

Exploratory Data Analytics & Data Preparation

Informatics Research Group
School of Electrical Engineering and Informatics
Institut Teknologi Bandung

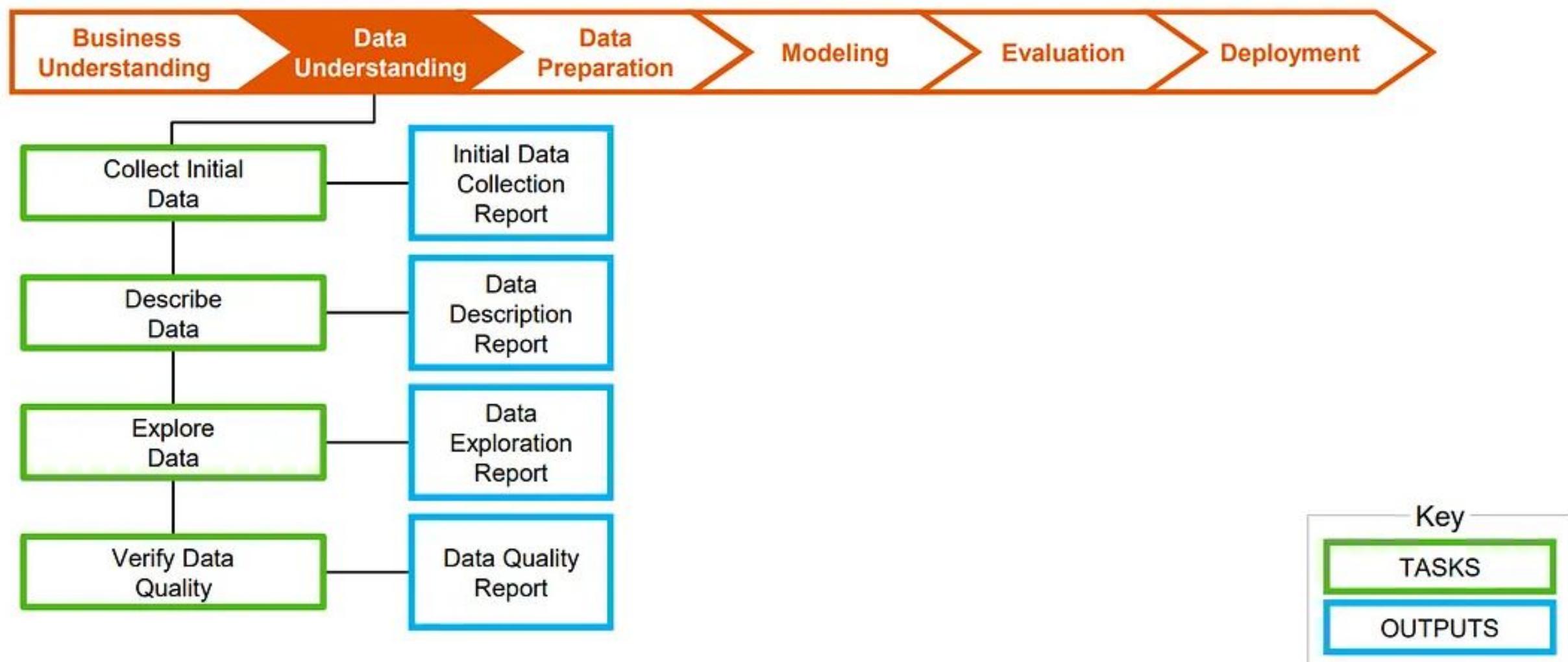
Sources:

Materi Pelatihan Associate Data Scientist – Pusat Artificial Intelligence ITB



Data Understanding Phase – Overview

CRISP-DM – Phase 2: Data Understanding



Sumber: <https://open.sap.com/courses/ds3>

Collect Initial Data

- **Collect Initial Data** or acquire the data and its access to the data listed in the projects resources. Collecting initial data also means you need to have a checklist of the dataset you have acquired, the dataset location, the methods to acquire the datasets, and record any problems encountered and any solutions to the problems for the other users or project members to be aware of.

Describe Data

- **Describe Data** by examining the properties of the data acquired, provide a description report regarding the format of the data, quantity of data and even the records and fields in each table or datasets.

Explore Data

- **Explore Data** by using data science questions that can be quickly answered through querying, visualization, and reporting or summary report. In this stage, you will be able to find your first or initial hypothesis and their impact on the project.

Verify Data Quality

- **Verify Data Quality** by examining if the data is complete. If the data has errors or are there missing values and if there is, what is the percentage of the missing values versus the overall data obtained.



Collect Initial Data

- Menentukan Kebutuhan Data
- Mengambil Data



Menentukan Kebutuhan Data

- Menentukan kebutuhan data adalah proses mengidentifikasi dan mendokumentasikan data yang dibutuhkan untuk melatih model AI yang disesuaikan dengan permasalahan bisnis yang sudah diketahui pada proses sebelumnya.
- Kebutuhan data tergantung pada:
 - Task pemodelan data science yang akan dilakukan (ditentukan pada langkah business understanding)
 - Domain permasalahan dapat menentukan apakah kebutuhan datanya berupa data terstruktur atau tidak terstruktur
 - Kompleksitas permasalahan atau model yang dibangun

Kebutuhan Data berdasarkan Task

- **Klasifikasi :**
Membutuhkan data berlabel dengan labelnya berupa kelas yang akan diprediksi
- **Regresi :**
Membutuhkan data berlabel dengan labelnya berupa nilai sebenarnya dari data
- **Clustering :** Membutuhkan data tidak berlabel



Data Terstruktur

Terdapat dua jenis informasi mengenai data adalah:

1. Informasi yang menjelaskan struktur data, seperti entitas, atribut, dan relasi.
Informasi ini biasanya dinyatakan dalam bentuk grafik seperti *Entity-Relationship Diagrams* (E-RD).
2. Informasi yang menggambarkan aturan atau batasan yang dapat menjaga integritas data.

Biasanya disebut aturan bisnis (*business rules*), batasan-batasan ini harus dituangkan dalam data dictionary/directory (atau *repository*) suatu organisasi.

Dataset

- A collection of separate (related) sets of information that is treated (manipulated) as a single unit by a computer
(Cambridge / Oxford Dictionary)
- Dataset terdiri atas objek data
 - Kolom merepresentasikan variabel (attributes, features, dimensions).
 - Baris merepresentasikan objek data (samples, examples, instances, data points, tuples).
Merepresentasikan entitas.

Contoh Structured Dataset: Automobile Dataset

Automobile Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: From 1985 Ward's Automotive Yearbook



Data Set Characteristics:	Multivariate	Number of Instances:	205
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	26
Associated Tasks:	Regression	Missing Values?	Yes

Predict price based on 25 attributes of automobile data

205 data objects =
205 rows

26 columns

Automobile Dataset:

Attributes (categorical, integer, real)

Attribute: Attribute Range:

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses:continuous from 65 to 256.
3. make:alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.

11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height:continuous from 47.8 to 59.8.
14. curb-weight:continuous from 1488 to 4066.
15. engine-type:dohc, dohc, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders:eight, five, four, six, three, twelve, two.
17. engine-size:continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore:continuous from 2.54 to 3.94.
20. stroke:continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower:continuous from 48 to 288.
23. Peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg:continuous from 16 to 54.
26. Price: continuous from 5118 to 45400.

Contoh Unstructured Dataset: YouTube Spam

YouTube Spam Collection Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: It is a public set of comments collected for spam research. It has five datasets composed by 1,956 real messages extracted from five videos that were among the 10 most viewed on the collection period.

Data Set Characteristics:	Text	Number of Instances:	1956	Area:	Computer
Attribute Characteristics:	N/A	Number of Attributes:	5	Date Donated	2017-03-26
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	63435

YouTube Spam Dataset: Attributes & Data (csv)

COMMENT_ID,AUTHOR,DATE,CONTENT,CLASS

z12pgdhovmrktzm3i23es5d5junftft3f,lekanaVEVO1,2014-07-22T15:27:50,i love this so much. AND also I Generate Free Leads on Auto Pilot & You Can Too! http://www.MyLeaderGate.com/moretraffic,1

z13yx345uxepetggz04ci5rjcxehz1rtf4,Pyunghee,2014-07-27T01:57:16,http://www.billboard.com/articles/columns/pop-shop/6174122/fan-army-f
ace-off-round-3 Vote for SONES please....we're against vips....please help us.. >.<,1

...

z12cdlswetvnejcri04cex0jfwy2u3tzj54,Rafi Hossain,2015-06-05T19:55:08,Honestly speaking except taylor swift and adele i don't like any of the modern day singers. But i must say whenever i hear this song i feel goosebumps. Its quite inspiring!! Thanks miss Perry!,0

z120e5uautvcuper304ccf4bjrjugdpbwrc0k,moaz adnan,2015-06-05T20:01:23,who is going to reach the billion first : katy or taylor ?,0

Pentingnya Menentukan Kebutuhan Data secara Tepat



<https://www.ibm.com/blogs/business-analytics/data-driven-analytics-vision/>

- Kualitas, kuantitas, dan keberagaman data penting dalam pembelajaran mesin.
 - Data berkualitas rendah akan menyebabkan hasil/model berkualitas rendah
- > Business goal yang tidak tercapai



Menentukan Kebutuhan Data Dilakukan Secara iteratif



Melakukan
Pengumpulan Data



Melakukan
Analisis Data



Mengevaluasi
Hasil



Dilakukan Secara **Iteratif**

- Apa yang diperlukan model pembelajaran mesin?
- Bagaimana hasil pemodelan menggunakan data tsb?
- Apakah diperlukan penambahan atau pengumpulan data lagi?



Sumber Data

Download dataset yang ada (open, publik)

- Website kumpulan dataset
- Link pada makalah

Manual labeling

- Anotasi manual oleh annotator manusia

Data observasi

- Log aktifitas
- Rekap data periodik

image	label
	cat
	not cat
	cat
	not cat

user ID	time	price (\$)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:30.15	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes



Metode Pengumpulan Data

1. Pengambilan data secara manual
2. Pengambilan data melalui API, contoh melalui API Kaggle atau Twitter
3. Pengambilan data melalui web scraping
4. Pengambilan data melalui akses langsung ke basis data relasional yang ada.



Pemberian Label Data



Label Data

Label/target/variable dependent : attribute/kolom/field yang menjadi sasaran/target untuk diprediksi.

- Disebut variable dependent, karena nilai dari attribute ini tergantung dari nilai atribut-atribut yang lain.
- Label/target biasanya disimbolkan dengan huruf y .
- y merupakan fungsi dari atribut independent (biasanya disebut x).
 - Jadi fungsi target:
 - persamaan y merupakan fungsi dari x : $y = f(x)$.

Data Berlabel

- Data berlabel (data beranotasi) adalah data yang sudah mengandung label yang bermakna, tag, atau kelas.
- Contoh:
 - Pada sistem pengenalan gambar:
 - Perlu dikumpulkan banyak (ribuan, jutaan) foto yang mengandung gambar target.
 - Gambar diberi label sesuai target, misal: '*orang*', '*pohon*', '*mobil*', dll.
 - Mesin akan belajar mengenali gambar berdasarkan input data pasangan gambar dan namanya

Contoh Data Berlabel



Computer Vision:

- Label pada gambar, piksel, atau *key point*, batas gambar digital.
 - Misal: produk vs. gaya hidup; objek wajah vs non wajah, objek hewan vs non hewan



Pemrosesan Bahasa Alami

- sentimen atau makna uraian teks
- Identifikasi bagian ucapan
- klasifikasikan kata benda
- Identifikasi teks, gambar, PDF, atau file lainnya.



Pemrosesan Audio

- Rangkaian teks dari audio.



