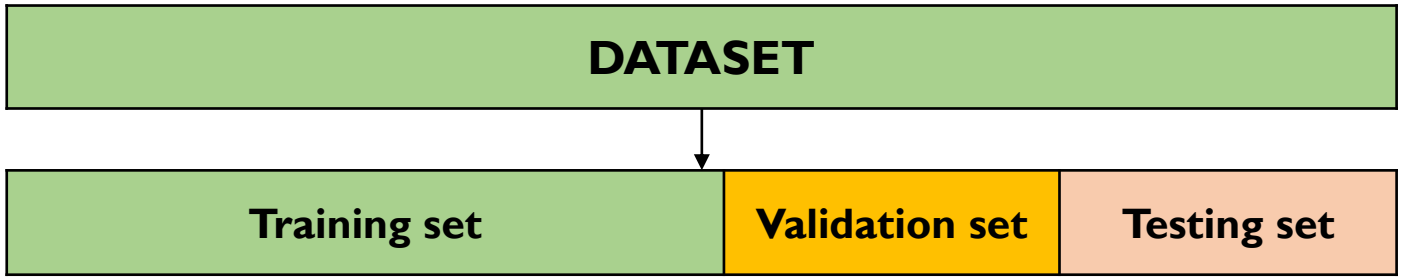
Untuk membagi data menjadi dua subset tersebut, terdapat beberapa metode yang dapat digunakan, yaitu :

- Hold out
- K-folds cross validation
- Leave-one-out

Hold Out



- Pada skema ini data dibagi menjadi dua.
- Training set > testing set
- Training set untuk membentuk model
- Testing set untuk menguji model

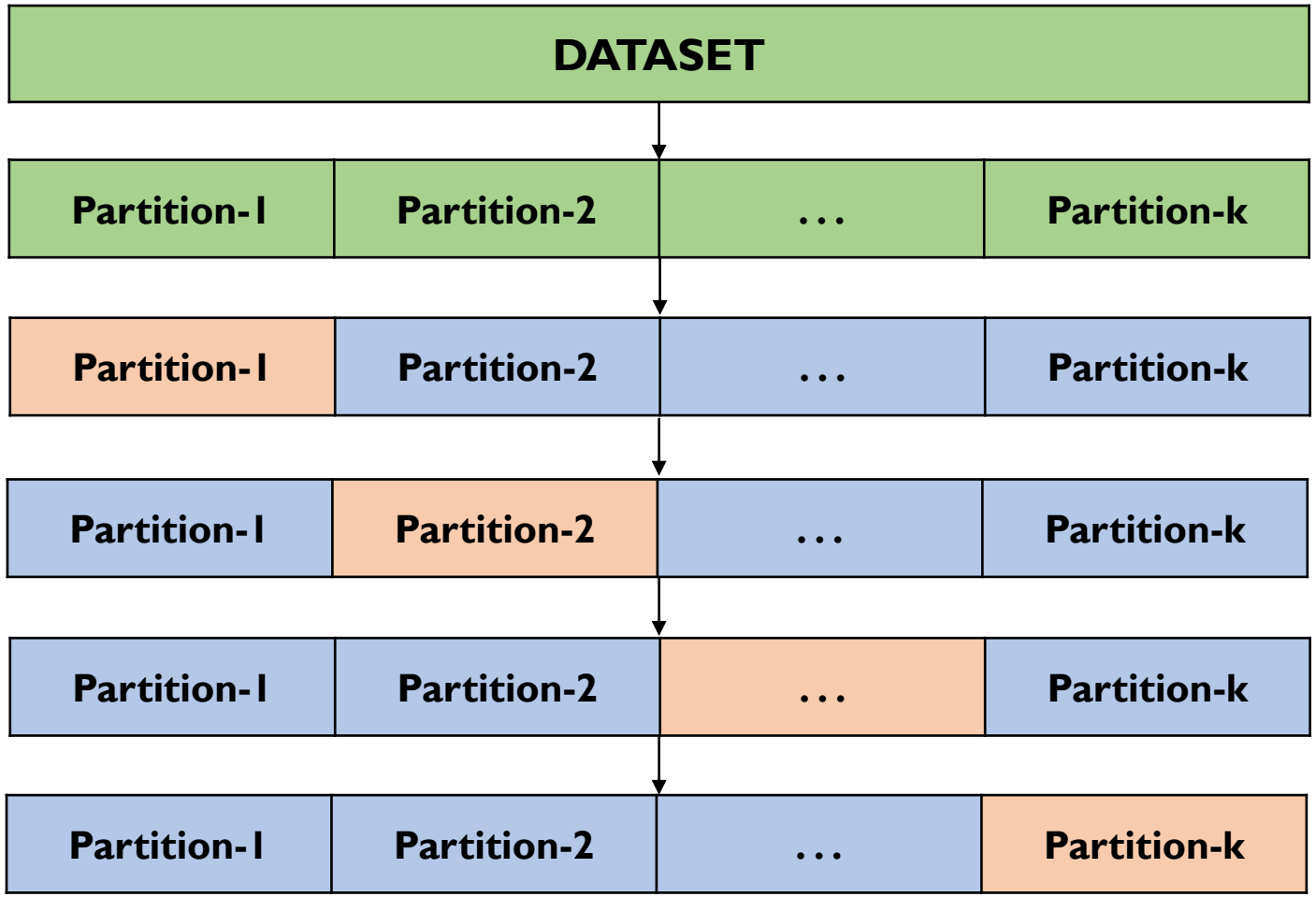


- Pada skema ini data dibagi menjadi tiga.
- Training set > testing set and validation set
- Testing set == Validation set
- Training set untuk membentuk model
- Validation set untuk tuning model
- Testing set untuk menguji model

