Session 29

Data Preprocessing for ML

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

1. What is Data Preprocessing
2. Data cleaning (imputation)
3. Data transformation (one-hot, label encoding)
4. Normalization, standardization

DigitalSkola
Uncover The World Of Digital Skills With Us

Data Preprocessing

Apa itu features/ variables

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

Sex	Age	BMI	DM type	DM duration	FBS	Sys BP	Dias BP	Retinopathy
Male	65	25	II	20	129	130	80	Yes
Male	42	27	II	300	210	140	90	No
Female	31	21	I	11	164	145	80	Yes
Male	70	32	II	29	208	160	100	Yes
Female	54	34	II	6	183	155	95	No
	46	29	II	7	198	160	100	No
Female	16	24	I	-1	250	135	80	No
Male	67	30	II	12	243	165	90	Yes
Female	51	28	II	7	163	130	85	No
Girl	70	36	II	20	250	150	90	Yes
Female	63	35	II	14	203	160	110	No
Male	44	39	II	3	149	140	90	No
Boy	51	24	II	9	160	155	80	No
Male	27	19	I	5	170	140	90	No

- Kualitas fitur dalam kumpulan data memiliki dampak besar pada kualitas wawasan yang akan diperoleh saat pemodelan machine learning.

Jenis-jenis Fitur

- Jenis Fitur Kategoris
  - Fitur Nominal
    - Fitur nominal yang sering juga disebut skala kualitatif adalah skala data yang berfungsi hanya untuk membedakan dan tidak ada tingkatan diantaranya.
    - Contoh : Gender, Warna Rambut, Warna Mata
  - Fitur Ordinal
    - Fitur Ordinal atau skala kualitatif di mana data dikelompokkan menjadi orde atau tingkatan tingkatan.
    - Contoh : Jenjang Pendidikan, Kepuasan Pelanggan
- Jenis Fitur Numerik
  - Fitur Discrete
    - Data diskrit mewakili item yang dapat dihitung.
    - Contoh : Jumlah Siswa, Jumlah Kendaraan, dll
  - Fitur Continuous
    - Data kontinu mewakili item yang dapat diukur.
    - Contoh: Tinggi, Suhu, Kecepatan, dll

What is Data preprocessing

Data

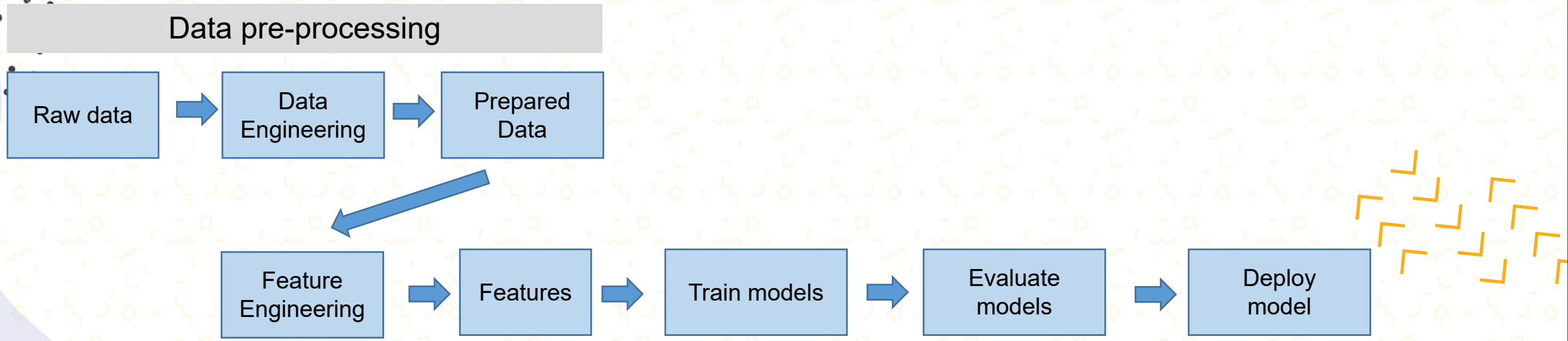
Preprocessing

Machine Learning Model

- Data cleaning
- Data integration
- Data reduction
- Data transformation

Data preprocessing

- *Data preprocessing* merupakan sekumpulan teknik yang diterapkan pada dataset untuk menghapus noise, meng-handle missing value, dan data yang tidak konsisten.
- *Feature engineering* adalah proses mengubah data mentah menjadi fitur yang siap dipakai oleh model ML.
- *Feature engineering* terdiri dari pembuatan fitur, sedangkan *data preprocessing* melibatkan pembersihan data.