Cara Pelabelan Data

Dua cara pelabelan data:

1. Pelabelan Manual

- Dilakukan oleh manusia
 - Pakar domain bisnis yang menjadi tujuan pemodelan
 - misal dokter radiologi melabeli foto thorax adanya infeksi paru-paru
 - Orang awal/biasa
- *Strong labelled*

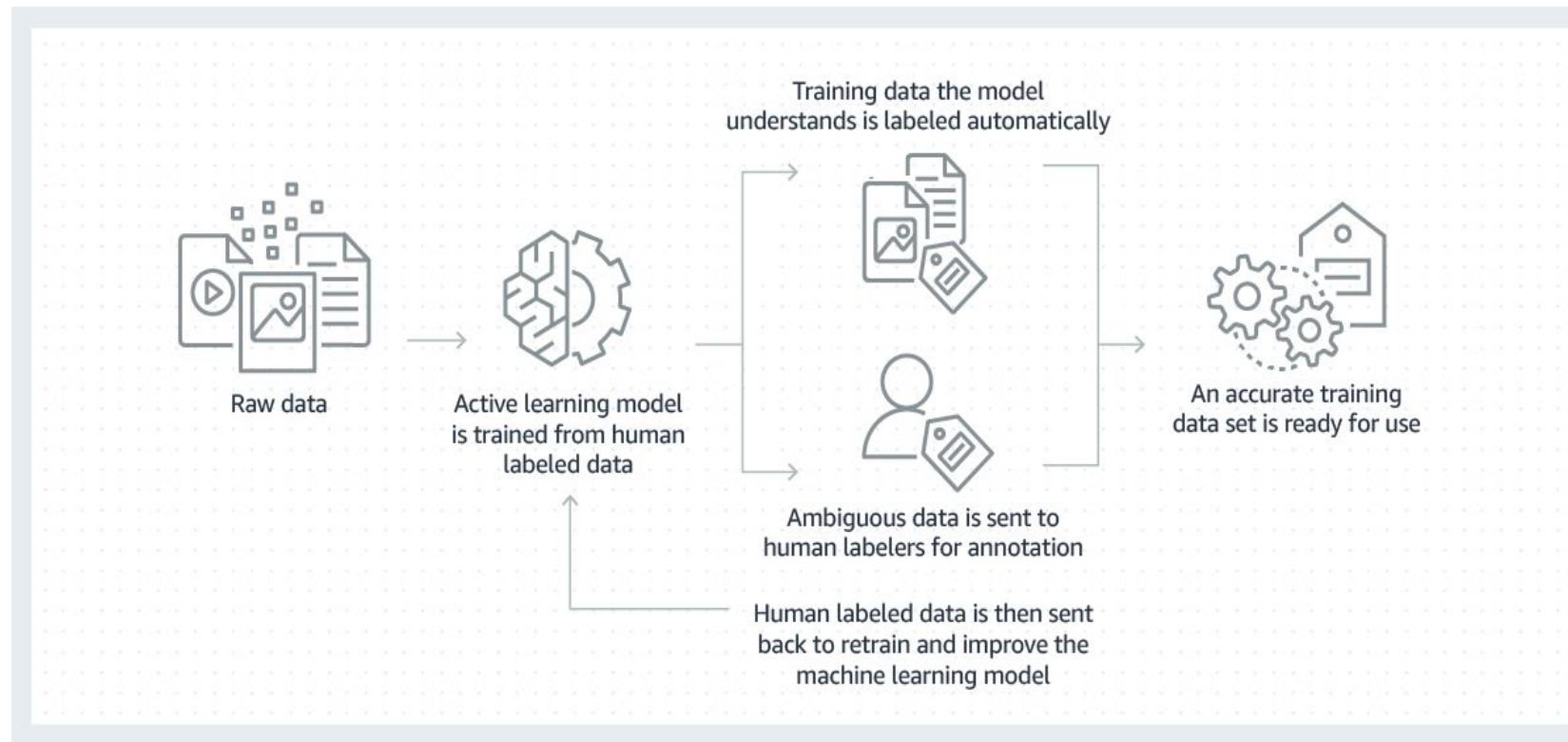
2. Pelabelan Otomatis:

- **Pelabelan sintetis (weakly labelled) :**
 - Data sintetis dihasilkan oleh model generatif yang dilatih dan divalidasi pada dataset asli (pelabelan oleh mesin)
 - Dilakukan jika pelabelan manual mahal
 - Diperlukan teknik pemodelan data yang sesuai untuk data weakly labelled (model yang robust)
- **Pemrograman data :**
 - Penulisan fungsi pelabelan — skrip yang secara terprogram melabeli data.
 - Contoh: Data transaksi fraud atau tidak yang sudah dilabeli oleh kode program



Pelabelan Data Menggunakan Pembelajaran Mesin

Pelabelan dapat dibuat lebih efisien dengan menggunakan model pembelajaran mesin untuk melabeli data secara otomatis.



Pendekatan Pelabelan Data

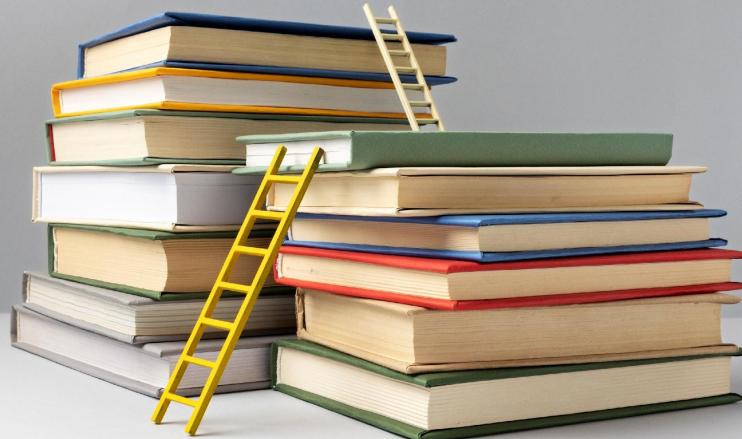
- **Inhouse Labeling** : Pelabelan dilakukan secara internal pengguna data latih
- **Crowdsourcing** : Penggunaan mekanisme Platform Crowdsourcing
- **Outsourcing** :
 - **ke individu** : Menggunakan pekerja lepas di berbagai situs web rekrutmen, pekerja lepas, dan jejaring sosial
 - **ke perusahaan** : Penggunaan jasa perusahaan outsourcing yang mengkhususkan diri dalam persiapan data pelatihan



Pertimbangan dalam Pelabelan Data

- Akurasi pelabelan data
- Jumlah data
- Biaya
- Waktu
- Teknik pemodelan yang digunakan (*weakly vs strongly labelled*)





Pemahaman Data

(describe data, explore data, verify data quality)



Tujuan Pemahaman Data

- Memahami isi data
- Menguji kualitas data
- Memastikan jumlah data
- Menemukan insight awal

Proses Pemahaman Data

1. Mendeskripsikan data
2. Melakukan analisis untuk memastikan kualitas data
3. Melakukan eksplorasi data





1. Mendeskripsikan Data

Proses Pemahaman Data: Mendeskripsikan Data

- Karakteristik Utama yang dideskripsikan:
 - a. **Sumber data**
 - Basis data lembaga
 - Artikel berita, dll
 - b. **Ukuran**
 - Jumlah instan
 - Jumlah atribut dan apa saja atributnya
 - c. **Mengidentifikasi Tipe Data dari Atribut Data**
 - Tipe atribut
 - Nilai atribut: *value range* (untuk numerik) atau value set (untuk nominal)
 - d. **Statistik Deskriptif: Perhitungan statistik dasar (mean, variance, dll)**



Mengidentifikasi Tipe Data dari setiap Atribut/Fitur

Tujuan mengidentifikasi tipe data dari fitur:

- Menjelaskan karakteristik sebuah fitur
- Menetapkan analisis statistik yang tepat untuk fitur tersebut.

Tipe Data sebuah fitur (Stanley Smith Stevens, 1940):

- Nominal
- Ordinal
- Interval
- Rasio



Tipe Data Fitur (Skala Pengukuran)

Kategori	Type Data	Karakteristik	Contoh Fitur
Kualitatif	Nominal	Tidak ada urutan dari kategori.	<ul style="list-style-type: none">• Jenis kelamin• Golongan darah• Agama/ Kepercayaan
	Ordinal	Kategori memiliki urutan tetapi jarak antar data tidak bermakna.	<ul style="list-style-type: none">• Nilai (A-E)• Status Sosial-Ekonomi• Tingkat pendidikan• Tingkat persetujuan• Skala Likert.
Kuantitatif	Interval	Memiliki urutan, jarak antara dua nilai bermakna, tetapi tidak memiliki nilai “nol” (nilai terkecil).	<ul style="list-style-type: none">• Suhu• Tahun• Skor test IQ• Umur
	Ratio	Memiliki semua karakteristik data interval dan memiliki nilai “nol”.	<ul style="list-style-type: none">• Tinggi objek• Panjang benda• Berat objek



Contoh Data Task Klasifikasi

Contoh kasus:

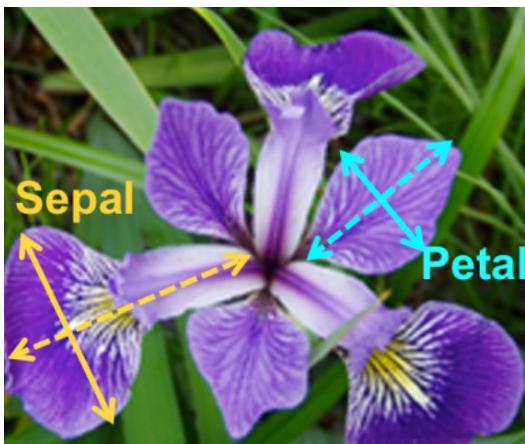
Memprediksi spesies bunga iris (kolom label: species)

apakah:

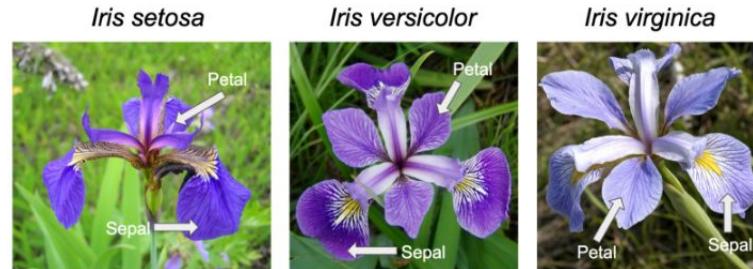
- Iris-setosa,
- Iris-versicolor, atau
- Iris-virginica

berdasarkan:

- Panjang sepal (SepalLengthCm)
- Lebar sepal (SepalWidthCm)
- Panjang daun bunga (PetalLengthCm)
- Lebar daun bunga (PetalWidthCm)



Pada task **klasifikasi**, label merupakan variabel **kategorikal**



SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa

x = variabel independen

y = variabel dependen / target / label



Contoh Data Task Regresi

Contoh kasus:

Memprediksi harga rumah (kolom label: price)

berdasarkan:

- luas (area)
- jumlah kamar tidur (bedrooms)
- jumlah kamar mandi (bathrooms)
- berapa kali rumah itu ditinggali sebelumnya (stories)
- apakah terletak di jalan raya atau tidak (mainroad)
- apakah memiliki ruang tamu atau tidak (guestroom)

Pada task regresi, label merupakan variabel kontinu

- apakah memiliki basement atau tidak (basement)
- apakah memiliki pemanas air atau tidak (hotwaterheating)
- apakah memiliki AC atau tidak (airconditioning)
- apakah memiliki tempat parkir atau tidak (parking)
- apakah berada di daerah yang favorit atau tidak (prefarea)
- apakah memiliki perabot lengkap / ada beberapa perabot / tidak ada perabot (furnishingstatus)

area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus	price
7420	4	2	3	yes	no	no	no	yes	2	yes	furnished	13300000
8960	4	4	4	yes	no	no	no	yes	3	no	furnished	12250000
9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished	12250000
7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished	12215000
7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished	11410000
7500	3	3	1	yes	no	yes	no	yes	2	yes	semi-furnished	10850000
8580	4	3	4	yes	no	no	no	yes	2	yes	semi-furnished	10150000
16200	5	3	2	yes	no	no	no	no	0	no	unfurnished	10150000
8100	4	1	2	yes	yes	yes	no	yes	2	yes	furnished	9870000

x = variabel independen

y = variabel dependen / target / label



Contoh Data Task Clustering

Contoh kasus:

Mengelompokkan pelanggan ke dalam beberapa kelompok

berdasarkan:

- jenis kelamin (Gender)
- umur (Age)
- pendapatan tahunan (Annual Income (\$))

Pada clustering, data tidak berlabel

- skor pengeluaran (Spending Score (1-100))
- pekerjaan (Profession)
- lama bekerja (Work Experience)
- Banyak anggota keluarga (Family Size)