- umumnya digunakan pada dataset dengan ukuran yang cukup besar

- computation cost paling minim

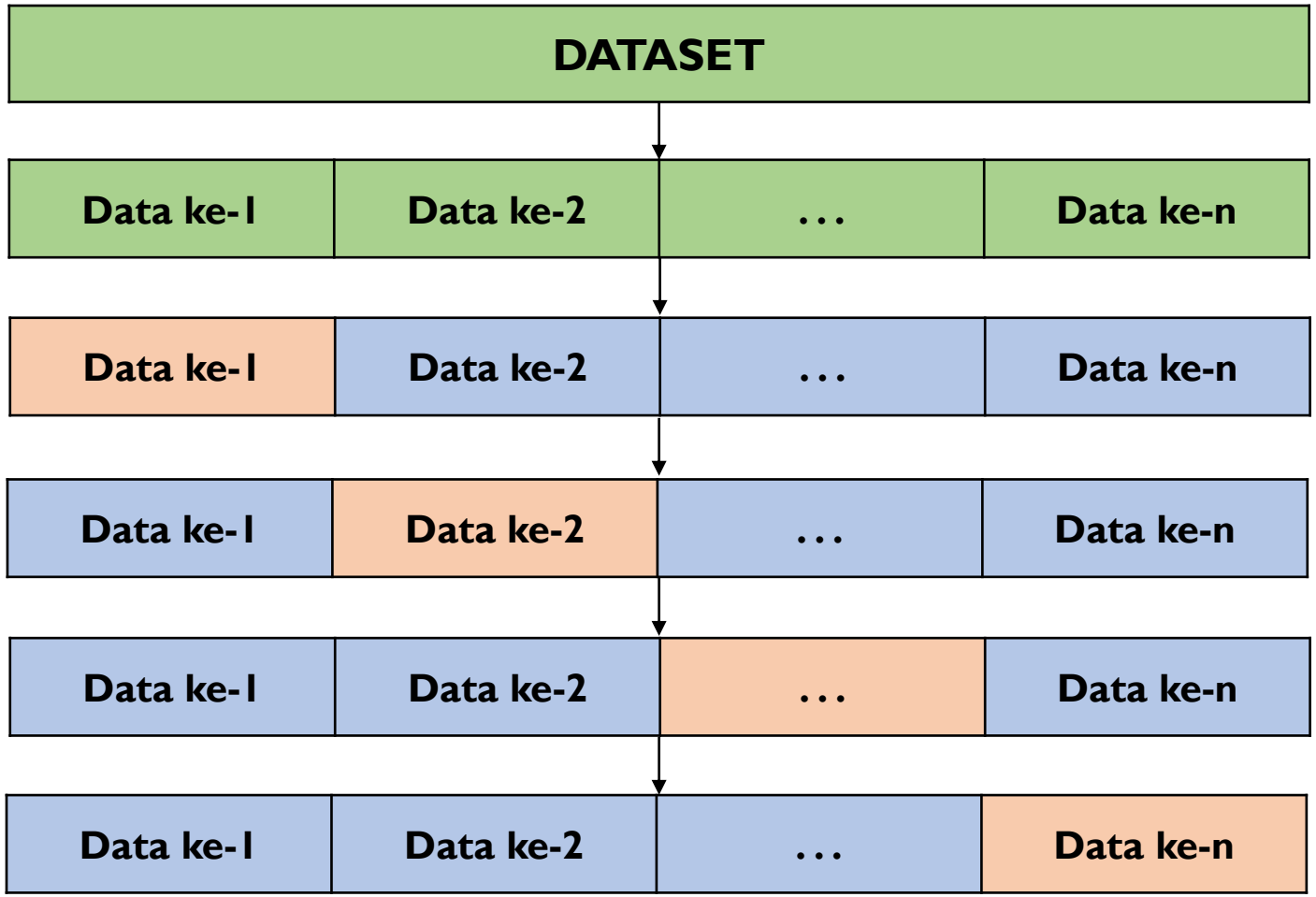
K-folds Cross Validation



- Dataset dibagi menjadi k-partisi/bagian
- Jumlah data pada setiap partisi diusahakan sama/mirip
- Distribusi antar kelas pada setiap partisi diusahakan sama/mirip
- Training terjadi sebanyak k-kali
- 1 partisi digunakan sebagai testing, partisi data k-1 digunakan sebagai training

- *computation cost lebih tinggi daripada holdout*
- *cocok digunakan untuk ukuran data yang menengah*

Leave One Out



- Dataset dibagi menjadi n bagian (n adalah jumlah data)
- Training terjadi sebanyak n-kali
- 1 data digunakan sebagai testing, partisi data n-1 digunakan sebagai training

- *computation cost paling tinggi dibandingkan ketiga metode*
- *cocok digunakan untuk ukuran data yang kecil*

pertanyaan/troubleshooting silahkan buat di channel Diskusi Teams