Tugas Utama dalam Data Preprocessing

- Data cleaning
  - Fill in missing values,
  - Smooth noisy data,
  - Identify or remove outliers, and
  - Resolve inconsistencies
- Data integration
  - Integration of multiple databases, or files
- Data reduction
  - Dimensionality reduction
- Data transformation
  - Normalization
  - Standardization
  - Encoding

Data Cleaning

- Data di dunia nyata itu ‘kotor’.
  - Kosong atau tidak lengkap
    - pekerjaan=“ ”
  - Noisy: nilai yg salah atau outliers
    - gaji=“-10”
  - Nilai tidak konsisten
    - jenis kelamin=“perempuan”
    - vs. jenis kelamin=“wanita”
  - Data yang sama/ duplicate

- *No quality data, no quality mining results*!

Sex	Age	BMI	DM type	DM duration	FBS	Sys BP	Dias BP	Retinopathy
Male	65	25	II	20	129	130	80	Yes
Male	42	27	II	300	210	140	90	No
Female	31	21	I	11	164	145	80	Yes
Male	70	32	II	29	208	160	100	Yes
Female	54	34	II	6	183	155	95	No
	46	29	II	7	198	160	100	No
Female	16	24	I	-1	250	135	80	No
Male	67	30	II	12	243	165	90	Yes
Female	51	28	II	7	163	130	85	No
Girl	70	36	II	20	250	150	90	Yes
Female	63	35	II	14	203	160	110	No
Male	44	39	II	3	149	140	90	No
Boy	51	24	II	9	160	155	80	No
Male	27	19	I	5	170	140	90	No

Incomplete (Missing) Data

- Data tidak selalu tersedia
  - Misalnya, banyak baris tidak memiliki nilai untuk beberapa atribut, seperti pendapatan pelanggan dalam data penjualan
- Data yang hilang mungkin karena
  - kerusakan peralatan
  - data tidak masuk karena ada kesalah pahaman
  - data tertentu mungkin tidak dianggap penting pada waktu *proses entri*

How to Handle Missing Data?

- Abaikan baris:
- Isi nilai yang hilang secara manual: butuh waktu lama?
- Isi secara otomatis dengan
  - konstanta global: misalnya, "unknown",
  - atribut mean, median (untuk numerik)
  - rata-rata atribut untuk semua sampel yang termasuk dalam kelas yang sama
  - nilai yang paling sering muncul (untuk kategoris)

Noisy Data

- Noise adalah data yang berisi nilai-nilai yang salah atau anomali, yang biasanya disebut juga outlier.
- Nilai atribut yang salah mungkin karena
  - instrumen pengumpulan data yang salah
  - terjadi masalah pada saat entri data
  - terjadi masalah pada transmisi data

How to Handle Noisy Data?

- Binning
  - urutkan data dan partisi terlebih dahulu ke dalam *bin* (frekuensi yang sama)
  - kemudian dapat mengganti nilai outlier dengan nilai rata rata atau median dalam *bin* tersebut.
- Regression
  - *smooth training data* dengan fungsi regresi / mengganti outlier berdasarkan fungsi regresi
- Clustering
  - mendeteksi dan menghapus outlier
- Combined computer and human inspection
  - mendeteksi nilai yang mencurigakan dan diperiksa oleh manusia (misalnya, menangani kemungkinan outlier)

DigitalSkola
Uncover The World Of Digital Skills With Us

Challenges

location	date_of_sale	property_size_sq_m	number of bedrooms	price	type
Clapham	12/4/1999	58	1	729000	apartment,1930s
Ashford	5/8/2017	119	3	699000	semi-detached,1970s
Stratford-on-Avon	29/3/2012	212	3	540000	detached,17th century
Canterbury	1/7/2009	95	2	529000	teraced,1960s
Camden	16/12/2001	54	1	616000	apartment,2000s
Rugby	1/3/2003	413	7	247000	detached, 19th century
Hampstead	5/3/2016	67	2	890000	terraced, 19th century

1. Numeric vs category
2. Measurement scale

Apa itu Feature Encoding?