Gender	Age	Annual Income (\$)	Spending Score (1-100)	Profession	Work Experience	Family Size
Male	19	15000	39	Healthcare	1	4
Male	21	35000	81	Engineer	3	3
Female	20	86000	6	Engineer	1	1
Female	23	59000	77	Lawyer	0	2
Female	31	38000	40	Entertainment	2	6
Female	22	58000	76	Artist	0	2
Female	35	31000	6	Healthcare	1	3
Female	23	84000	94	Healthcare	1	3
Male	64	97000	3	Engineer	0	3

x = variabel independen

Statistik Deskriptif

- **Rangkuman informasi atau karakteristik dari sejumlah data.**
- **Ukuran Pemusatan:** ukuran yang menjelaskan titik pusat data
 - Mean (\bar{x}) = $\sum_{i=1}^n (x_i/n)$
 - Kuartil ke-1 (Q_1) adalah nilai data dimana 25 % dari data setelah disortir menaik bernilai lebih kecil dari nilai tersebut.
 - Kuartil ke-2 atau Median (Q_2) adalah nilai data dimana separuh dari data setelah disortir menaik bernilai lebih kecil dari nilai tersebut.
 - Kuartil ke-3 (Q_3) adalah nilai data dimana 75 % dari data setelah disortir menaik bernilai lebih kecil dari nilai tersebut.
 - Mode (Modus) adalah nilai yang paling sering muncul pada sekumpulan data
- **Ukuran Variabilitas:** ukuran variabilitas data
 - Varians atau variance (s^2) = $\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$
 - Standar deviasi (s) = $\sqrt{s^2}$
 - Kisaran atau range = $x_{max} - x_{min}$

- Alat yang sangat penting dalam dunia pengumpulan, penyajian, dan interpretasi data
- Digunakan untuk mengenali pola, tren, dan informasi yang terkandung dalam data.



Statistik Deskriptif

Jenis data	Ukuran Pemusatan	Ukuran Variabilitas	Sebaran Data
Nominal	Modus	--	Frekuensi nilai, Proporsi
Ordinal	Median, Modus	--	Frekuensi nilai, Proporsi
Interval	Mean, Median, Mode	Variance, Kuartil, Persentil, Range	Frekuensi interval nilai
Ratio	Mean, Median, Mode	Variance, Kuartil, Persentil, Range	Frekuensi interval nilai

Contoh Mendeskripsikan Data

Automobile Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: From 1985 Ward's Automotive Yearbook



Data Set Characteristics:	Multivariate	Number of Instances:	205
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	26
Associated Tasks:	Regression	Missing Values?	Yes

205 data objects = 205 rows

26 columns

Predict price based on 25 attributes of automobile data

<http://archive.ics.uci.edu/ml/datasets/Automobile>





Contoh Mendeskripsikan Data

Automobile Dataset: Attributes

Attribute: Attribute Range:

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses:continuous from 65 to 256.
3. make:alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.

11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height:continuous from 47.8 to 59.8.
14. curb-weight:continuous from 1488 to 4066.
15. engine-type:dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders:eight, five, four, six, three, twelve, two.
17. engine-size:continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore:continuous from 2.54 to 3.94.
20. stroke:continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower:continuous from 48 to 288.
23. Peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg:continuous from 16 to 54.
26. Price: continuous from 5118 to 45400.





Menganalisis dan Mengeksplorasi Data

Analisis Data

Analisis yang dilakukan untuk memeriksa data dan menentukan bagaimana data harus diproses sebelum digunakan untuk pemodelan

Tujuan analisis data:

Memeriksa kesiapan data untuk digunakan pada pemodelan, baik dari sisi kelengkapannya maupun kualitasnya



Pemeriksaan Kelengkapan Data

- Apakah semua data yang diperlukan sudah ada?
- Apakah semua variabel/kolom yang diperlukan sudah ada?
- Apakah data yang ada sudah mencakup periode waktu yang diperlukan?
- Apakah jumlah instans sudah mencukupi
- Apakah sudah mencakup sebanyak mungkin kasus dunia nyata?
- Apakah jumlah data untuk setiap kelompok label sudah berimbang?



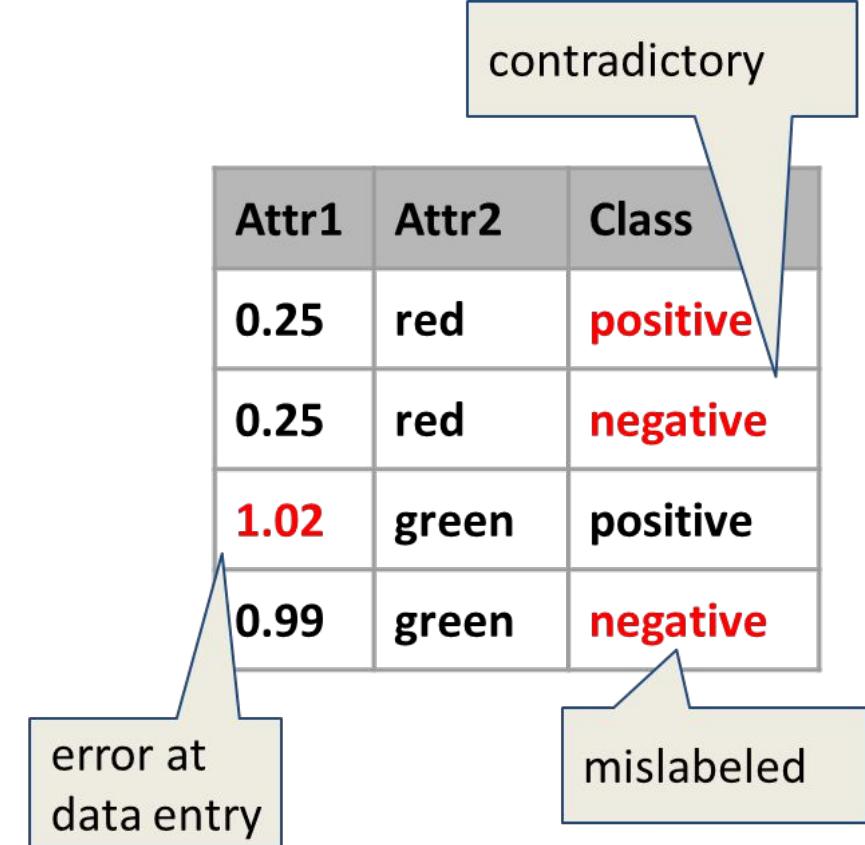
Pemeriksaan Kualitas Data

- Apakah ada inkonsistensi data?
 - Apakah ada klasifikasi atau pengkodean yang tidak dapat dijelaskan?
 - Apakah ada masalah dalam format seperti tanggal yang tidak biasa, karakter ASCII?
- Apakah ada item yang tidak lengkap atau hilang (*missing values*)?
- Apakah ada duplikat data?
- Apakah ada nilai yang tidak biasa atau penciran yang jelas?
- Seperti apa hubungan antar atribut data dan hubungan antara setiap atribut dengan labelnya? (ada korelasi atau tidak)



Noisy Data and Inconsistent Data

- Tipe data kotor:
 - Class (label) noise
 - Sampel yang kontradiksi
 - Sampel yang salah label
 - Attribute noise
 - Kesalahan entri data
 - Pelanggaran terhadap konstrain
 - Pengulangan white space (karakter yang tidak terbaca atau tidak terlihat)
 - Perbedaan unit pengukuran (contoh: centimeter dengan meter)
 - Perbedaan level agregasi (contoh: akumulasi per hari dengan per minggu)



The diagram illustrates three types of noisy data using a table:

Attr1	Attr2	Class
0.25	red	positive
0.25	red	negative
1.02	green	positive
0.99	green	negative

Annotations point to specific rows:

- A callout labeled "contradictory" points to the second row where Attr1 is 0.25 and Attr2 is red, but the Class is negative.
- A callout labeled "error at data entry" points to the fourth row where Attr1 is 1.02 and Attr2 is green, but the Class is positive.
- A callout labeled "mislabeled" points to the fifth row where Attr1 is 0.99 and Attr2 is green, but the Class is negative.

<https://sci2s.ugr.es/noisydata>



Temukan Data yang Salah!

ID	A1	A2	A3	A4	C
1	1	2	1	1	y1
2	2	1	1	1	y2
3	3	5	2	4	y1
4	2	2	2	2	y2
5	3	5	2	4	y2
6	6	7	5	8	y1
7	9	4	3	1	y1

<https://pdfs.semanticscholar.org/ee88/ae92d19e208f15d3613016266bb13da8fc98.pdf>

Temukan Data yang Salah!

	A	B	C	D	E	F	G
1	First name	Last name	January	February	March	Q1 Sales	Region
2	Darrel	Alston	\$ 11,896	\$ 2,552	\$ 11,350	\$ 25,798	East
3	David	Terrell	\$ 9,763	\$ 1,749	\$ 8,678	\$ 20,190	South
4	Gwendolyn	Cameron	\$ 9,421	\$ 5,585	\$ 10,423	\$ 25,429	East
5	Katell	Hall	\$ 3,291	\$ 2,610	\$ 13,692	\$ 19,593	Norht
6	Honorato	Howard	\$ 11,746	\$ 6,756	\$ 10,471	\$ 28,973	West
7	Nehru	Rose	\$ 8,603	\$ 5,907	\$ 1,682	\$ 16,192	West
8	Upton	Shields	\$ 10,955	\$ 4,914	\$ 11,539	\$ 27,408	West
9	Germane	Holman	\$ 11,561	\$ 8,547	\$ 8,433	\$ 28,541	North
10	Elliott	Hall	\$ 9,318	\$ 5,857	\$ 4,935	\$ 20,110	North
11	Illana	Erickson	\$ 3,709	\$ 13,401	\$ 3,431	\$ 20,541	Wst
12	Lani	Spears	\$ 5,620	\$ 14,252	\$ 8,894	\$ 28,766	East
13	Clementine	Pope	\$ 8,901	\$ 10,143	\$ 13,573	\$ 32,617	South

<https://edu.gcfglobal.org/en/excel-tips/a-trick-for-finding-inconsistent-data/1/>



Missing values Data

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.233	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

<https://freecontent.manning.com/pre-processing-data-for-modeling/>



Data yang Tidak Konsisten

Data yang tidak konsisten: mengandung perbedaan dalam nama atau kode, atau perbedaan antara instan yang sama yang biasanya terjadi ketika data dikumpulkan dari berbagai sumber

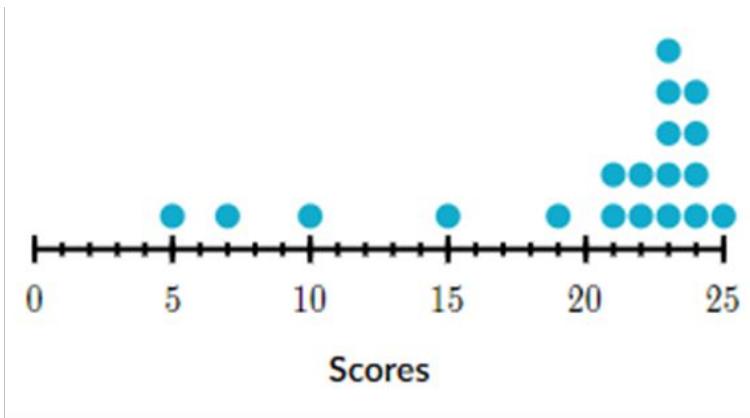
Age	BirthDate	...
18	30 June 2000	...
19	30 June 2000	...
...

ID	GPA	...
01	3.25	...
01	3.67	...
...

ID	Rating	...
1	A	...
2	B	...
3	1	...
4	3.5	...



Identifikasi Pencilan



19 Data:

5, 7, 10, 15, 19, 21, 21,
22, 22, 23, 23, 23, 23, 23,
24, 24, 24, 24, 25

Aturan Umum:

Low outlier $< Q1 - 1.5 * IQR$

High outlier $> Q3 + 1.5 * IQR$

$IQR = Q3 - Q1$

$Q1: 19 ; Q3: 24$

$IQR = Q3 - Q1 = 24 - 19 = 5$

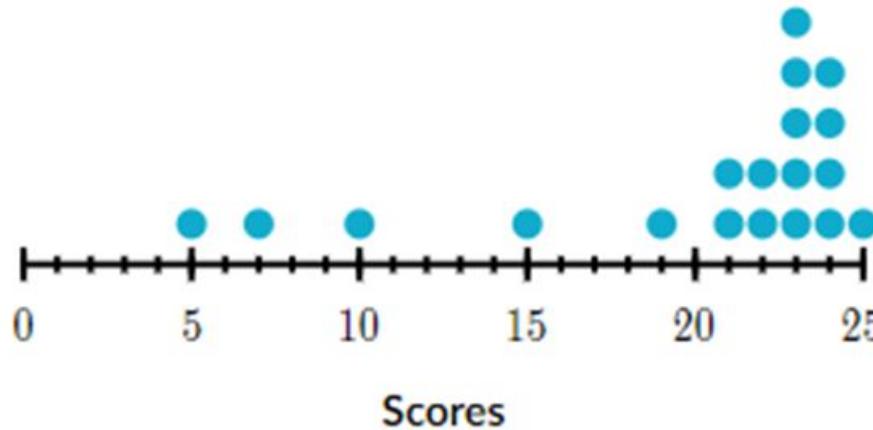
$\text{Low outlier} < 19 - (1,5 * 5)$

$\text{High outlier} > 24 + (1,5 * 5)$

<https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule>



Identifikasi Penculan menggunakan Visualisasi: Whisker Plot

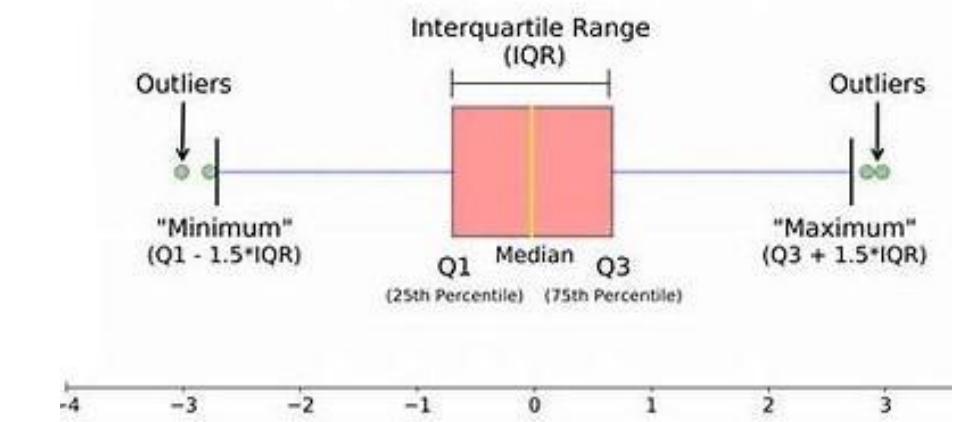


Median: 23; Q1: 19 ; Q3: 24

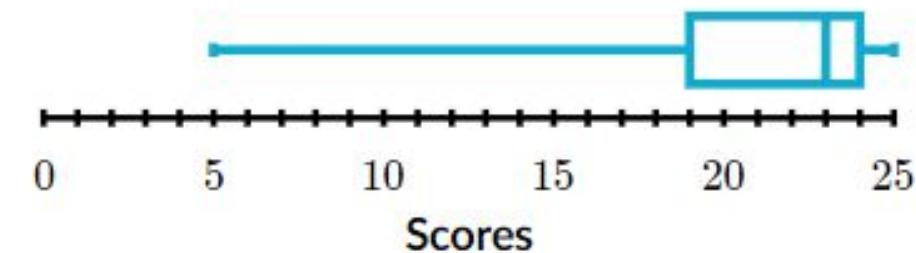
$$\text{IQR} = \text{Q3}-\text{Q1}=24-19=5$$

$$\text{Min} = 19-7.5=11.5$$

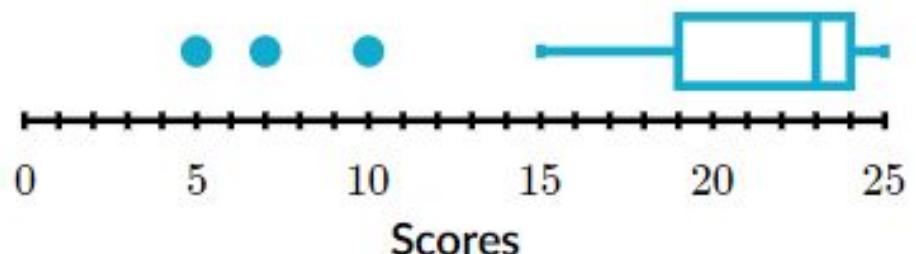
$$\text{Max} = 24+7.5=31.5$$



Whisker Plot tanpa Penculan



Whisker Plot dengan Penculan



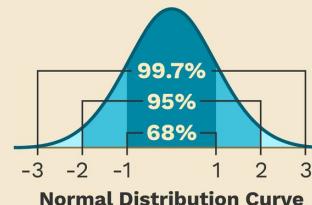
Identifikasi Penculan Menggunakan STD Fitur

- STD (*standard deviation*) = $\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$,
dimana: x_i adalah nilai fitur ke- i dan n adalah banyaknya sampel data
- Sebuah data x dikategorikan sebagai *outlier* apabila: $x < (\bar{x} - 3 \text{ STD})$ atau $x > (\bar{x} + 3 \text{ STD})$

Calculating Standard Deviation

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

n = The number of data points
X_i = Each of the values of the data
X̄ = The mean of X_i



A normal distribution curve centered at zero. Three vertical lines extend from the center to the curve, creating three nested regions. The innermost region is shaded blue and labeled '68%'. The middle region is also shaded blue and labeled '95%'. The outermost region is shaded blue and labeled '99.7%'. The x-axis is labeled 'Normal Distribution Curve' and has tick marks at -3, -2, -1, 1, 2, and 3.

ThoughtCo.



Data yang Tidak Berimbang

Class	Frequency
A	1000
B	10

Majority class

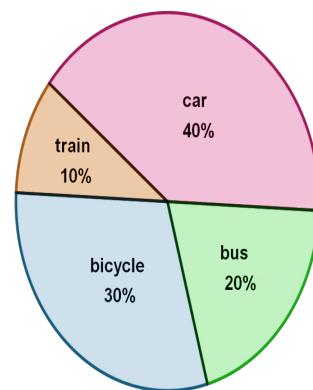
Minority class

Eksplorasi Data Menggunakan Grafik

Diagram Univariat

Contoh Pie Chart:

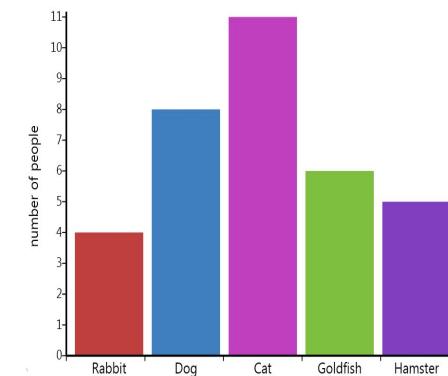
Pengguna Mode Transportasi di DKI Jakarta,
1 Agustus 2022



Tipe data: Nominal, Ordinal

Contoh Bar Chart:

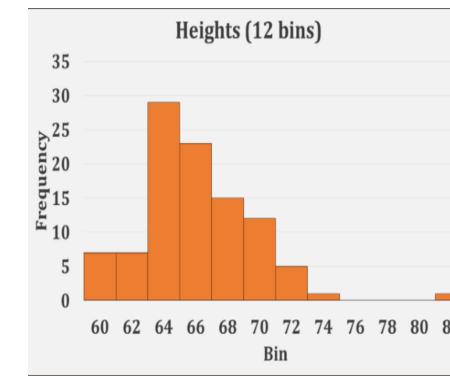
Sebaran Pemilik Hewan di Kota Bogor,
1 Agustus 2022



Tipe data: Nominal, Ordinal

Contoh Histogram:

Sebaran Usia Mahasiswa Univ. XYZ,
1 Agustus 2022



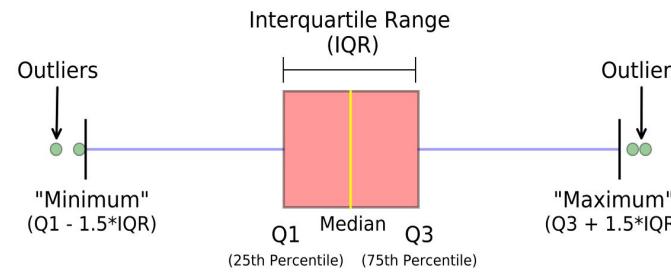
Tipe data: Interval, Rasio



Eksplorasi Data Menggunakan Grafik

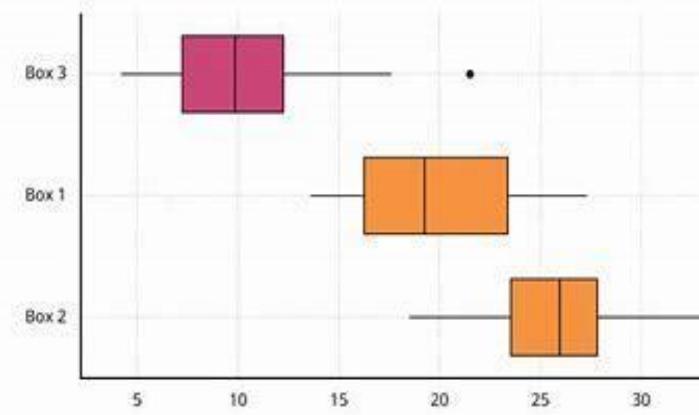
Diagram Bi/Multivariat

Contoh Box Plot 1 Fitur:



Tipe data: Interval, Rasio

Contoh Box Plot 3 Fitur:

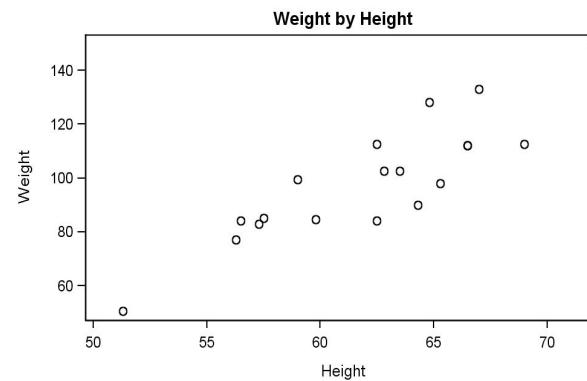


Eksplorasi Data Menggunakan Grafik

Diagram Bi/Multivariat

Contoh Scatter Plot:

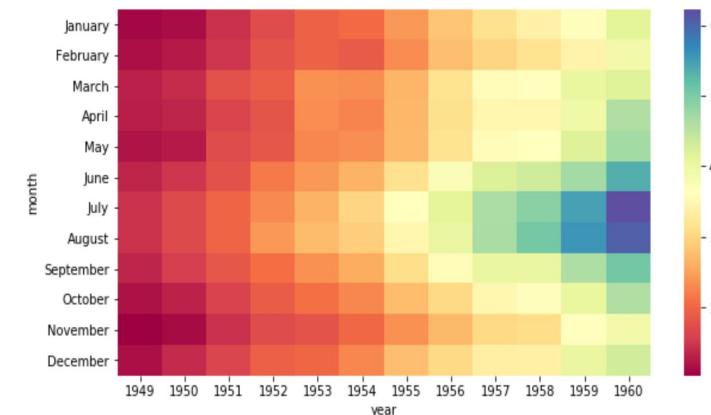
Sebaran Tinggi Badan dan Berat Badan dari Sampel Mahasiswa Univ. XYZ,
1 Agustus 2022



Tipe data: Interval, Rasio

Contoh Heatmap

Sebaran Suhu Udara di Kota X
1949-1960



Tipe data: Interval, Rasio





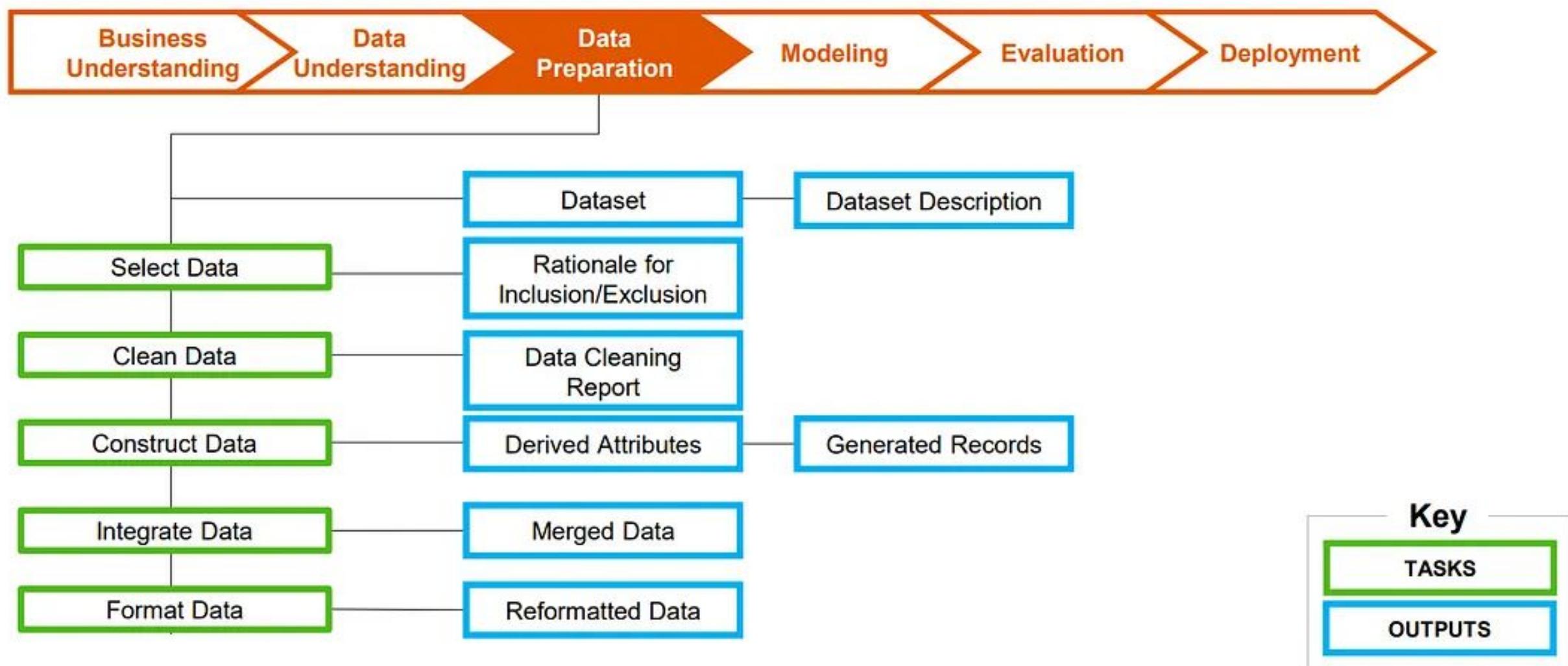
Membuat Laporan Telaah Data

Laporan Telaah Data

- Dibuat setelah selesai mendeskripsikan, menganalisis, dan mengeksplorasi data
- Laporan berisi:
 - Deskripsi data
 - Deskripsi kualitas data
 - Deskripsi kuantitas data
- Laporan telaah data akan digunakan pada tahap berikutnya, yaitu persiapan data untuk pemodelan untuk menentukan langkah apa saja yang perlu dilakukan di tahap persiapan data.

Data Preparation Phase – Overview

CRISP-DM – Phase 3: Data Preparation



Persiapan Data

- Proses dimana data yang sesuai dikumpulkan, dipilih, dibersihkan, dan diorganisir sesuai dengan kebutuhan bisnis untuk digunakan pada tahap pemodelan.
- Dilakukan setelah tahap pemahaman data.
 - Laporan hasil pemahaman data digunakan sebagai dasar untuk menentukan aksi apa yang harus dilakukan pada tahap ini.



Select Data

Select data or decide on the data to be used for analysis. One of the criteria in selecting the data is that it should be relevant to the data science goal that was identified in the business understanding phase. In selecting data, you also need to list the data to be excluded and included and the reasons for these decisions

Clean data

Clean data by raising the data quality to the level required by the selected analysis techniques. Here, you also need to describe what decisions and actions were taken to address the data quality problems reported during the Verify Data Quality task of the Data Understanding Phase

Construct data

Construct data by including derived attributes, entire new records, or transformed values for existing attributes. This may be conducting encoding methods especially for categorical variables or feature engineering

Integrate data

Integrate data by combining from multiple tables or records to create records or values. SQL knowledge and skill is very important and would come in handy in this part

Format data

Format data by transforming the data but not necessarily change its meaning but might be required by the modeling tool. An example would be transforming your data either by standardization or normalization



Tujuan Persiapan Data

- Meningkatkan kualitas data
- Memudahkan pemodelan

Proses dalam Persiapan Data

1. Pemilihan Data

- a. *Record selection*
- b. *Feature selection*

2. Perbaikan Data

- a. Mengisi *missing values*
- b. Perbaikan error
- c. Penanganan *outlier*
- d. Penghapusan duplikasi

3. Konstruksi Data

- a. Reduksi data
- b. Mengubah representasi data
- c. Encoding

4. Integrasi Data

- a. *Data Join*
- b. *Append*





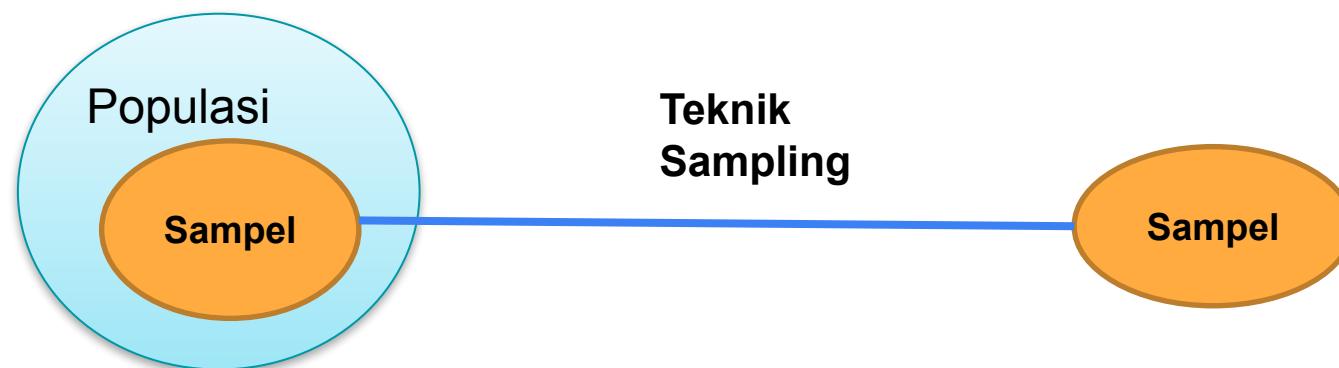
Pemilihan Data

- a. Record selection
- b. Feature selection



Record Selection (Sampling)

- **Sampling** adalah proses dalam analisis statistik dimana peneliti mengambil sejumlah **pengamatan** yang telah ditentukan sebelumnya dari **populasi** yang lebih **besar**.



- Pengambilan sampel memungkinkan peneliti untuk melakukan studi tentang kelompok besar dengan menggunakan sebagian **kecil** dari populasi.



Kategori Metode Sampling

- **Probability Sampling:**

Teknik pengambilan sampel dimana sampel dari populasi yang lebih besar dipilih dengan menggunakan metode berdasarkan teori probabilitas.

- **Non-Probability Sampling**

Teknik pengambilan sampel dimana peneliti memilih sampel berdasarkan penilaian subjektif, bukan pemilihan acak.





Feature Selection

Feature Selection

- Pemilihan fitur adalah proses pengurangan jumlah variabel masukan saat mengembangkan model machine learning.

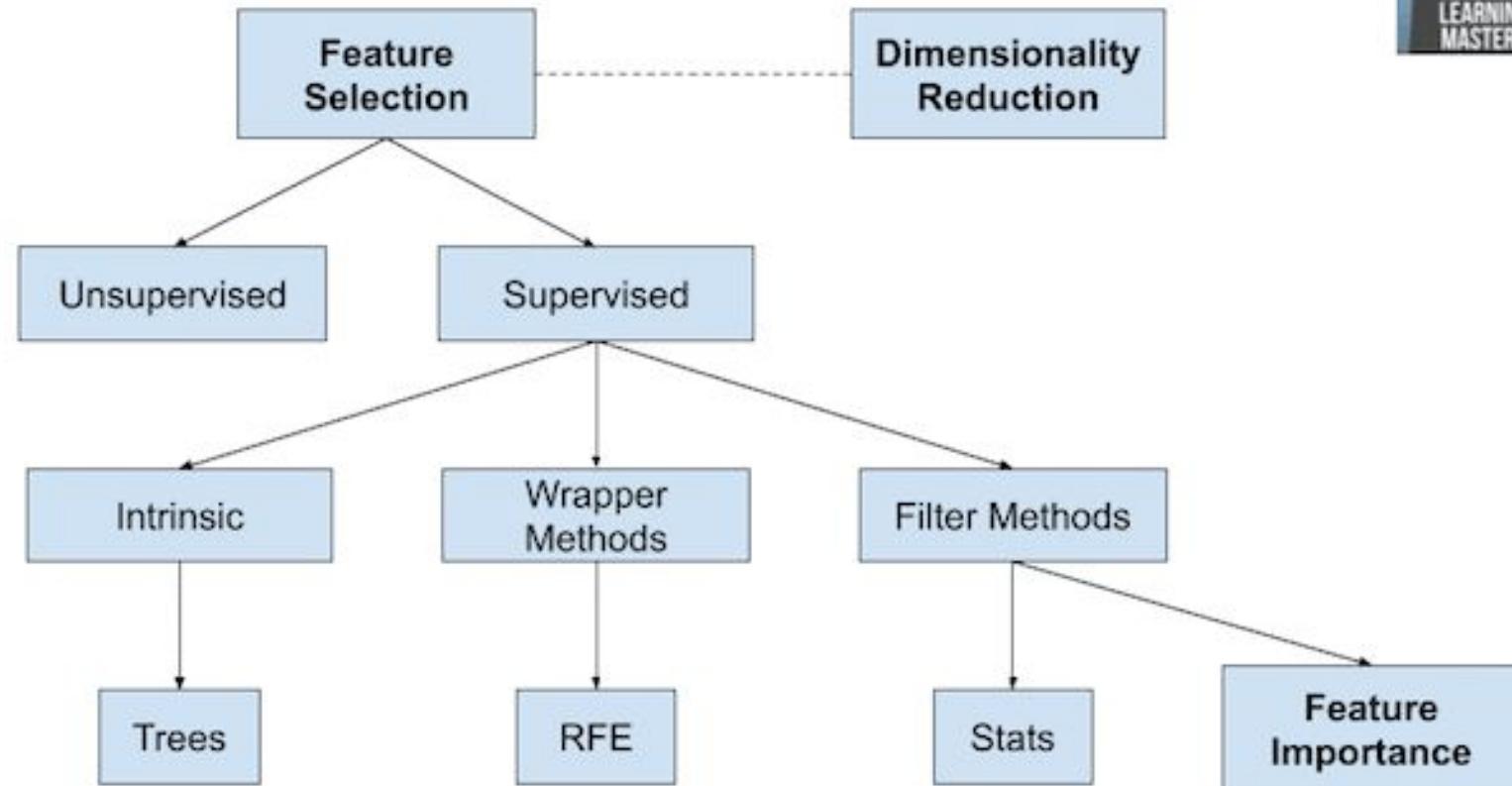


Tujuan Feature Selection

- Mengurangi jumlah variabel masukan untuk mengurangi biaya komputasi pemodelan
- Dalam beberapa kasus, ditujukan untuk meningkatkan kinerja model.

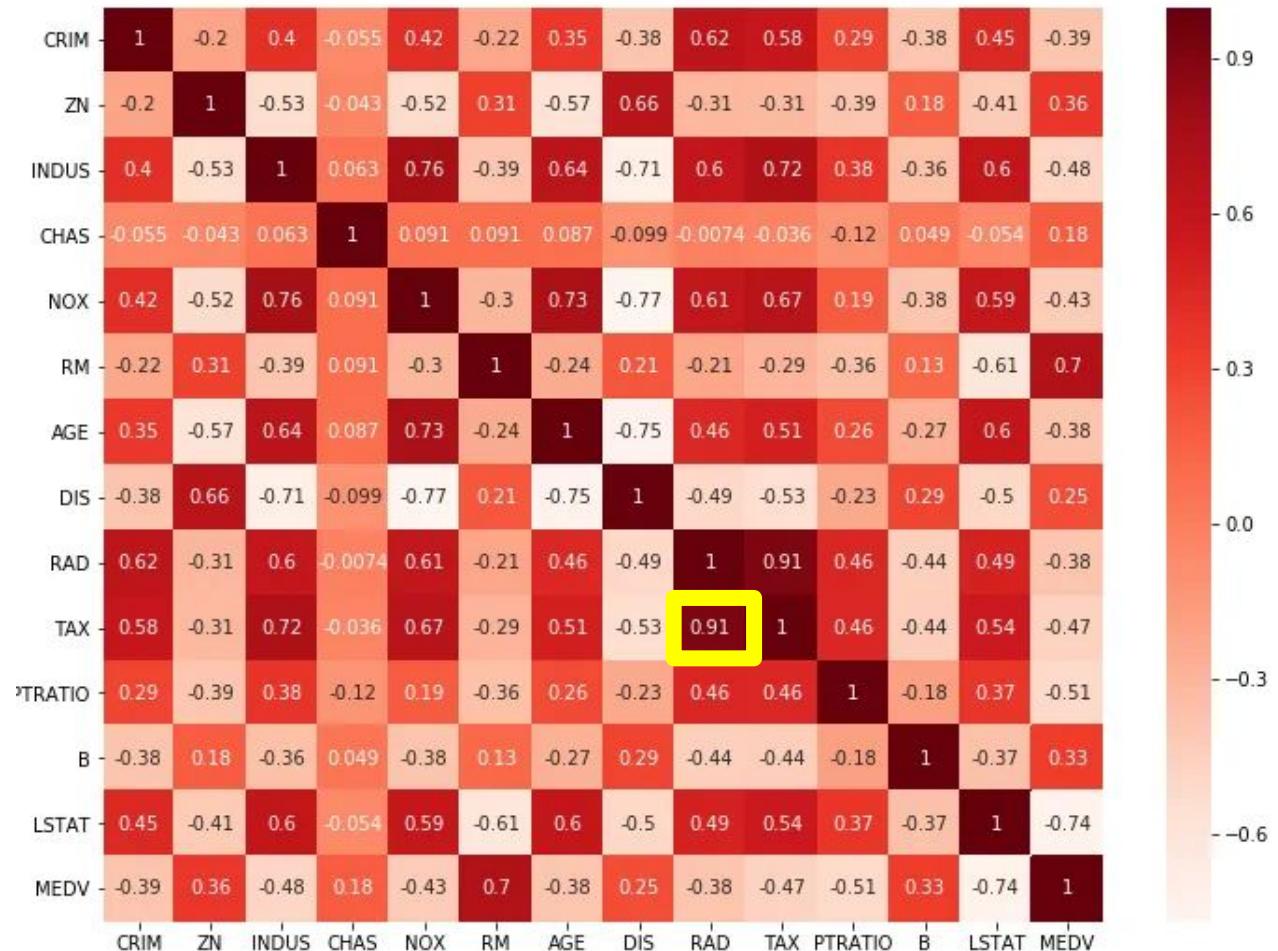
Teknik Feature Selection

Overview of Feature Selection Techniques



Teknik Feature Selection berbasis Unsupervised

- Tidak melibatkan variable target
- Jika ada dua fitur yang memiliki korelasi yang kuat, hilangkan salah satu fitur



Teknik Feature Selection berbasis Filter

- Mengevaluasi hubungan antara setiap variabel masukan dan variabel target (kelas) menggunakan:
 - statistik
 - memilih fitur yang penting (*feature importance*)
- Memilih variabel masukan yang memiliki hubungan paling kuat dengan variabel target.



Teknik Statistik untuk Feature Selection berbasis Filter

- Memilih variabel masukan yang memiliki hubungan yang paling kuat dengan variabel target secara statistik.
- Pilihan ukuran statistik bergantung pada tipe data variabel masukan dan keluaran.

Tipe Variable Masukan	Tipe Variable Target	Teknik Statistik
Numerik	Numerik	<ul style="list-style-type: none">• Pearson's correlation coefficient (linear)• Spearman's rank coefficient (nonlinear)
Numerik	Categorical	<ul style="list-style-type: none">• ANOVA correlation coefficient (linear)• Kendall's rank coefficient (nonlinear)
Categorical	Numerik	<ul style="list-style-type: none">• ANOVA correlation coefficient (nonlinear)• Kendall's rank coefficient (linear)
Categorical	Categorical	<ul style="list-style-type: none">• Chi-Squared test (contingency tables)• Mutual Information



Proses dalam Persiapan Data (review)

1. Pemilihan Data

- a. *Record selection*
- b. *Feature selection*

2. Perbaikan Data

- a. Mengisi *missing values*
- b. Perbaikan error
- c. Penanganan *outlier*
- d. Penghapusan duplikasi

3. Konstruksi Data

- a. Reduksi data
- b. Mengubah representasi data
- c. Encoding

4. Integrasi Data

- a. *Data Join*
- b. *Append*



Hukum Persiapan Data

Data Preparation Law (Data Mining Law 3)

Data preparation is more than half of every data mining process

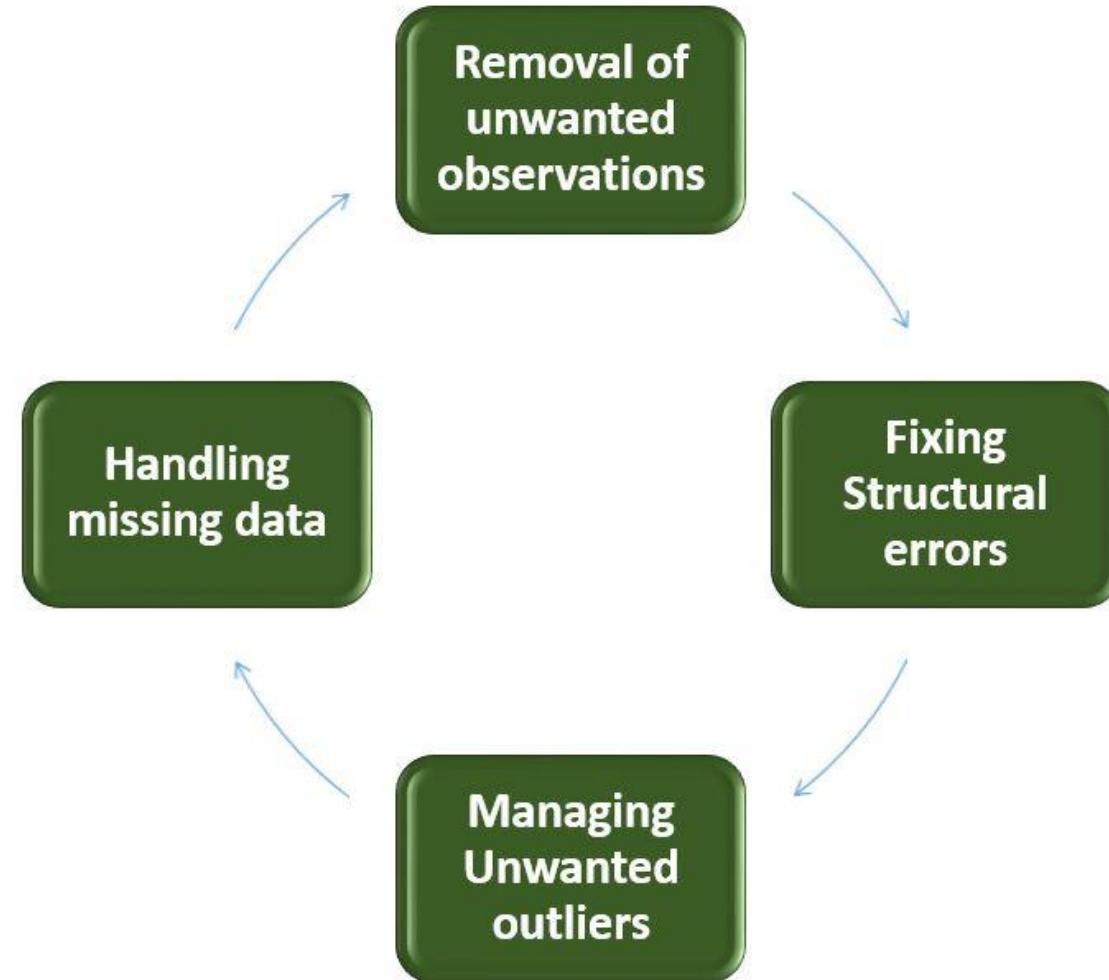
- Maxim of data mining: most of the effort in a data mining project is spent in data acquisition and preparation, and informal estimates vary from 50 to 80 percent





Proses Pembersihan Data

Langkah-langkah Pembersihan Data



Jenis Kesalahan Data dan Alternatif Cara Mengatasinya

1. Kesalahan nilai fitur di dalam sebuah dataset (1)

Jenis Error

- Kesalahan selama proses data entry
- Nilai fitur yang meragukan/tidak mungkin (*impossible values*)
- Pengulangan white space (*unreadable or undetected characters*)

Tindakan Mengatasinya

- Meningkatkan kapasitas staf data entry
- Menggunakan dukungan software untuk memvalidasi data.
- Perbaikan Data
- Menggunakan software untuk menghilangkan unreadable atau undetected characters dari data input.



Jenis Kesalahan Data dan Alternatif Cara Mengatasinya

1. Kesalahan nilai fitur di dalam sebuah dataset (2)

Jenis Error

- Tidak ada nilai fitur (*missing value*)
- Pencilan data (*outlier*)

Tindakan Mengatasinya

- Perbaikan Data
- Menghapus sampel
- Perbaikan Data
- Menghapus sampel



Jenis Kesalahan Data dan Alternatif Cara Mengatasinya

2. Ketidakkonsistenan nilai fitur di dalam sebuah dataset (2)

Jenis Error

- Deviasi dari nilai fitur yang standar
- Perbedaan unit pengukuran
(contoh: centimeter dengan meter)
- Perbedaan level agregasi (Contoh: akumulasi per hari dengan per minggu)

Tindakan Mengatasinya

- Meningkatkan kapasitas staf data entry
- Menggunakan dukungan software untuk memvalidasi data
- Perbaikan data
- Menghitung ulang.
- Menyamakan tingkat pengukuran menggunakan teknik agregasi atau ekstrapolasi



Contoh Data Kotor: Data Mengandung Missing Values

Row No.	age	education	balance	duration	campaign	y
1	58	tertiary	2143	261	1	no
2	44	secondary	29	151	1	no
3	33	secondary	?	76	1	no
4	47	unknown	1506	92	1	no
5	33	unknown	1	198	1	no
6	35	tertiary	231	139	1	no
7	28	tertiary	?	217	1	no
8	42	tertiary	2	380	1	no
9	58	primary	121	50	1	no
10	43	secondary	593	55	1	no
11	41	secondary	270	222	1	no
12	29	secondary	?	137	1	no
13	53	secondary	?	517	1	no
14	58	unknown	71	71	1	no
15	57	secondary	162	174	1	no
16	51	primary	229	353	1	no
17	45	unknown	13	98	1	no
18	57	primary	52	38	1	no

ExampleSet (45,211 examples, 0 special attributes, 6 regular attributes)



Contoh Data Kotor: Data Tidak Konsisten

Row No.	nama_nas...	jenis_kelamin	umur	jml_pinjaman	jkw
1	x1	P	40	345000	1
2	x2	L	31	350000	7
3	x3	L	29	649926	6
4	x4	P	2	459168	19
5	x5	WANITA	34	3055499	8
6	x6	L	49	2000000	19
7	x7	L	29	8333334	10
8	x8	L	27	4435001	8
9	x9	L	29	560000	19
10	x10	LAKI-LAKI	49	1443750	15
11	x11	LAKI-LAKI	42	3066000	10
12	x12	PRIA	26	4071669	20
13	x13	L	29	228655000	19
14	x14	L	55	840000	4
15	x15	L	38	3000000	24
16	x16	WANITA	29	1640000	19
17	x17	L	41	930000.010	4



Pembersihan Data Kuantitatif

- Data kuantitatif: bilangan bulat atau bilangan floating point dalam berbagai bentuk (set, tensor, deret waktu)
- Tantangan: konversi unit (terutama untuk unit yang mudah berubah seperti mata uang)
- Teknik pembersihan: Normalisasi data
 - perbaikan data missing value
 - perbaikan data outlier
 - perbaikan data salah
 - perbaikan data tidak konsisten

Pembersihan Data Kategori (Kualitatif)

- Data kategori: nama atau kode untuk menetapkan data ke dalam grup, tidak ada urutan atau jarak yang ditentukan
- Masalah umum: salah mengeja saat entri data
- Dasar teknik pembersihan: Normalisasi data

Pembersihan Data Text

- Tantangan utama:
 - Duplikasi data
 - Salah ketik
 - Karakter yang salah
- Cara pembersihan data:
 - Penghilangan duplikat
 - Perbaikan salah ketik
 - Penghilangan karakter yang error



Teknik Pembersihan Data: Duplikat Data

Menghapus duplikat data

Teknik Pembersihan Data: Missing value, tidak konsisten, outlier

Urutan pembersihan berdasarkan prioritas:

1. Data yang salah/kosong diisi dengan nilai sebenarnya
 - Lakukan validasi data dan data diisi/diganti dengan nilai sebenarnya
 - Validasi bisa dilakukan secara manual atau menggunakan kode program
 - Contoh menggunakan kode program:
 - Jenis kelamin laki-laki, atribut “hamil/tidak” pasti berisi “Tidak”
 - Usia bisa dihitung ulang dari tanggal lahir



Teknik Pembersihan Data: Missing value, tidak konsisten, outlier

Urutan pembersihan berdasarkan prioritas:

2. Jika tidak diketahui nilai seharusnya:

- Diisi dengan nilai yang paling mungkin
 - Nilai yang sama untuk label yang sama
 - Data non time series:
 - Nilai tengah
 - mean : data numerik jika outlier sudah dihilangkan
 - median : data numerik jika outlier belum dihilangkan
 - modus : data kategorikal

Dari:

- Kelompok data yang sama, misal nilai gaji dari level pekerjaan yang sama
- Data dengan label yang sama
- Keseluruhan data untuk data non time series
- Data time series:
 - Nilai data sebelumnya atau setelahnya



Teknik Pembersihan Data: Missing value, tidak konsisten, outlier

Urutan pembersihan berdasarkan prioritas:

3. Dihapus, jika tidak memungkinkan diperbaiki (jumlah data masih cukup banyak)

Jika dihapus data menjadi sangat sedikit (tidak cukup digunakan sebagai data latih untuk membangun model):

Harus mengumpulkan atau mencari lagi data lain

Proses dalam Persiapan Data (review)

1. Pemilihan Data

- a. *Record selection*
- b. *Feature selection*

2. Perbaikan Data

- a. Mengisi *missing values*
- b. Perbaikan error
- c. Penanganan *outlier*
- d. Penghapusan duplikasi

3. Konstruksi Data

- a. Reduksi data
- b. Mengubah representasi data
- c. Encoding

4. Integrasi Data

- a. *Data Join*
- b. *Append*





Konstruksi Data: Reduksi Data

Reduksi Data

- Reduksi data dilakukan untuk memperoleh dataset yang lebih sedikit dari sisi volume, namun tetap menghasilkan analisis yang sama
- Mengapa perlu dilakukan:
 - Jumlah data yang dikumpulkan (dari basis data atau data warehouse) sangat besar (orde terabytes)
 - Analisis data terhadap data yang sangat besar akan membutuhkan waktu yang sangat lama

Metode Reduksi Data

- **Reduksi Dimensi (*Dimensionality reduction*)**
Mengurangi jumlah kolom/atribut data
- **Pengurangan Data (*Numerosity reduction*)**
Mengurangi jumlah instan atau sample data



Teknik Reduksi Dimensi (Dimensionality reduction)

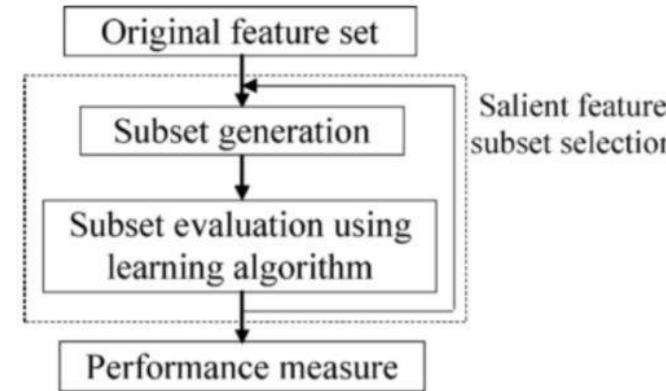
1. Feature Extraction
2. Feature Selection
 - a. Filter Approach
 - b. Wrapper Approach
 - c. Embedded Approach

Reduksi Dimensi dengan Ekstraksi Fitur: Principal Component Analysis (PCA)

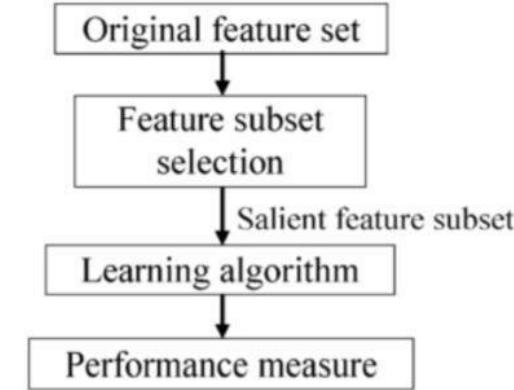
1. Normalisasi data input: Setiap atribut berada dalam kisaran yang sama
2. Hitung vektor ortonormal (unit), yaitu Komponen utama
3. Setiap data input (vektor) adalah kombinasi linear dari nilai k vektor komponen utama.
4. Komponen utama diurutkan dalam urutan decreasing (menurun) “Signifikansi” atau kekuatan
5. Karena komponen diurutkan, ukuran data dapat dikurangi dengan menghilangkan komponen yang lemah, yaitu komponen-komponen dengan varian rendah.



Reduksi Dimensi dengan Seleksi Fitur



Wrapper Approach



Filter Approach

1. Dalam pendekatan wrapper, fitur-fitur digunakan untuk melatih model pembelajaran yang telah ditentukan. Fitur dikurangi secara bertahap dengan melihat kinerja model menaik atau menurun ketika fitur tersebut dihilangkan. Dapat menggunakan forward selection, backward elimination, randomized hill climbing, dll.
2. Dalam pendekatan filter, analisis statistik dari set fitur diperlukan, tanpa menggunakan model pembelajaran apapun. Dapat menggunakan information gain, chi square, log likelihood ratio, dll. (sudah dibahas sebelumnya)
3. Pendekatan yang embedded memanfaatkan kekuatan pelengkap pendekatan wrapper dan filter. Dapat menggunakan decision tree, weighted naïve bayes, dll.

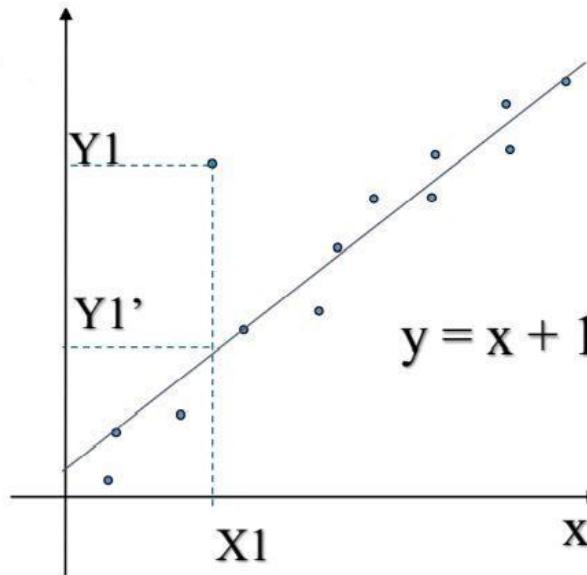


Teknik Pengurangan Data (Numerosity Reduction)

1. Parametrik:
 - a. Regresi
 - b. log-linear model, dll
2. Non-Parametrik
 - a. Histogram
 - b. Clustering
 - c. Sampling

Numerosity Reduction dengan Pendekatan Parametrik

Regresi:



Regresi linier memodelkan hubungan antara dua atribut dengan memodelkan persamaan linier ke kumpulan data.

Misalkan kita perlu memodelkan fungsi linier antara dua atribut.

$$y = wx + b$$

y adalah atribut respons

x adalah atribut prediktor.

Jika kita membahas dari segi data mining, atribut x dan atribut y adalah atribut numerik database, sedangkan w dan b adalah koefisien regresi.

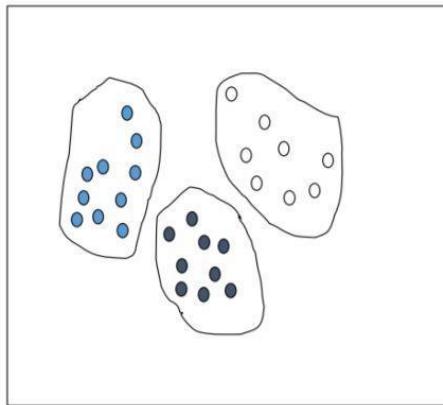
Nilai w dan b dijadikan data baru untuk menggantikan n data pada kelompok data yang bisa diwakili oleh model regresi



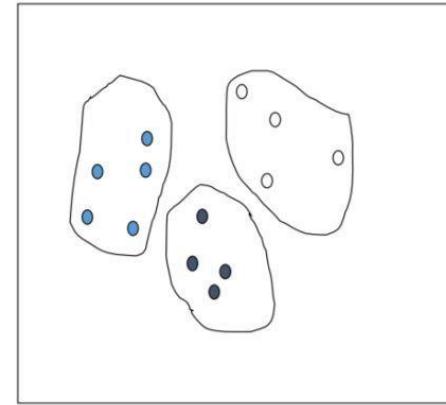
Numerosity Reduction dengan Pendekatan Non-Parametrik

Clustering:

Raw Data



Cluster/Stratified Sample



Teknik clustering mengelompokkan objek-objek yang mirip, sehingga objek-objek dalam satu cluster akan mirip satu sama lain, tetapi berbeda dengan objek-objek di cluster lain.

Seberapa mirip objek di dalam cluster dapat dihitung menggunakan fungsi jarak.

Centroid dari cluster digunakan sebagai data baru untuk mewakili data-data lain pada cluster yang sama





Konstruksi Data: Mengubah Representasi Data (Transformasi Data)

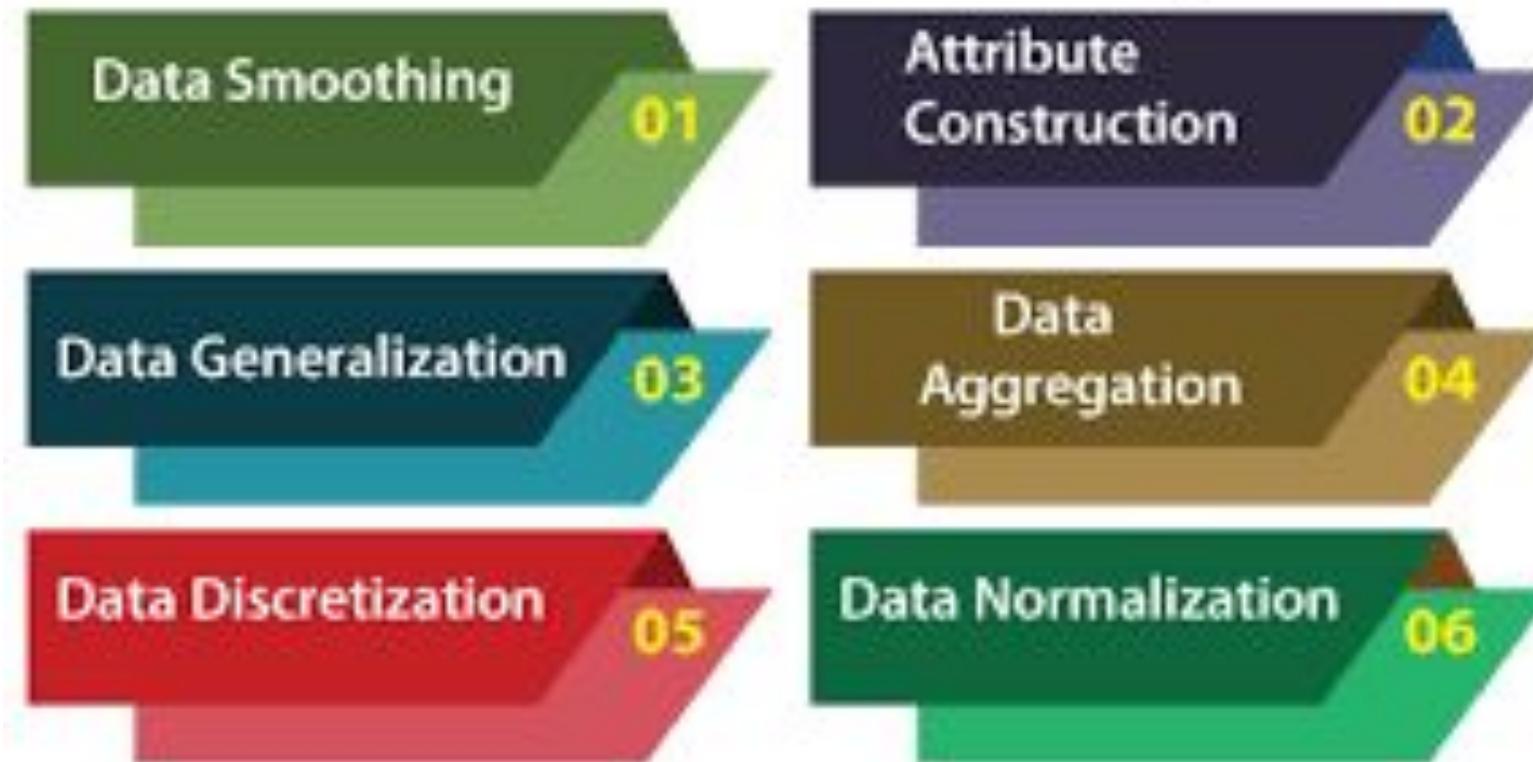


Transformasi Data

Prosedur untuk:

- mengubah
- memformat
- menskalakan
- membersihkan data mentah dalam format tertentu yang diperlukan, baik itu untuk:
 - aplikasi
 - sistem
 - algoritme
 - atau model pembelajaran mesin.

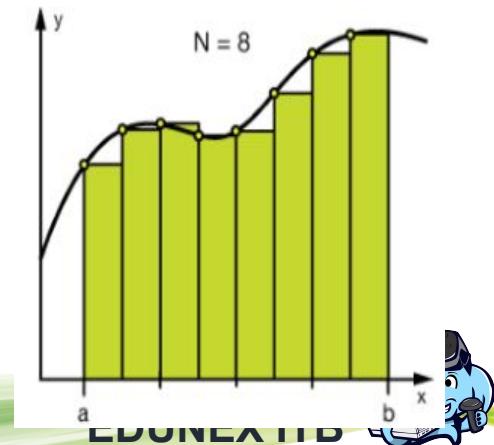
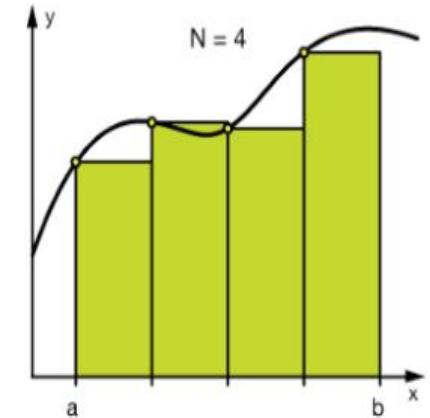
Teknik Transformasi Data



1. Data Smoothing (Binning)

Data binning adalah proses mengelompokkan nilai-nilai data kontinu menjadi interval atau "bin" yang lebih besar, yang mewakili rentang nilai tertentu.

- **Pro:**
 - Dapat diterapkan pada data kategorik dan numerik.
 - Model lebih robust dan mencegah *overfitting*.
- **Kontra:**
 - Meningkatnya biaya kinerja perhitungan.
 - Mengorbankan informasi.
 - Untuk kolom data numerik, dapat menyebabkan redundansi untuk beberapa algoritma.
 - Untuk kolom data kategorik, label dengan frekuensi rendah berdampak negatif pada robustness model statistik.
 - Untuk ukuran data dengan 100 ribu baris, disarankan menggabungkan label/kolom dengan record yang < 100 menjadi kategori baru, misal "Lain-lain".



2. Attribute Construction

- Atribut baru dibuat untuk membantu proses data mining dari atribut yang sudah ada.
- Contoh: Membuat atribut baru 'area' dari atribut 'tinggi' dan 'lebar'.



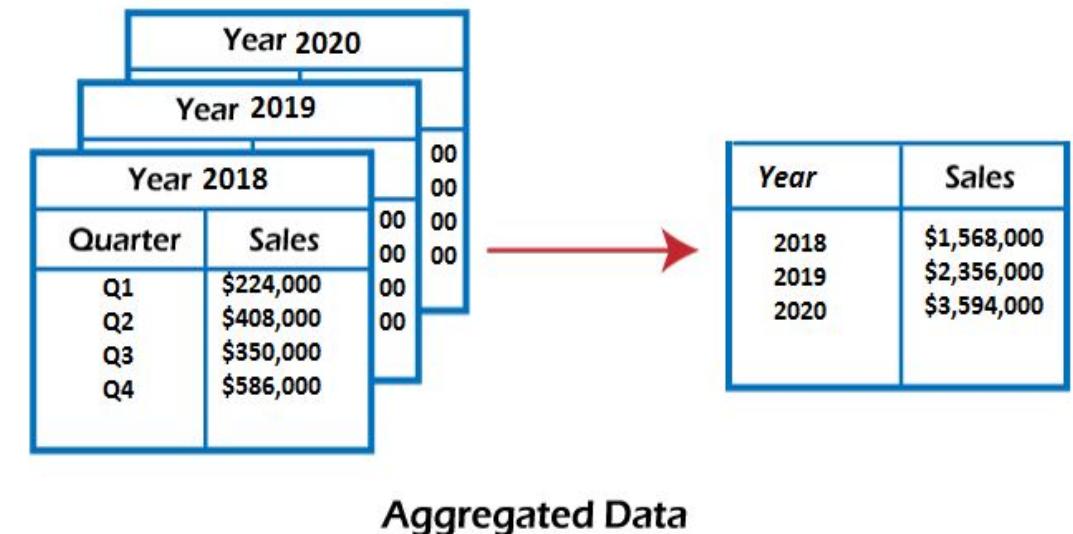
3. Data Generalization

- Mengubah atribut data tingkat rendah menjadi atribut data tingkat tinggi menggunakan hierarki konsep.
- Kegunaan:
 - Mendapatkan gambaran data yang lebih jelas.
 - Mengurangi kedetailan data
- Contoh:
 - Data umur dapat berupa (10, 20, 30, 40) dalam sebuah dataset dapat ditransformasikan ke tingkat konseptual yang lebih tinggi menjadi nilai kategoris anak-anak, remaja, dewasa, tua.
 - Menggantikan alamat rumah individu dengan nama kota atau negara bagian.



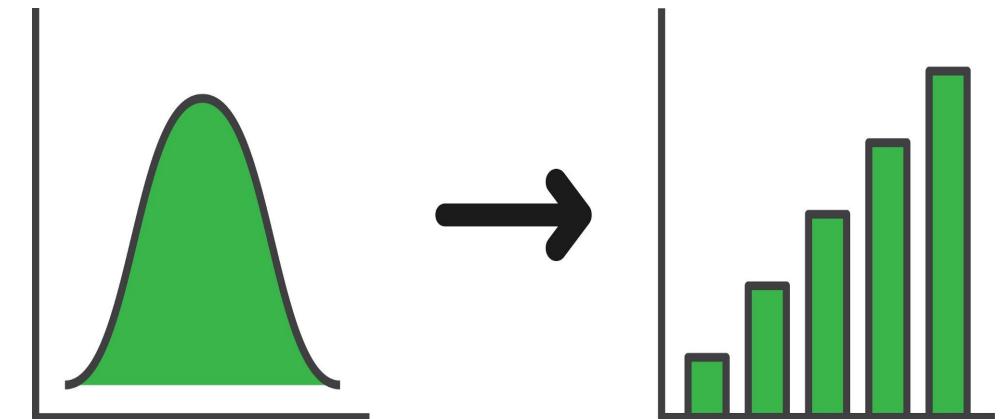
4. Data Aggregation

- **Metode menyimpan dan menyajikan data dalam format ringkasan.**
- Contoh:
Kumpulan data laporan penjualan suatu perusahaan yang memiliki data penjualan triwulanan setiap tahun dapat di-agregasi untuk mendapatkan laporan penjualan tahunan perusahaan.



5. Data Discretization

- Proses mengubah fungsi, model, dan variabel kontinu menjadi diskrit.
- Contoh:
 - misal berat badan < 65 kg (ringan); 65 – 80 kg (mid); > 80 kg (berat).
 - pembulatan sebuah nilai riil ke nilai terdekat

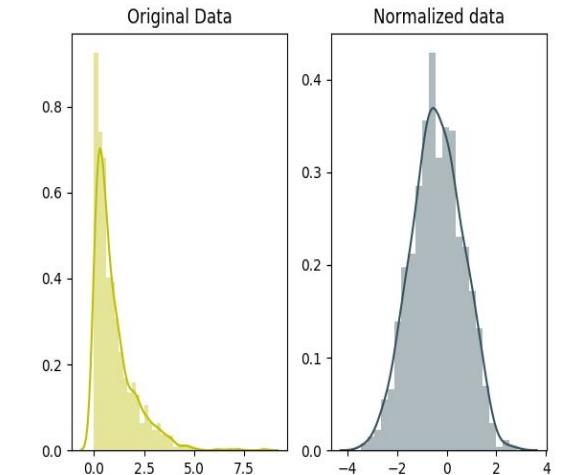


Discretization Process

6. Data Normalization

- Definisi: Teknik penskalaan di mana nilai-nilai digeser dan diubah skalanya sehingga nilainya berkisar antara 0 dan 1 (rentang [0,1]).
- Min-Max Normalization:
Misal X_{max} dan X_{min} masing-masing adalah nilai maksimum dan minimum dari fitur.
 - Ketika nilai X adalah nilai minimum dalam kolom, pembilangnya adalah 0, dan karenanya X' adalah 0.
 - Sebaliknya, ketika nilai X adalah nilai maksimum dalam kolom, pembilangnya sama dengan penyebutnya sehingga nilai X' adalah 1.
 - Jika nilai X berada di antara nilai minimum dan maksimum, maka nilai X' berada di antara 0 dan 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$



Integrasi Data

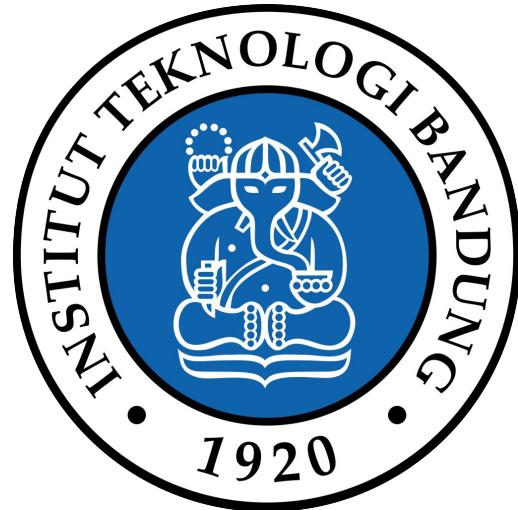
- **Join**

Menggabungkan dua data yang memiliki satu fitur sama.

- **Append**

Menambah instans dari data yang persis sama fitur-fiturnya.





Salam
Semoga Bermanfaat



EDUNEX ITB