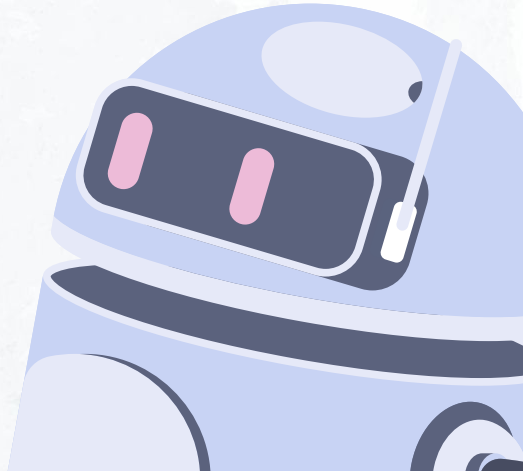
Learning Progress Review

“Introduction to Machine Learning”



Group 5

TIM SERU

Table of contents

01 →

Google Looker Studio

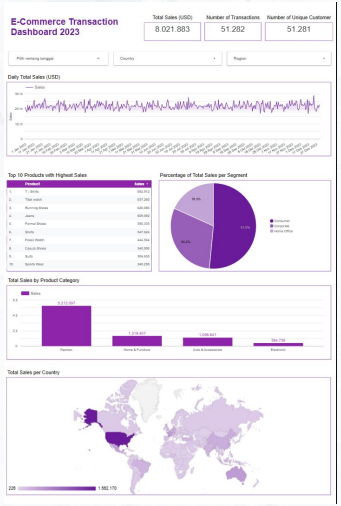
(AI)

Google Looker Studio

Google looker studio adalah alat visualisasi data gratis dari google yang digunakan untuk membuat laporan dan dashboard interaktif dengan mudah dan dapat dipahami. Hal mendasar dalam membuat visualisasi data adalah dengan menggunakan business question. Dengan adanya business question kita mampu membuat visualisasi data menjadi lebih terarah dan menggunakan diagram yang tepat untuk merepresentasikan sebuah data.

Business questions

1. Berapa total nilai sales selama 2023? – Score card (krn single value)
2. Berapa banyaknya transaksi di tahun 2023? – score card
3. Berapa banyaknya customer unik yg bertransaksi di th 2023? – score card
4. Bagaimana tren total sales harian selama 2023? – line chart (tren)
   a. Sb. x? Order date
   b. sb. Y? Sales
5. Berikan daftar 10 produk dengan total sales tertinggi. – tabel
6. Bagaimana proporsi/persentase total sales per segment? – pie chart
7. Berikan urutan total sales per kategori produk. – bar chart vertikal
8. Total sales per country – geo chart/geo map



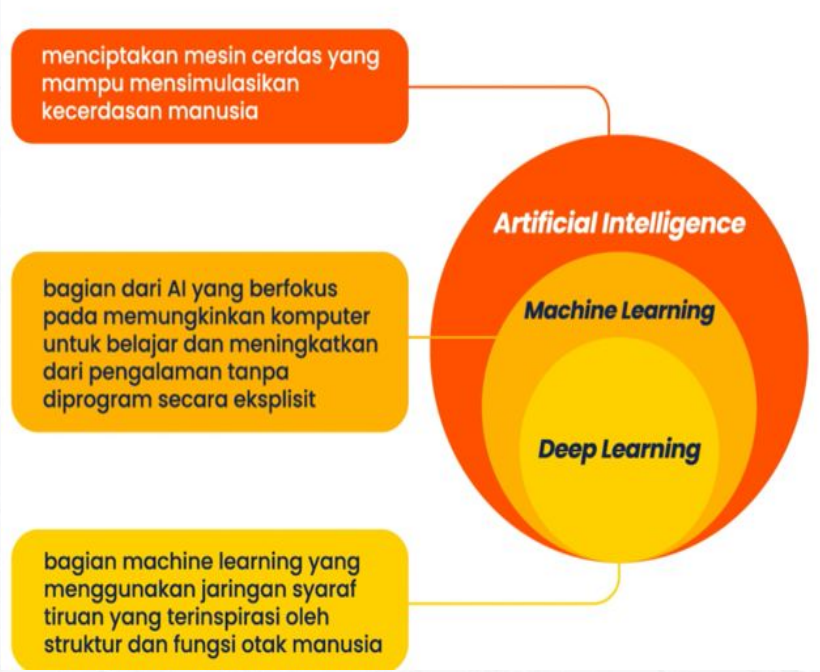
02 →

Introduction to Machine Learning

(AI)

Definisi Machine Learning

Machine learning adalah **sub-bidang kecerdasan buatan** yang **fokus** pada pengembagan **algoritma** dan **model** yang mampu belajar dan **membuat prediksi** atau keputusan **tanpa diprogram** secara eksplisit. Machine learning menggunakan teknik statistik untuk memungkinkan komputer belajar dari data dan beradaptasi.



3 Konsep Yang Harus Dipahami Dalam Machine Learning

(1) Data

Data adalah dasar dari machine learning, baik yang terstruktur (terorganisisr dan diberi label) maupun todak terstruktur (teks, gambar, audio)

(2) Model

Model machine learning adalah algoritma yang dilatih dengan menggunakan data untuk mengidentifikasi dan membuat prediksi atau keputusan

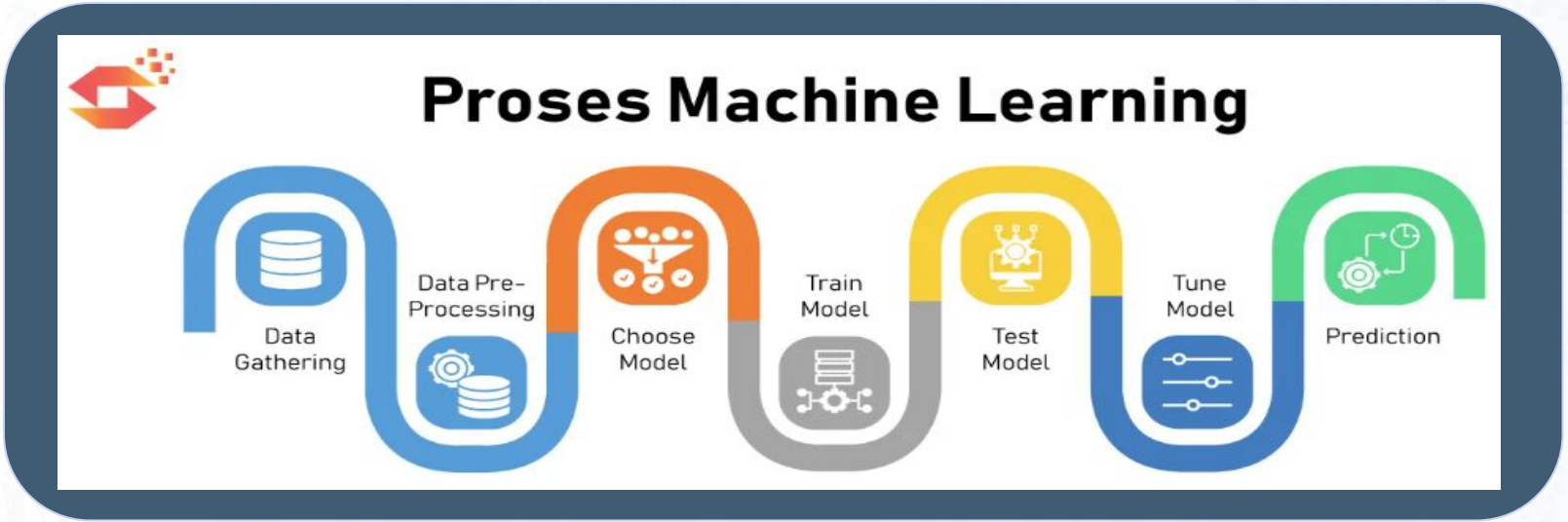
(3) Prediksi

Tujuan utama dari machine learning adalah membuat prediksi atau keputusan yang akurat berdasarkan pola yang dipelajari oleh data

7 Tahap Proses Umum dari Machine Learning

1. **Pengumpulan Data**: Mesin mengumpulkan data yang relevan untuk membangun model dan mengidentifikasi pola yang sesuai.
2. **Pra-pemrosesan Data:** Data yang telah dikumpulkan kemudian diolah, divisualisasikan, dan dipisahkan menjadi dua bagian: Training Set (Set Pelatihan) dan Testing Set (Set Pengujian).
3. **Pemilihan Model**: Pilih model machine learning yang paling tepat berdasarkan output yang diharapkan.
4. **Pelatihan Model:** Ini adalah tahap krusial di mana machine learning menggunakan algoritma yang telah dirancang untuk menemukan pola dan membuat prediksi.
5. **Pengujian Model**: Machine learning menguji pola yang telah ditemukan untuk mengevaluasi kinerja dari proses pelatihan model.
6. **Penyempurnaan Model** (Tune Model): Setelah model dibuat dan diuji, periksa apakah akurasi model dapat ditingkatkan lebih lanjut melalui penyesuaian parameter.
7. **Prediksi**: Pada tahap terakhir, model machine learning yang sudah disempurnakan siap digunakan untuk membuat prediksi berdasarkan data yang telah diproses.

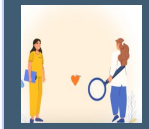
7 Tahap Proses Umum dari Machine Learning



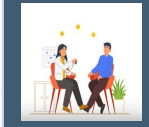
Source : https://sasanadigital.com/perbedaan-artificial-intelligence-machine-learning-dan-deep-learning-serta-contohnya/

Contoh Machine Learnig

Healthcare



Identifying Disease



Personalized Medicine

Drug Discovery

Finance

Fraud Detection

Credit Scoring

Stock Market Prediction

Retail

Customer Segmentation

Demand Forecasting

Recommendation System

03 →

Type of Machine Learning

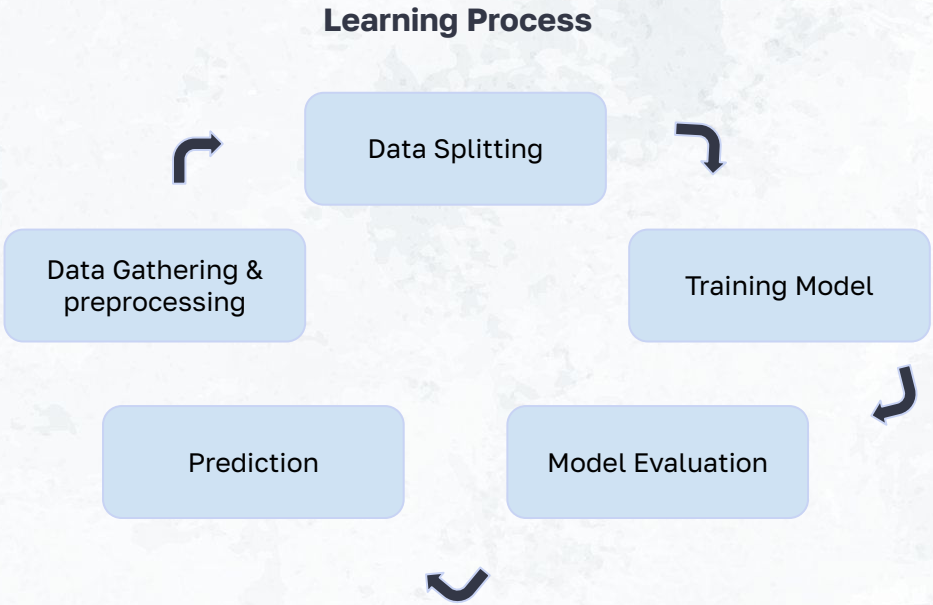
(AI)

Supervised Learning

Melibatkan penggunaan data yang sudah diberi label, bertujuan untuk mempelajari hubungan antara input dan output untuk membuat prediksi atau klasifikasi data baru.

Supervised Learning Example

- Regresi: memprediksi nilai kontinu berdasarkan data input. Contoh → Memprediksi harga rumah berdasarkan luas, jumlah kamar, dan lokasi.
- Klasifikasi: Membagi data menjadi kategori/kelas yang telah ditentukan. Contoh → Mengklasifikasikan email menjadi “spam” atau “non-spam”.
- Time-series Forecasting: Memprediksi hasil berdasarkan data historis. Contoh → Memprediksi penjualan

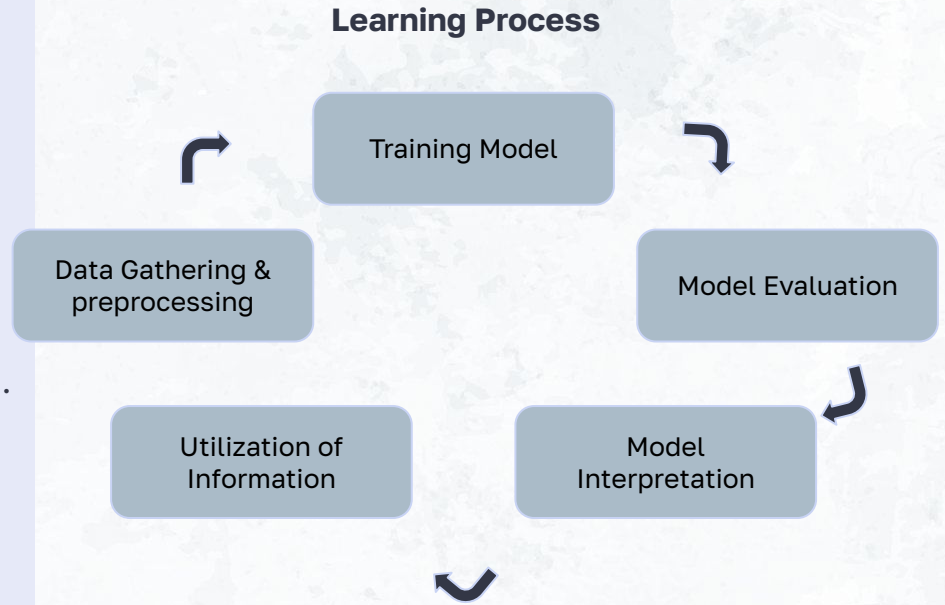


Unsupervised Learning

Melibatkan penggunaan data yang tidak memiliki label, bertujuan untuk menemukan pola, struktur, atau kelompok yang tersembunyi dalam data.

Unsupervised Learning Example

- Clustering: Mengelompokkan data berdasarkan kemiripan fitur/karakteristik yang ada. Contoh → Mengelompokkan pelanggan berdasarkan perilaku pembelian mereka.
- Dimensi Reduksi: Mengurangi dimensi/variabel dalam data untuk memperoleh representasi yang lebih sederhana tetapi tetap mempertahankan informasi penting. Contoh → Principal Component Analysis (PCA).
- Anomaly Detection: Mengidentifikasi data yang tidak biasa/sesuai dengan pola yang ada.
- Association Rule Mining: Mengungkapkan asosiasi/keterkaitan antara utem/variabel dalam data. Contoh → Algoritma apriori untuk analisis keranjang belanja.

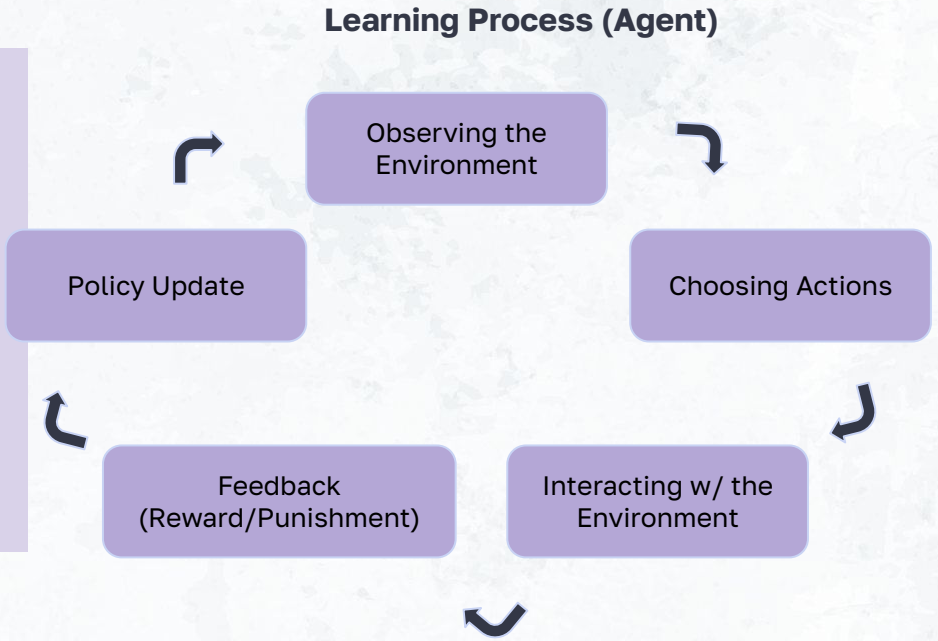


Reinforcement Learning

Agen belajar berinteraksi dengan lingkungan dengan mencoba tindakan-tindakan dan menerima umpan balik (feedback) berupa hadiah/hukuman, bertujuan untuk mempelajari keputusan yang harus diambil oleh agen dalam suatu lingkungan untuk mencapai tujuan tertentu.

Reinforcement Learning Example

- Game: Mengajari agen komputer untuk bermain game seperti catur/video game.
- Robotic: Mengajarkan robot untuk melakukan tugas-tugas kompleks seperti berjalan, mengambil objek, dan berkomunikasi dengan manusia.
- Navigation: Mengajarkan agen untuk mempelajari rute optimal dalam pemetaan/pengiriman barang.
- Finance: Menggunakan reinforcement learning dalam pengambilan keputusan investasi/manajemen risiko.



04 →

Data Preparation

(AI)

Data Preparation Overview

Data Preparation adalah tahap awal untuk memastikan data siap digunakan dalam analisis dan pemodelan.

2 Tahapan Utama

Data Understanding

- Memahami karakteristik, struktur, dan sumber data.
- Meliputi identifikasi tipe variabel, pola distribusi data, deskriptif statistik, dan visualisasi awal untuk menemukan missing values, outliers, serta inkonsistensi.

Data Cleaning

- Membersihkan data dari duplikasi, kesalahan format, nilai ekstrem, missing values, dan inkonsistensi lain yang dapat mengganggu analisis dan performa model.
- Tujuannya meningkatkan kualitas data dan memastikan hasil analisis lebih akurat serta dapat diandalkan.

Data Cleaning Treatment

Proses treatment untuk memperbaiki masalah kualitas data meliputi:

2 Tahapan Utama

Identifikasi Kesalahan

- Duplikasi: mendeteksi dan menghapus entri yang berulang.
- Kesalahan Format: memastikan format data sesuai, seperti tanggal dan tipe data.
- Nilai Ekstrem: mendeteksi nilai di luar kewajaran yang dapat disebabkan kesalahan input.

Penanganan Missing Values

- Menghapus baris data: jika jumlahnya kecil dan tidak mempengaruhi analisis.
- Imputasi nilai: menggantikan nilai hilang dengan mean, median, modus, atau metode prediktif seperti regresi atau algoritma pengisi data.

Data Transformation & Feature Engineering

Proses transformasi dan rekayasa fitur untuk meningkatkan kualitas data serta performa model meliputi:

2 Tahapan Utama

Data Transformation

- Transformasi Logaritmik: mengubah distribusi data yang miring agar lebih simetris.
- Standarisasi: mengubah data agar memiliki mean 0 dan standar deviasi 1 untuk konsistensi skala.
- Normalisasi: mengubah nilai ke rentang tertentu seperti 0–1 agar variabel mudah dibandingkan.
- Encoding Kategorikal: mengubah kategori menjadi angka menggunakan one-hot atau label encoding.

Feature Engineering

- Feature Interaction Creation: menggabungkan fitur untuk membentuk hubungan baru yang lebih prediktif.
- Feature Selection: memilih fitur paling relevan yang berdampak besar terhadap performa model.
- Ekstraksi Fitur Teks: mengubah teks menjadi fitur numerik menggunakan TF-IDF atau word embedding.
- Ekstraksi Fitur Gambar: mengekstraksi fitur dari citra seperti tekstur, bentuk, dan warna.

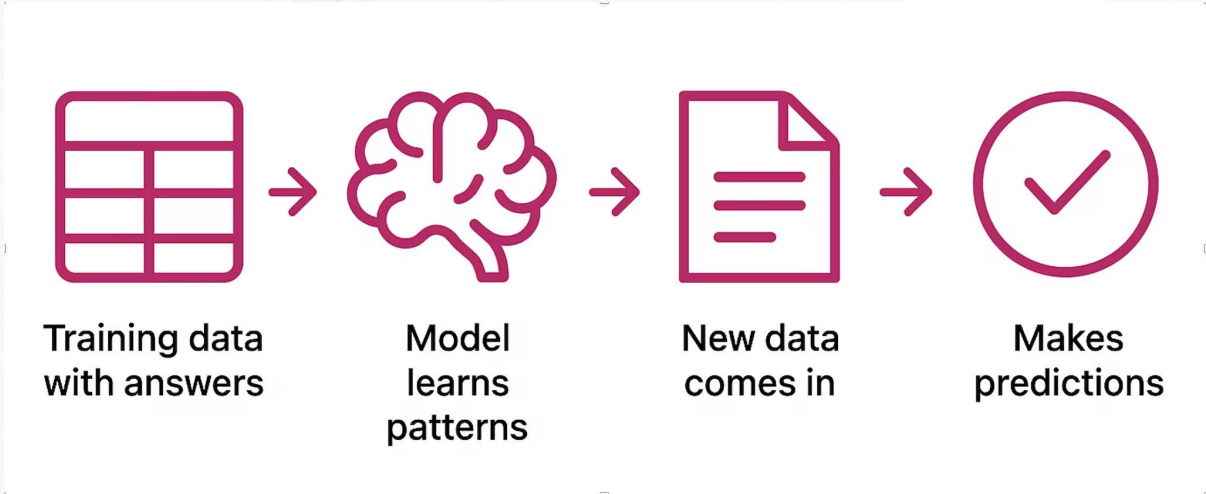
05 →

Model Training

(AI)

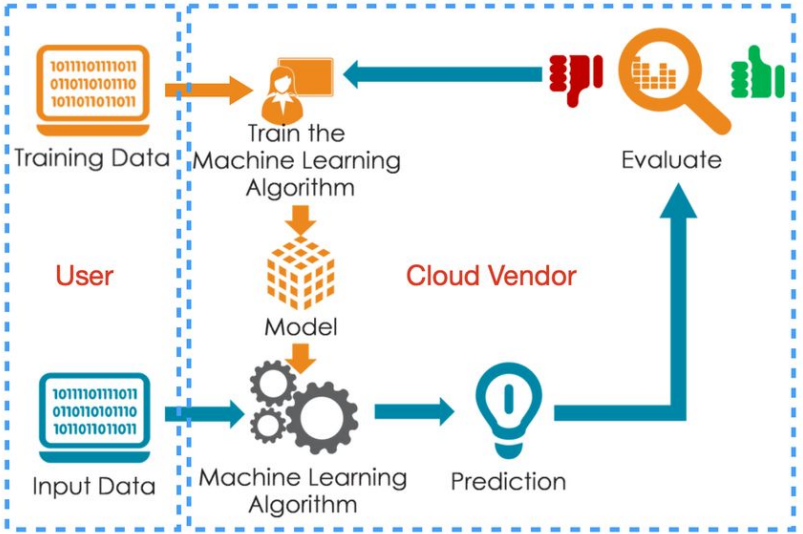
a. Apa itu Model Training

- proses “mengajari” model machine learning untuk mengoptimalkan kinerja pada kumpulan data pelatihan yang relevan dengan contoh penggunaan akhir model.
- model belajar untuk mengenali pola dan hubungan dalam data dengan menghubungkan input dan output yang diberikan.
- Komputer mengenali pola, hubungan antar data, dan akhirnya membentuk sebuah model prediktif.



b. Bagaimana proses belajar dalam Model Training

Proses belajar dalam model training dilakukan dengan menyesuaikan parameter dan bobot model secara bertahap hingga mencapai nilai yang optimal, sehingga model mampu membuat prediksi atau klasifikasi yang lebih akurat.



c. Paradigma Mekanisme Model Training

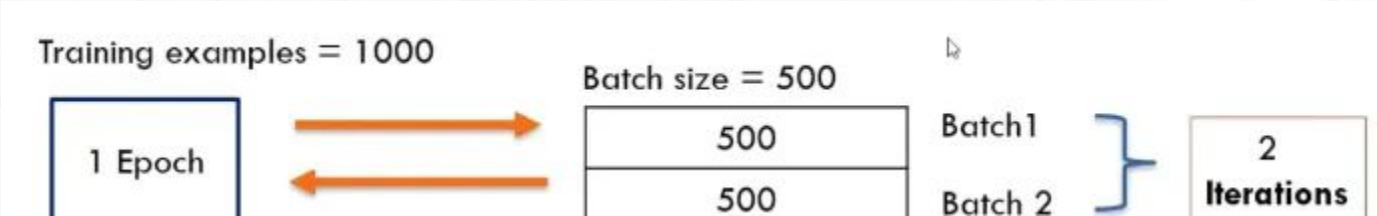
1. Iteration or Epoch

Epoch

- satu siklus lengkap di mana model telah "melihat" dan memproses seluruh kumpulan data pelatihan (dataset).
- Dengan kata lain, Satu epoch = seluruh dataset digunakan sekali untuk melatih model.
- Jika dataset ada 1.000 baris, maka satu epoch berarti model memproses semua 1.000 baris satu kali.
- Jika dianalogikan, epoch seperti membaca satu buku dari awal sampai akhir satu kali.

Iteration

- 1 iteration = 1 kali update model berdasarkan 1 batch data
- Contoh: Jika menggunakan batch size (Banyaknya data dalam sekali update):
  dataset = 1.000 sampel
  batch size = 100
  → 1 epoch = 10 iteration
  Analogi: 1 batch = membaca 1 bab.
  Iteration = selesai baca 1 bab → otak mendapatkan update pengetahuan sedikit.



c. Paradigma Mekanisme Model Training

2. Hyperparameter

a) Variabel konfigurasi eksternal yang digunakan Data Scientist untuk mengelola pelatihan model *machine learning*.
b) Parameter yang *kita tentukan di awal* sebelum training dimulai, dan *tidak dipelajari oleh model secara otomatis*.
c) Hyperparameter menentukan **bagaimana proses training berlangsung**.
d) contoh *hyperparameter* umum:
   - *Learning rate* adalah tingkat perkiraan pembaruan algoritma
   - *Learning rate decay* adalah pengurangan bertahap dalam tingkat pembelajaran dari waktu ke waktu untuk mempercepat pembelajaran
   - *Momentum* adalah arahan langkah berikutnya sehubungan dengan langkah sebelumnya
   - *Neural network nodes* mengacu pada jumlah simpul di setiap lapisan tersembunyi
   - *Neural network layers* mengacu pada jumlah lapisan tersembunyi dalam jaringan neural
   - *Mini-batch size* adalah ukuran *batch* data pelatihan
   - *Epochs* adalah frekuensi seluruh set data pelatihan ditampilkan ke jaringan selama pelatihan
   - *Eta* adalah penyusutan ukuran langkah untuk mencegah *overfitting*

06 →

Model Tuning & Optimization

(AI)

Model Tuning

Ada beberapa metode untuk melakukan model tuning, antara lain:

(1) Grid Search

Memilih kombinasi nilai yang diinginkan untuk setiap hyperparameter dan menguji semua kombinasi secara sistematis

(2) Random Search

Menggunakan pendekatan acak untuk mencoba berbagai kombinasi hyperparameter dalam rentang yang ditentukan

(3) Bayesian Optimization

Menggunakan teknik pemodelan statistik untuk menggantikan random search dan memperkirakan kombinasi hyperparameter yang lebih baik secara iteratif

(4) Gradient-Based Optimization

Menggunakan algoritma optimasi berbasis gradien, seperti algoritma Adam atau RMSprop, untuk menentukan nilai optimal parameter dalam model

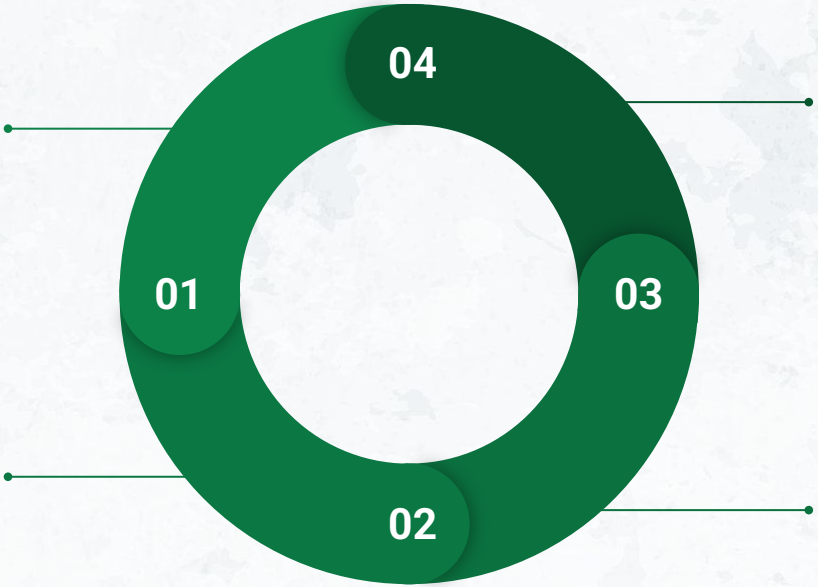
Model Optimization

Melakukan Tuning

Menjalankan penyetelan model dengan mencoba berbagai kombinasi hyperparameter dalam rentang yang ditentukan

Performance Evaluation

Menggunakan evaluation metrics yang relevan untuk mengevaluasi performa model pada setiap kombinasi hyperparameter



Validasi Akhir

Menguji model terpilih pada data yang belum pernah dilihat sebelumnya untuk memvalidasi performa yang sebenarnya

Memilih Model Terbaik

Memilih model dengan kombinasi hyperparameter yang memberikan performa terbaik pada data validasi

Model Evaluation Metrics - Regression

Root Mean Squared Error (RMSE)

Mengukur akar rata-rata dari selisih kuadrat antara nilai prediksi dan nilai sebelumnya

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

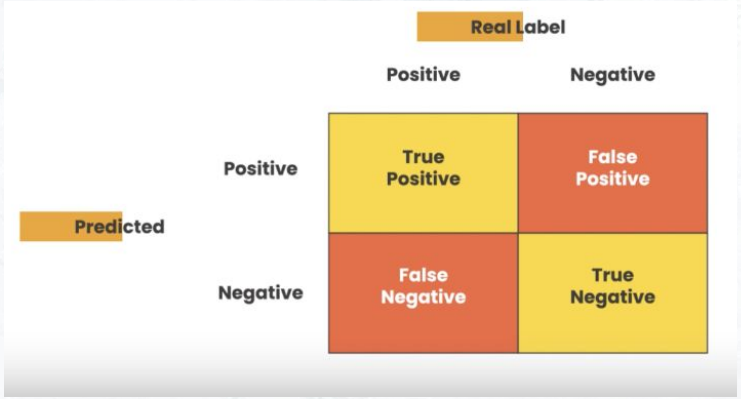
Mean Absolute Error (MAE)

Mengukur rata-rata dari selisih absolut antara nilai prediksi dan nilai sebelumnya

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Model Evaluation Metrics - Classification

Accuracy	Predictions/ Classifications	$\frac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\frac{\text{2 * True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$
IoU	Object Detections/ Segmentations	$\frac{\text{Pixel Overlap}}{\text{Pixel Union}}$



Confusion Matrix

Source:

https://www.researchgate.net/figure/Evaluation-metrics-accuracy-precision-recall-F-score-and-Intersection-over-Union_fig2_358029719

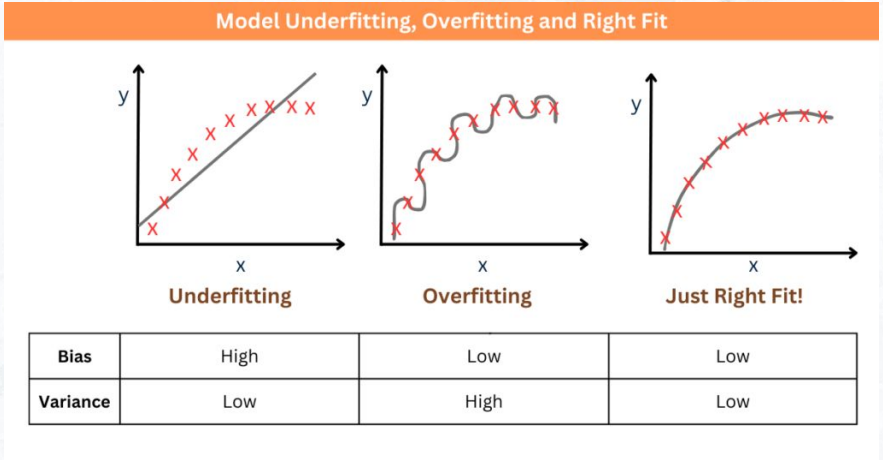
Bias and Variance Tradeoff

Bias

Mengukur kesalahan sistematis suatu model dalam mempelajari hubungan antara fitur dan target.

Variance

Mengukur sejauh mana model sensitif terhadap fluktuasi data pelatihan



Bias	High	Low	Low
Variance	Low	High	Low

Source:

https://machinemindscape.com/overfitting-underfitting-and-models-capacity-in-deep-learning/

Underfitting

Underfitting terjadi ketika model machine learning tidak cukup rumit atau adaptif terhadap data pelatihan sehingga tidak dapat mempelajari pola yang ada.

Tanda-tanda Underfitting:

- Performa yang buruk pada data training dan data testing
- Hasil Prediksi yang terlalu sederhana dan tidak akurat
- Learning curve yang lambat atau stagnan

Penanganan Underfitting:

- Menggunakan model yang lebih kompleks atau algorima machine learning yang lebih kuat/robust
- Menambahkan fitur yang lebih relevan atau menggunakan teknik transformasi data
- Memperluas dataset dengan mengumpulkan lebih banyak data yang representatif
- Menguji dan memodifikasi hyperparameter model untuk meningkatkan kompleksitas

Overfitting

Underfitting terjadi ketika model machine learning terlalu rumit atau terlalu adaptif terhadap data pelatihan sehingga tidak dapat umum dalam memprediksi data baru.

Tanda-tanda Overfitting:

- Performa yang baik pada data training, tetapi performa buruk pada data testing atau data baru
- Variasi yang tinggi antara hasil prediksi pada setiap iterasi pelatihan
- Model terlalu menyesuaikan noise atau kesalahan dalam data pelatihan

Penanganan Overfitting:

- Menggunakan lebih banyak data training untuk membantu model melihat pola yang lebih umum
- Menggunakan teknis regularisasi seperti pinalti L1 atau L2 untuk mengurangi kompleksitas model
- Mengurangi jumlah fitur yang tidak relevan atau menggunakan seleksi fitur
- Menggunakan teknik ensemble learning seperti penggabungan model

07 →

Model Deployment

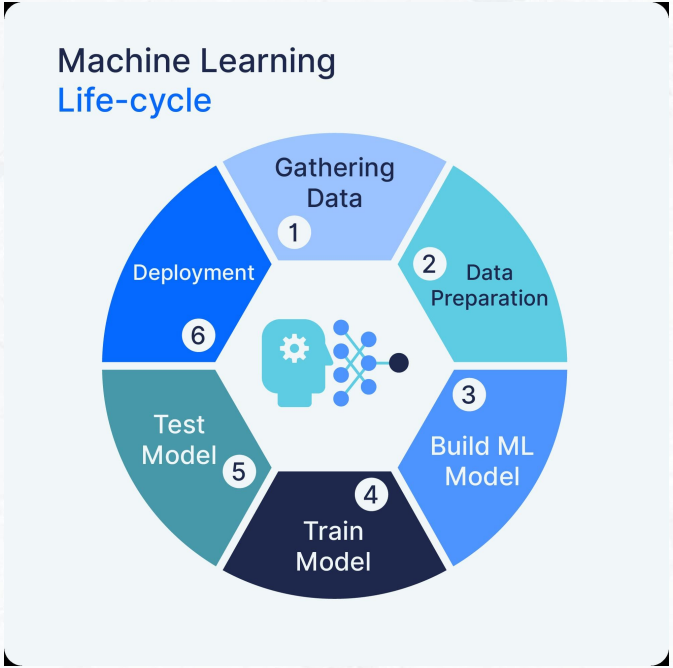
(AI)

Model Deployment

Model Deployment adalah proses mengimplementasikan model machine learning yang telah dilatih ke dalam produksi atau lingkungan yang dapat digunakan secara nyata.

Tujuan dari Model Deployment

Membuat model dapat digunakan secara efisien dan memberikan manfaat dalam menjawab pertanyaan bisnis atau menyelesaikan masalah yang dihadapi.



Source:
https://www.akkio.com/post/deploying-ml-models-a-guide

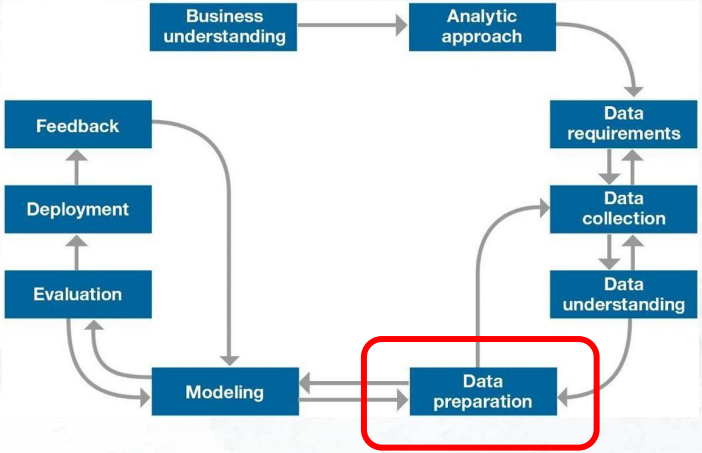
08 →

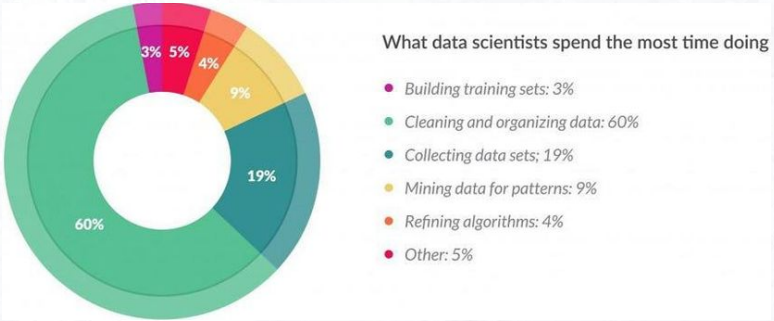
Data Preprocessing

(AI)

Data Preprocessing

Data Science Life Cycle





Data Preprocessing

Data → **Machine Learning Model**

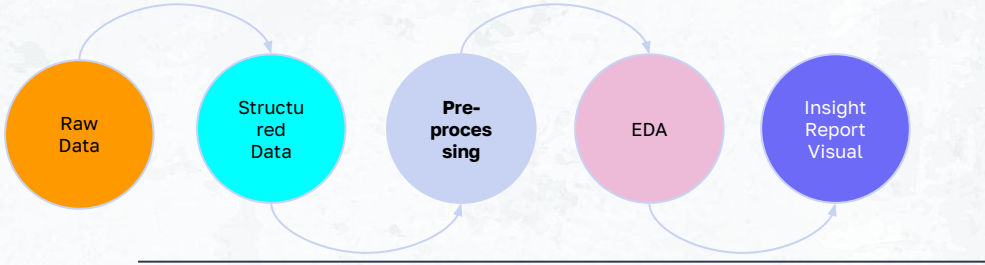
Aktivitas Data Preprocessing:

1) Pembersihan data kotor
2) Pengisian data kosong
3) Penyesuaian format
4) Perubahan format
5) Disaggregasi

Positive impact:

1) Data bisa dijalankan
2) Performa model meningkat
3) Model lebih efisien

Preprocessing sebagai bagian paling menantang



09 →

Feature and Feature Engineering

(AI)

Feature (Variable)

Definisi

Fitur adalah properti terukur dari objek yang Anda coba analisis. Dalam kumpulan data, fitur muncul sebagai kolom.

Jenis-jenis fitur

1. Fitur Kategoris
   - Nominal: skala kualitatif, hanya untuk membedakan, tidak ada tingkatan. Contoh: gender, warna rambut, warna mata.
   - Ordinal: skala kualitatif dengan tingkatan-tingkatan. Contoh: jenjang pendidikan, kepuasan pelanggan.
2. Fitur Numerik
   - Discrete: dapat dihitung, finite maupun infinite. Contoh: jumlah siswa, jumlah kendaraan.
   - Continuous: mewakili pengukuran, tidak dapat dihitung, dijelaskan menggunakan interval bilangan real. Contoh: tinggi, suhu, kecepatan.

Predictor dan Target

- Predictor = independent variable, variable penjelas, features, disimbolkan dengan X.
- Target = variable respon, dependent variable.

Feature Engineering

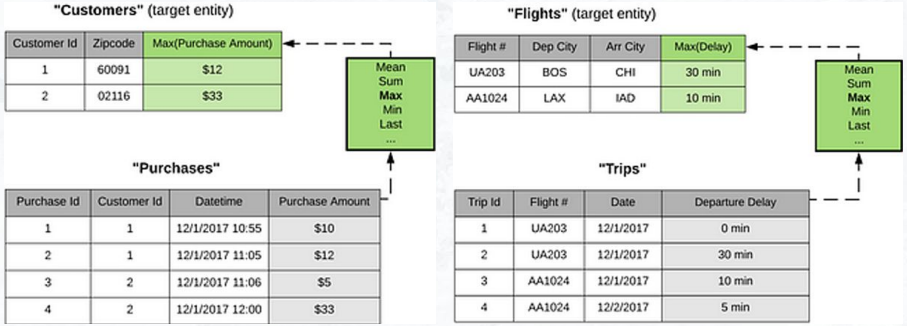
Definisi

- Teknik machine learning yang memanfaatkan data untuk membuat variabel baru yang tidak ada dalam dataset yang telah disediakan.
- Ini dapat menghasilkan fitur baru dengan tujuan menyederhanakan dan mempercepat transformasi data sambil juga meningkatkan akurasi model.

Sq Ft.	Amount
2400	9 Million
3200	15 Million
2500	10 Million
2100	1.5 Million
2500	8.9 Million

→

Sq Ft.	Amount	Cost Per Sq Ft
2400	9 Million	4150
3200	15 Million	4944
2500	10 Million	3950
2100	1.5 Million	510
2500	8.9 Million	3600



10 →

Pengecekan Kualitas Data

(AI)

Missing Value Handling

Jumlah data yang missing > 60%

- Menghapus kolom tersebut

Categoric

- Mengisi dengan kategori paling banyak
- Mengisi dengan “other”

Numeric

- Mengisi dengan rata-rata/median
- Mengisi dengan nilai 0

```python
# importing pandas as pd
import pandas as pd

# importing numpy as np
import numpy as np

# dictionary of lists
dict = {'First Score':[100, 90, np.nan, 95],
        'Second Score': [30, 45, 56, np.nan],
        'Third Score':[np.nan, 40, 80, 98]}

# creating a dataframe from dictionary
df = pd.DataFrame(dict)

# filling missing value using fillna()
df.fillna(0)
```

Duplicate Data Handling

Dampaknya Model akan lebih dominan memprediksi data yang duplikat
Cara Mengatasi Menghapus baris yang duplikat adalah dibuang (drop duplicate)

DATA SCIENCE PARICHAY

pandas

	Pet	Color	Eyes
0	Cat	Brown	Black
1	Dog	Golden	Black
2	Dog	Golden	Black
3	Dog	Golden	Brown
4	Cat	Black	Green

→

	Pet	Color	Eyes
0	Cat	Brown	Black
1	Dog	Golden	Black
3	Dog	Golden	Brown
4	Cat	Black	Green

Drop duplicates

Incosistent Data Handling

Dampaknya

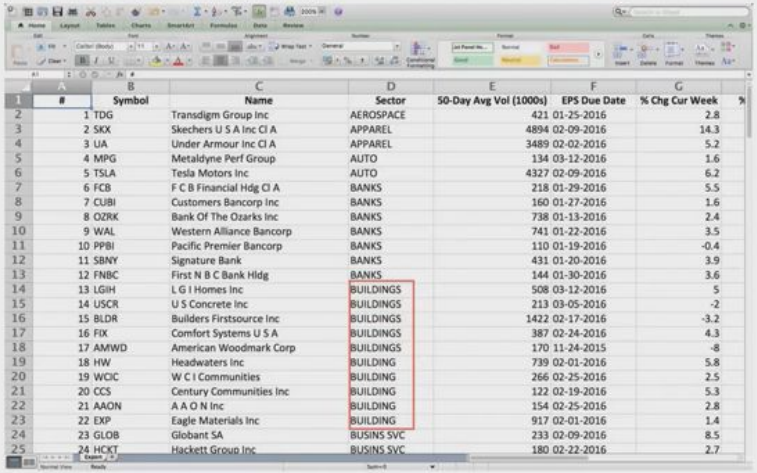
- Model tidak akan berjalan (error)
- Performa model menjadi berkurang

Ciri-ciri inconsistency value

- Missing value diisi dengan ‘?’ atau ‘-’ atau karakter lainnya
- Adanya data dengan tipe kategorik di fitur umur
- Beda karakter dan standard penulisan

Cara Mengatasi

- Tiap kasus berbeda-beda



#	Symbol	Name	Sector	50-Day Avg Vol (1000s)	EPS Due Date	% Chg Cur Week
1	TDG	Transdigm Group Inc	AEROSPACE	421	01-25-2016	2.8
2	SKX	Skechers U S A Inc Cl A	APPAREL	4894	02-09-2016	14.3
3	UA	Under Armour Inc Cl A	APPAREL	3489	02-02-2016	5.2
4	MPG	Metaldyne Perf Group	AUTO	134	03-12-2016	1.6
5	TSLA	Tesla Motors Inc	AUTO	4327	02-09-2016	6.2
6	FCB	F C B Financial Hdg Cl A	BANKS	218	01-29-2016	5.5
7	CUBI	Customers Bancorp Inc	BANKS	160	01-27-2016	1.6
8	OZRK	Bank Of The Ozarks Inc	BANKS	738	01-13-2016	2.4
9	WAL	Western Alliance Bancorp	BANKS	741	01-22-2016	3.5
10	PPBI	Pacific Premier Bancorp	BANKS	110	01-19-2016	-0.4
11	SBNY	Signature Bank	BANKS	431	01-20-2016	3.9
12	FNBC	First N B C Bank Hldg	BANKS	144	01-30-2016	3.6
13	LGIH	L G I Homes Inc	BUILDINGS	508	03-12-2016	5
14	USCR	U S Concrete Inc	BUILDINGS	213	03-05-2016	-2
15	BLDR	Builders Firstsource Inc	BUILDINGS	1422	02-17-2016	-3.2
16	FIX	Comfort Systems U S A	BUILDINGS	387	02-24-2016	4.3
17	AMWD	American Woodmark Corp	BUILDINGS	170	11-24-2015	-8
18	HW	Headwaters Inc	BUILDING	739	02-01-2016	5.8
19	WCIC	W C I Communities	BUILDING	266	02-25-2016	2.5
20	CCS	Century Communities Inc	BUILDING	122	02-19-2016	5.3
21	AAON	A A O N Inc	BUILDING	154	02-25-2016	2.8
22	EXP	Eagle Materials Inc	BUILDING	917	02-01-2016	1.4
23	GLOB	Globant SA	BUSINS SVC	233	02-09-2016	8.5
24	HCKT	Hackett Group Inc	BUSINS SVC	180	02-22-2016	2.7

11 →

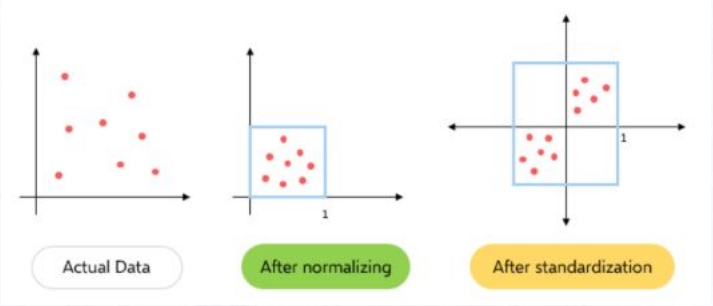
Metode Normalization & Standardization Data

(AI)

Normalization and Standardization

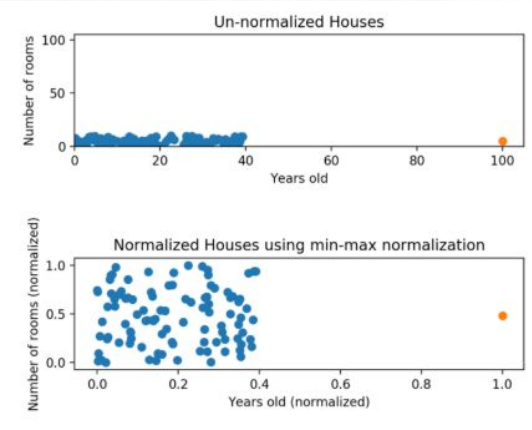
Normalization → Proses mengubah nilai-nilai suatu feature menjadi skala tertentu.

Standardization → Proses mengubah nilai-nilai feature sehingga mean = 0 dan standard deviation = 1



Kenapa perlu ini?

- Menjamin algoritma pembelajaran memperlakukan semua feature dengan adil
- Mempercepat algoritma pembelajaran
- Mempermudah interpretasi beberapa model ML



Kapan menggunakan normalization atau standardization?

Sl. No	Normalization	Standardization
1	Feature scaling method to bring the data into common range such as [0, 1], [-1, 1], etc.	Feature scaling method bring the data with mean 0 and unit variance
2	Scikit-learn provides MinMaxScaler, MaxAbsScaler and RobustScaler methods for normalization	Scikit-learn provides StandardScaler for standardization
3	MinMaxScaler and MaxAbsScaler are sensitive to outliers whereas RobustScaler is more robust to outliers	Standardization is less sensitive to outliers compared to MinMaxScaler and MaxAbsScaler
4	Useful when we don't know about the distribution of features and there are no or little outliers - MinMaxScaler: if features don't follow normal distriubtion and if there are no or less outliers - MaxAbsScaler: if the data is sparse - RobustScaler: if the data contains outliers	Useful when we know features are normally distributed (Gaussian distribution)

12 →

Label Encoder and One-Hot Encoder

(AI)

Binary Encoding vs One-Hot Encoding

Proses pengubahan fitur kategorikal untuk meningkatkan kualitas input model serta mencegah kesalahan interpretasi oleh algoritma.

2 Tahapan Utama

Binary Encoding

- Mengubah kategori menjadi angka ordinal, lalu direpresentasikan dalam bentuk biner.
- Jumlah kolom lebih sedikit sehingga lebih efisien untuk dataset dengan kategori sangat banyak.
- Cocok digunakan ketika terdapat ratusan hingga ribuan kategori, seperti produk di e-commerce.
- Kelemahan: model dapat salah menganggap kategori memiliki hubungan matematis karena nilai ordinalnya.

One-Hot Encoding

- Membuat kolom baru untuk setiap kategori, nilai berisi 1 atau 0.
- Menghilangkan hubungan antar kategori sehingga lebih aman untuk model berbasis jarak atau regresi.
- Cocok untuk dataset dengan jumlah kategori sedikit.
- Kelemahan: jumlah kolom dapat membengkak (curse of dimensionality) bila kategori terlalu banyak.

13 →

Imbalance Data

(AI)

a. Pengertian Imbalance Data

- Merupakan masalah umum dalam Machine learning, dimana distribusi sampel di antara kelas yang berbeda tidak seragam, yang menyebabkan hasil yang bias dalam prediksi model.
- Atau dengan kata lain, kondisi ketika sebuah dataset memiliki distribusi jumlah sampel yang tidak seimbang antara satu kelas dengan kelas lainnya. Artinya, ada kelas yang jumlah datanya sangat banyak (mayoritas) dan ada kelas yang jumlahnya sangat sedikit (minoritas).
- Kondisi ini bisa menimbulkan masalah:
  1. Biased Learning: Model menjadi baik dalam memprediksi apa yang umum (mayoritas) tetapi kesulitan memahami hal-hal yang kurang umum (minoritas). Model machine learning mengalami bias terhadap kelas mayoritas.
  2. Misleading Accuracy: Model machine learing cenderung memiliki akurasi tinggi ketika memprediksi kelas yang mayoritas dan ini bisa menyesatkan karena tidak menunjukkan seberapa baik komputer bekerja di kelas minoritas.
- Contoh:
  - Pada kasus pengumpulan data pelatihan untuk model yang memprediksi kondisi medis.
  - Sebagian besar data pasien yang dikumpulkan, misalkan terdapat 95% data yang masuk ke dalam kelompok "sehat", sementara pasien yang sakit hanya mencakup sebagian kecil (5%) dari data tersebut.
  - Selama pelatihan, model klasifikasi belajar bahwa ia dapat mencapai akurasi 95% ketika memprediksi "sehat" untuk setiap data yang dihadapinya.
  - Ini adalah masalah besar karena yang benar-benar diinginkan oleh dokter adalah agar model dapat mengidentifikasi pasien yang menderita kondisi medis (sakit).

b. Cara *handling* Imbalance Data

1. **Collect more data**

- Mencari/melengkapi data yang kurang dengan data baru atau data sintetis.
- Cara ini membuat distribusi kelas menjadi lebih seimbang secara alami

2. **Undersampling**

- Menghapus sampel dari kelas mayoritas sampai memiliki distribusi data yang setara.
- memiliki keuntungan: relatif mudah diterapkan, dapat meningkatkan waktu proses model dan biaya komputasi.
- Harus dilakukan dengan hati-hati karena menghapus sampel dari dataset asli dapat mengakibatkan hilangnya informasi yang berguna.

3. **Oversampling**

- Meningkatkan jumlah sampel di kelas minoritas hingga komposisinya sama dengan kelas mayoritas.
- Hal ini dapat dilakukan dengan membuat salinan sampel di kelas yang kurang terwakili sehingga model menghadapi jumlah sampel yang sama dari setiap kelas.
- Untuk mengurangi resiko overfitting dapat dibuat berbagai augmentasi untuk mensimulasikan dataset yang lebih beragam.

14 →

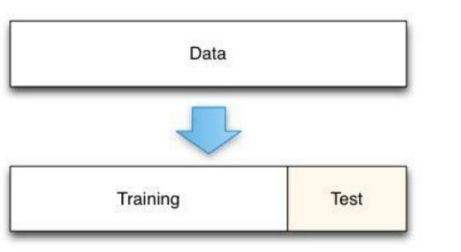
Data Split

(AI)

Data Split Method

Train Test Split

Metode membagi dataset menjadi dua bagian yaitu data training dan data test. Dengan komposisi 70%(training)/30%(test) atau 80(training)/20(test).



Train Validation Test split

Metode membagi dataset menjadi tiga bagian yaitu training set, dev set dan test set. Metode ini digunakan untuk dataset yang besar. Dev set digunakan untuk hyperparameter tuning.

Cross Validation

Membagi dataset menggunakan train test split kemudian dilakukan iterasi/pengulangan untuk menguji model kemudian hasil setiap iterasi akan dijumlah dan dirata-ratakan untuk menjadi hasil akhir cross validation.