- One-Hot Encoding
  - Mengubah setiap kategori sehingga memiliki nilai angka 1 atau angka 0

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

- Label Encoding
  - Mengubah setiap kategori menjadi angka 1,2,3, … dst

petallength	petalwidth	iris_class
1.4	0.2	Iris-setosa
1.4	0.2	Iris-versicolor
1.3	0.2	Iris-virginica

petallength	petalwidth	iris_class
1.4	0.2	1
1.4	0.2	2
1.3	0.2	3

Normalizalization dan Standardization

- Normalization adalah proses mengubah nilai-nilai suatu feature menjadi skala tertentu [0,1].
- Standardization adalah proses mengubah nilai-nilai feature sehingga mean = 0 dan standard deviation = 1

- **Min-Max Scaling**

  Uses MinMaxScaler

  Transform to defined range

$$y = \frac{x - \min x_i}{\max x_i - \min x_i}$$

- **Standardization**

  Uses StandardScaler

  Transform to mean=0, sd=0

$$y = \frac{x - \bar{x}}{s}$$

Where

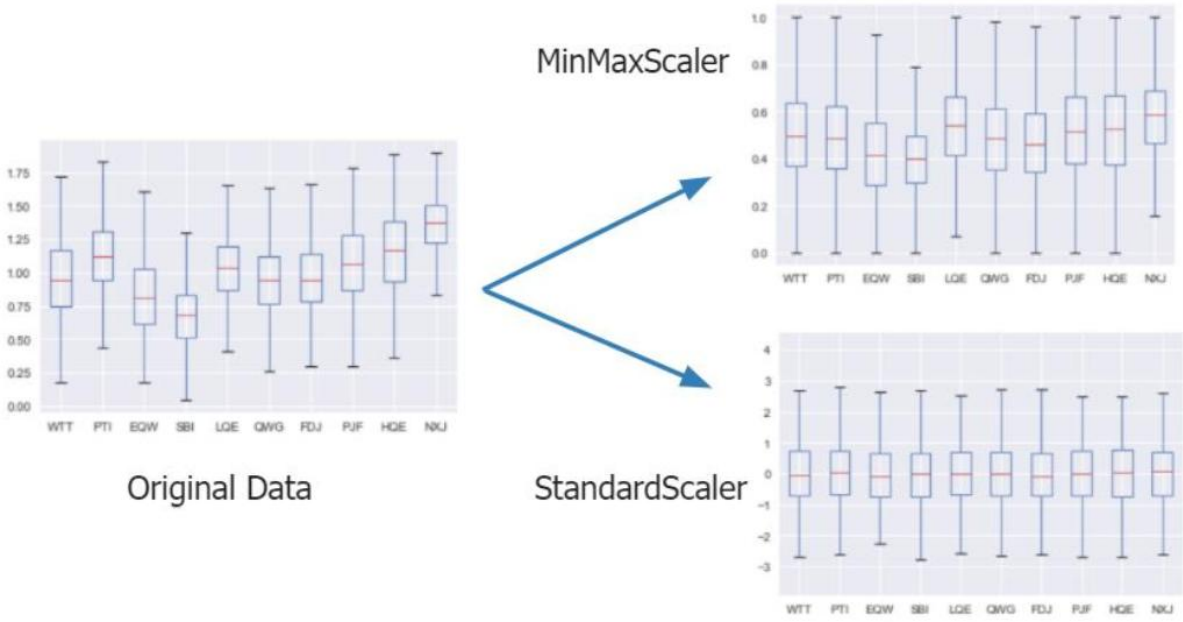
$\bar{x}$ = mean

s = Standard deviation

Normalizalization dan Standardization

- Tujuan
  - Data dengan skala yang sama akan menjamin algoritma pembelajaran memperlakukan semua feature dengan adil
  - Data dengan skala yang sama dan centered akan mempercepat algoritma pembelajaran
  - Data dengan skala yang sama akan mempermudah interpretasi beberapa model ML

- Kapan penggunaan:
  - Gunakan standardization bila kita tahu data punya sebaran normal/gaussian



Train test split

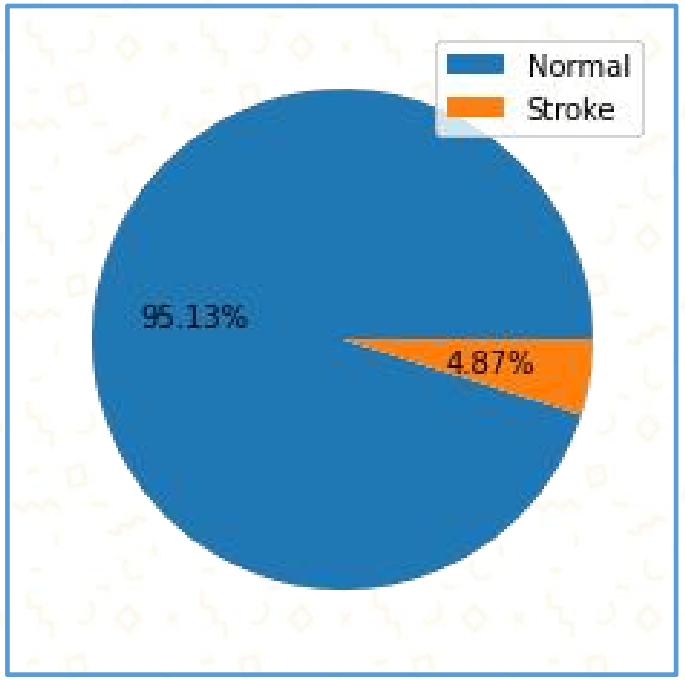
- Training adalah proses ketika model mempelajari data
- Hasil dari training disebut model machine learning (trained model)
- Untuk membuktikan keakuratan model, diperlukan data uji (test data)
- Training set : subset untuk melatih model.
- Test set : subset untuk menguji model yang dilatih.
- Karena kurangnya data, kita bisa memisahkan dataset menjadi dua bagian yaitu training dan testing

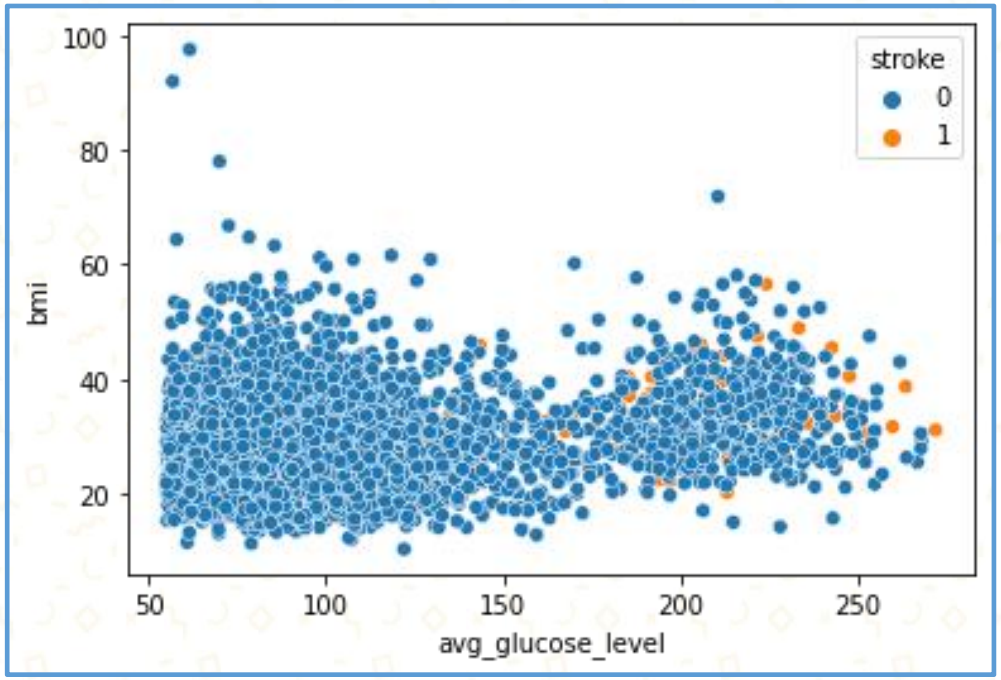
Training

Testing / Proving

Imbalanced dataset

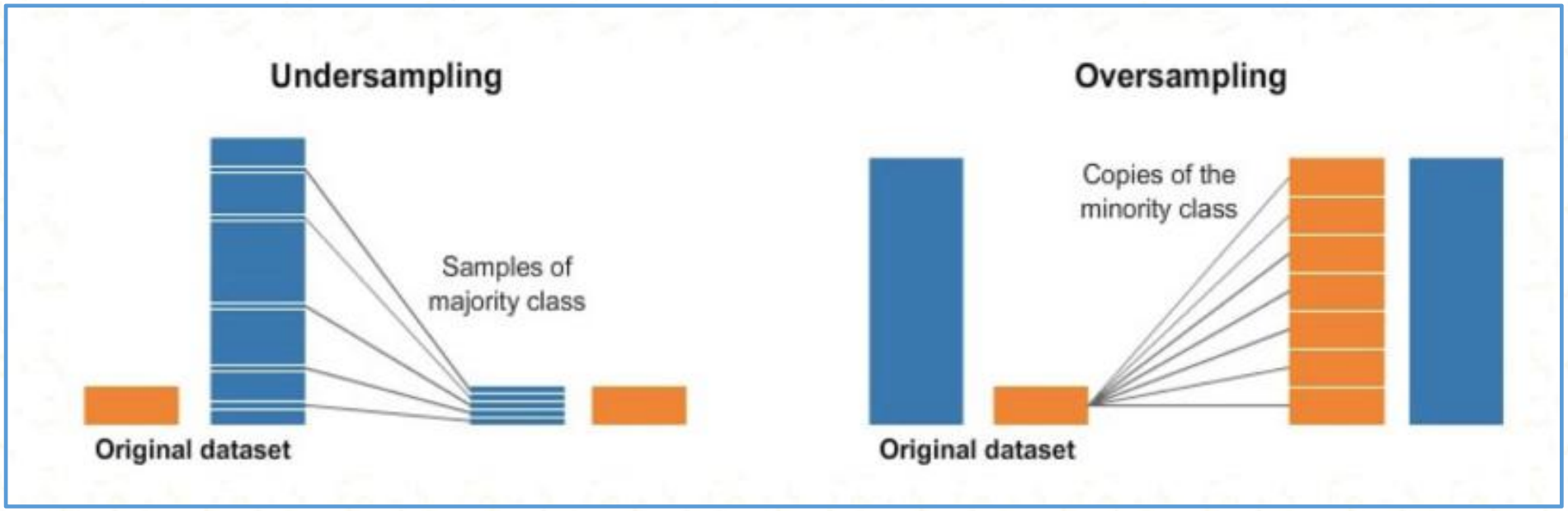
- Imbalanced data mengacu pada masalah klasifikasi di mana jumlah pengamatan per kelas tidak merata.
- https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv





How to handle imbalanced dataset

- Under sampling = Menyeimbangkan distribusi kelas dengan menghilangkan contoh kelas mayoritas secara acak.
- Oversampling = Meningkatkan jumlah instance di kelas minoritas dengan mereplikasinya secara acak



DigitalSkola
Uncover The World Of Digital Skills With Us

Let’s practice

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU