

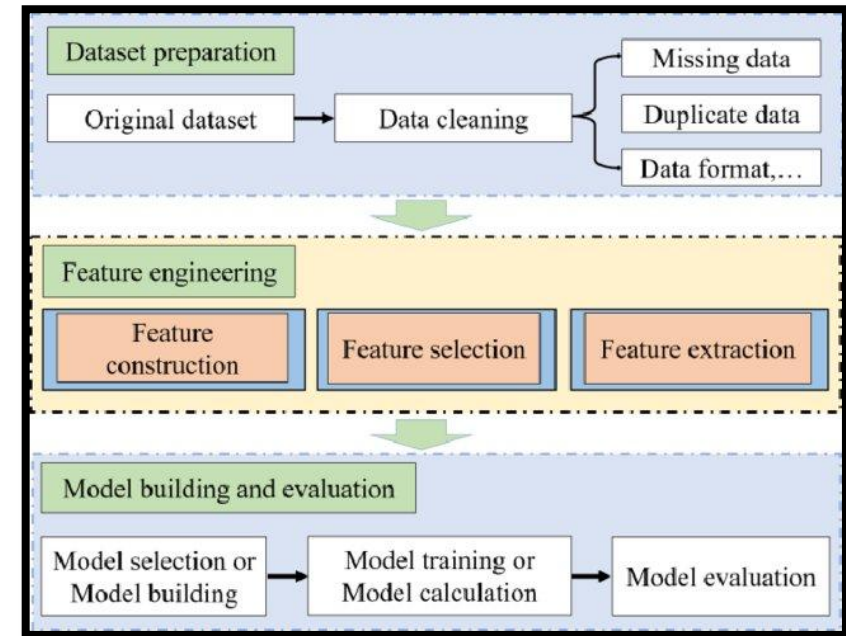
data preparation, dan pre processing data

Pertemuan 4

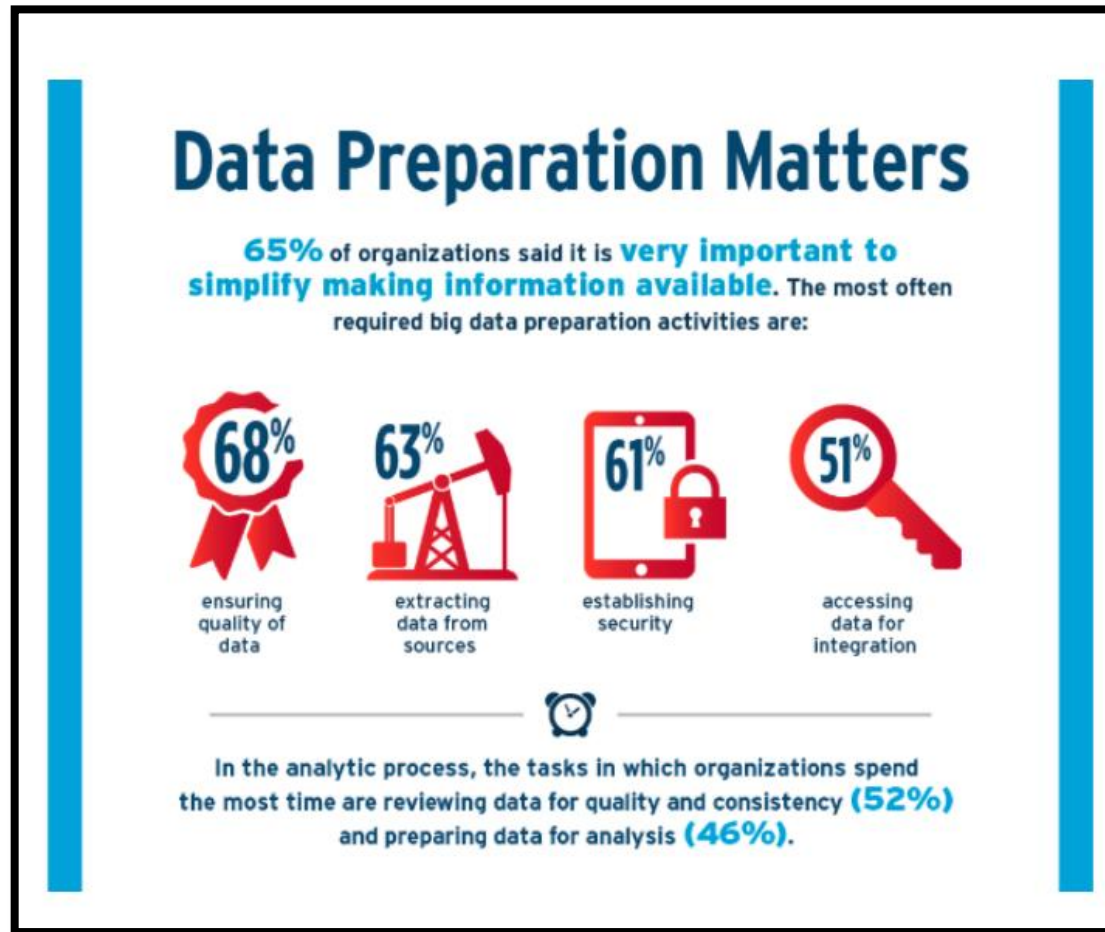
Data Preparation

- *Data Preparation* dilakukan setelah rangkaian proses *business understanding* dan *data understanding*
- Proses *Data Preparation* setidaknya terdiri atas: Memilah data, membersihkan data, mengkontruksi data, menentukan label data dan mengintegrasikan data.
- Untuk Tahapan *Data Preparation* digunakan rujukan SKKNI 299 terkait AI-Data Science yang juga dapat digunakan sebagai kerangka *Data Analytics*.

TUJUAN UTAMA	FUNGSI KUNCI	FUNGSI UTAMA	FUNGSI DASAR
Menemukan pengetahuan, <i>insight</i> atau pola yang bermanfaat dari data untuk berbagai keperluan (orang mengambil keputusan atau sistem memproses lebih lanjut)	Menganalisis Kebutuhan (Requirements) Organisasi	<i>Business Understanding</i>	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
		<i>Data Understanding</i>	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
		<i>Data Preparation</i>	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
		<i>Modeling</i>	12. Membangun skenario pengujian 13. Membangun model
		<i>Model Evaluation</i>	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
		<i>Deployment</i>	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
	Menggunakan model yang dihasilkan	<i>Evaluation</i>	20. Melakukan review proyek 21. Membuat laporan akhir proyek



Urgensi *Data Preparation*



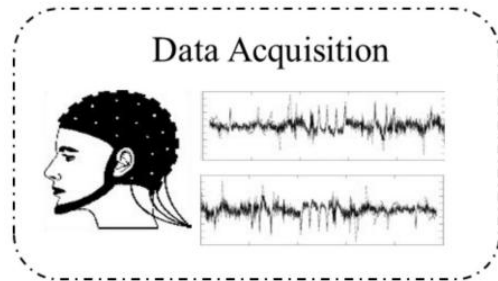
- *Data Preparation* seringkali mengambil porsi kurang lebih 60% dari seluruh proses project berbasis data (*Data Analytics*). Sehingga keberhasilan project *data analytics* seringkali dipengaruhi sangat vital dari proses *Data Preparation*.
- Proses *Data Analytics* bertujuan merubah data menjadi informasi bermanfaat yang mendukung Keputusan strategis bisnis.
- Hasil Analisa data valid, jika didukung oleh data yang berkualitas.
- Oleh karenanya, *data preparation* ditujukan untuk menghasilkan Analisa yang valid yang mendukung Keputusan bisnis

Urgensi *Data Preparation*

Case of Failure Due to Data Preparation

Deteksi ASD (Autism spectrum disorder) melalui Analisis rekam EEG menggunakan Random Forest

- Data diambil dari 15 pasien (4 normal, 11 ASD)
- Signal direkam melalui 15 channel EEG
- *First order statistics* digunakan untuk ekstraksi fitur



1	Anak_Channel	Mean	Variance	Skewness	Kurtosis	Kondisi
2	amerFP1	0,001625845	2819,300414	0,388569232	12,62257062	Normal
3	amerF3	-0,000160359	1663,379895	-0,1088111	14,31750639	Normal
4	amerF7	6,44833E-05	1618,864577	-1,487005131	36,34590976	Normal
5	amerT3	6,5804E-05	839,1332168	-0,329921397	15,90329851	Normal
6	amerT5	8,27872E-05	702,9375687	-1,914153844	71,22939993	Normal
7	amerO1	-5,54296E-05	660,8209878	-0,058461045	4,33513732	Normal
8	amerC4	0,000188206	219,1661912	3,14679394	57,00550545	Normal
9	amerFP2	4,9322E-05	85,40465013	-0,018590194	3,035755416	Normal
10	amerFZ	0,000194675	44,90004573	-0,248984657	7,353518709	Normal
11	amerF4	0,000119959	52,17514946	0,060609521	0,087413313	Normal
12	amerF8	0,000302437	23,89251052	0,09572071	3,74004792	Normal
13	amerC3	1,34068E-05	33,50399758	0,027604081	1,311062702	Normal
14	amerCZ	0,000225541	27,89194903	0,035711643	0,257191446	Normal
15	amerPZ	0,000303029	13,0898407	-0,011293946	0,560008267	Normal
16	amerOZ	0,000208415	13,0898407	-0,011293946	0,560008267	Normal
17	baderFP1	4,10794E-05	1555,008001	-0,282258297	28,73907626	Autism
18	baderF3	-0,000310223	447,1542274	-1,12225671	216,1231373	Autism
19	baderF7	-1,64774E-05	481,3650347	2,744673139	23,95638428	Autism
20	baderT3	-0,000294799	394,1237034	-0,116256322	4,462771723	Autism
21	baderT5	2,61057E-05	175,2856712	28,53596145	1387,499012	Autism
22	baderO1	4,9797E-05	210,8495705	-2,579784799	89,69251698	Autism
23	baderC4	0,000394642	146,6346402	1,047424499	8,751996036	Autism
24	baderFP2	0,000356232	156,4813777	2,339591043	31,26598051	Autism
25	baderFZ	0,000377359	123,2009437	0,530040872	10,53435248	Autism



Random
Forest

Can You Spot what is the mistakes here?

Data Preparation

- A. Memilah Data
- B. Membersihkan data
- C. Mengkontruksi- Transformasi data
- D. Menentukan label data
- E. Mengintegrasikan data

A. MEMILAH DATA (Menentukan Object Data)

KODE UNIT : J.62DMI00.007.1

JUDUL UNIT : Menentukan Objek Data

DESKRIPSI UNIT: Unit kompetensi ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memilah dan memilih data yang sesuai permintaan atau kebutuhan.

ELEMEN KOMPETENSI	KRITERIA UNJUK KERJA
1. Memutuskan kriteria dan teknik pemilihan data	1.1 Kriteria pemilihan data diidentifikasi sesuai dengan tujuan teknis dan aturan yang berlaku
	1.2 Teknik pemilihan data ditetapkan sesuai dengan kriteria pemilihan data.
2. Menentukan <i>attributes (columns)</i> dan <i>records (row)</i> data	2.1 Attributes (columns) data diidentifikasi sesuai dengan kriteria pemilihan data.
	2.2 Records (row) data diidentifikasi sesuai dengan kriteria pemilihan data.

A. MEMILAH DATA (Menentukan Object Data)

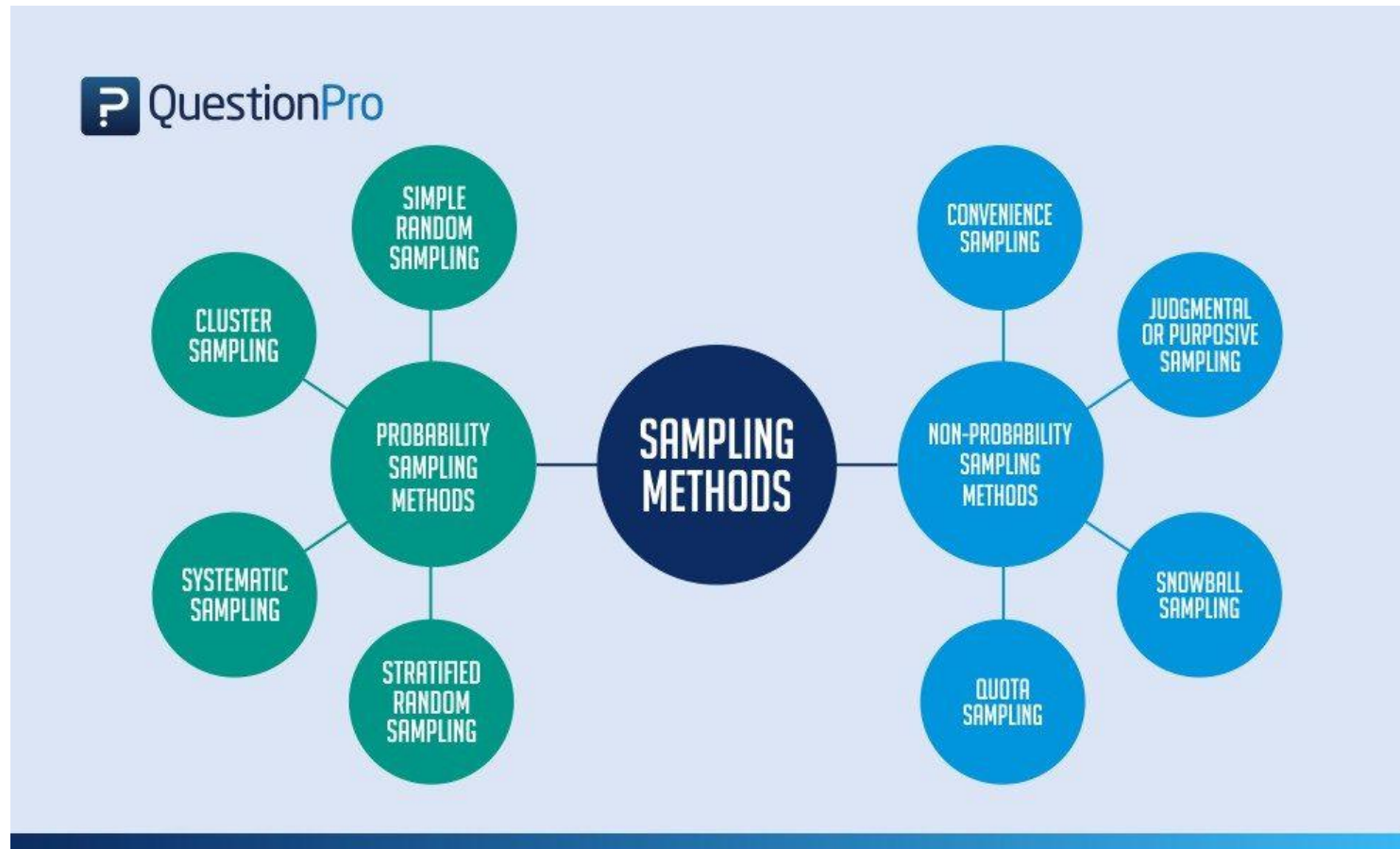
- Hal ini terkait dengan menentukan dataset yang digunakan untuk membangun model.
- Data analytics adalah proses mengubah data menjadi informasi yang bermanfaat.
- Oleh karenanya dataset yang berkualitas sangat diperlukan
- Dataset berkualitas memiliki kriteria: relevan dengan permasalahan bisnis, representative, memiliki fitur yang lengkap sesuai dengan tujuan teknis dan tidak mengandung *missing value* ataupun *duplicate data*.

A. Memilah Data (Menentukan Object Data)

- Jika menggunakan data menggunakan data sekunder maka seringkali diperlukan dari beberapa sumber sehingga memerlukan proses integrasi data (akan dijelaskan berikutnya)
- Jika menggunakan data primer maka beberapa hal berikut wajib diperhatikan
 - Tentukan informasi yang diproyeksikan untuk diperoleh sesuai dengan tujuan teknis
 - Tentukan fitur yang representative dan hendak diukur sesuai dengan tujuan teknis
 - Tentukan mekanisme pengambilan data
 - Tentukan metode sampling

A. Memilah Data (Menentukan Object Data)

Metode Sampling



B. MEMBERSIHKAN DATA

KODE UNIT : J.62DMI00.008.1

JUDUL UNIT : Membersihkan Data

DESKRIPSI UNIT : Unit kompetensi ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membersihkan data yang sesuai permintaan atau kebutuhan.

ELEMEN KOMPETENSI	KRITERIA UNJUK KERJA
1. Melakukan pembersihan data yang kotor	1.1 Strategi pembersihan data ditentukan berdasarkan hasil telaah data. 1.2 Data yang kotor dikoreksi berdasarkan strategi pembersihan data.
2. Membuat laporan dan rekomendasi hasil membersihkan data	2.1 Masalah dan teknis koreksi data dideskripsikan sesuai dengan kondisi data dan strategi pembersihan data. 2.2 Evaluasi dihasilkan berdasarkan analisis koreksi yang telah dilakukan. 2.3 Evaluasi proses dan hasilnya didokumentasikan.

B. MEMBERSIHKAN DATA

- Data seringkali dikatakan “data kotor” jika terdapat beberapa kondisi tidak ideal yang nantinya akan mempengaruhi validitas pembangunan model ataupun Analisa statistic terkait dengan data tersebut.
- Data dikatakan data kotor jika memuat setidaknya salah satu dari:
 - *Missing value* (nilai yang hilang)
 - Nilai Outlier (nilai pencilan)
- Secara Umum tujuan Membersihkan data adalah memastikan seluruh elemen data valid digunakan. Dalam hal ini data sudah tidak memuat nilai *outliers* ataupun nilai *missing value* sudat tertangani.
- Proses membersihkan data memastikan data valid akan membantu performasi model pembelajaran lebih baik dan hasil yang lebih robust terhadap bias.

B. MEMBERSIHKAN DATA

Missing value – define dan jenisnya

- Sebuah kondisi dimana sebuah nilai yang harusnya tersedia pada sebuah variable tidak ditemukan dengan beberapa alasan.
- Namun secara umum kondisi ini terjadi disebabkan oleh non-sampling error (interviewer recording error, respondent inability error, dan respondent unwillingness error).

B. MEMBERSIHKAN DATA

Missing value – define dan jenisnya

Jenis Missing value

- Missing Completely At Random (MCAR):
 - Probabilitas nilai yang hilang tidak bergantung pada nilai yang ada ataupun pada nilai yang hilang itu sendiri
 - Nilai yang hilang diasumsikan mengikuti distribusi nilai yang diketahui
- Missing at Random (MAR):
 - Probabilitas nilai yang hilang mungkin bergantung pada nilai yang diketahui namun tidak pada nilai yang hilang itu sendiri. Hilangnya nilai bergantung pada factor diluar dari nilai tersebut
- Missing not at Random (MNAR)
 - Hilangnya sebuah nilai bergantung pada nilai variable itu sendiri

B. MEMBERSIHKAN DATA

Missing value – Cara Menangani

- Penghapusan data yang hilang secara lengkap. Hal ini bisa dilakukan jika missing value hanya terjadi pada sedikit field data sehingga tidak terlalu mempengaruhi informasi dataset secara keseluruhan
- Menggunakan beberapa Teknik Imputasi (dijelaskan lebih detail dalam Teknik kontruksi data). Pada Teknik ini penting dipahami tipe data dan keterhubungan informasi.
- Metode Khusus. Sebenarnya cara ini sama dengan imputasi, hanya Teknik proyeksi nilai yang hilang tidak hanya menggunakan parameter *first order statistic* saja. Beberapa teknik dalam metode ini: *Hot-Deck-Imputation, K-NN, Expectation Maximization, Full Information Maximum Likelihood (FIML)*

B. MEMBERSIHKAN DATA

Penanganan *Missing Value*

IMPUTASI adalah Teknik yang digunakan untuk mengatasi missing value. Secara definisi imputasi adalah mengganti nilai yang hilang dengan sebuah nilai pengganti

Imputasi Data Numerik

- Imputasi Mean atau Median
- Imputasi Nilai suka-suka (arbitrary)
- Imputasi nilai ujung (end of tail)
- Imputasi K-NN
- Imputasi Regresi



Imputasi Data Kategori

- Imputasi Mode
- Imputasi *New Missing category*
- Imputasi nol

B. MEMBERSIHKAN DATA

Penanganan *Missing Value* – **IMPUTASI MEAN**

Mengganti nilai yang hilang dengan nilai mean data

Age		Age
34	Mean	34
37	35,2	37
NA		35,2
29		29
33		33
NA		35,2
43		43

B. MEMBERSIHKAN DATA

Penanganan *Missing Value* – *IMPUTASI* Nilai Suka suka (*Arbitrary value*)



Mengganti nilai yang hilang dengan nilai suka suak

Kelebihan

- Mudah diimplementasi
- Cocok untuk dataset numerik berukuran kecil
- Cocok untuk MCAR

Kekurangan

- Tidak mempertimbangkan factor korelasi antara fitur
- Kurang akurat
- Tidak memperhitungkan unsur probabilitas
- Tidak cocok untuk penggantian *missing value* lebih dari 5%

Age		Age
34	Mean	34
37	35,2	37
NA		35,2
29		29
33		33
NA		35,2
43		43

B. MEMBERSIHKAN DATA

Penanganan *Missing Value* – *IMPUTASI* Nilai Nilai Ujung (*End of Tail*)

Mengganti nilai yang hilang dengan nilai *end of tail*



Kelebihan

- Mudah diimplementasi
- cocok digunakan untuk data numerik

Ketentuan Khusus

Dalam menentukan nilai end-of tail perlu diperhatikan bentuk sebaran data

- Jika persebaran data normal
 - Nilai end of tail = $\text{mean} + 3 \times \text{std}$
- Jika persebaran data skewed
 - Nilai end of tail menggunakan aproksimasi inter quartile (IQR)
- Imputasi ini hanya diberlakukan untuk data training saja

Age	Mean	35,20	Age
34	Std	5,22	34
37	Mean + 3 std	50,85	37
NA			50,82
29			29
33			33
NA			50,82
43			43

* Asumsi data terdistribusi Normal

B. MEMBERSIHKAN DATA

Penanganan *Missing Value* – *IMPUTASI Modus*



Mengganti nilai yang hilang dengan nilai modus data

Kelebihan

- Mudah diimplementasi
- cocok digunakan untuk data kategori
- Cocok untuk data Missing at random
- Cocok untuk data dengan persebaran skew

Kekurangan

- Mendistorsi relasi label dengan frekuensi tertinggi vs variable lain
- Menghasilkan over-representation jika banyak data yang hilang

Data Jenjang Peneliti		Data Jenjang Peneliti
Peneliti Muda		Peneliti Muda
Peneliti Pertama		Peneliti Pertama
Peneliti Pertama	Mode	Peneliti Pertama
Peneliti Muda	Peneliti Muda	Peneliti Muda
Peneliti Utama		Peneliti Utama
Peneliti Utama		Peneliti Utama
Peneliti Madya		Peneliti Madya
		Peneliti Muda
Peneliti Utama		Peneliti Utama
Peneliti Madya		Peneliti Madya
Peneliti Muda		Peneliti Muda
Peneliti Muda		Peneliti Muda
Peneliti Muda		Peneliti Muda
Peneliti Madya		Peneliti Madya
Peneliti Pertama		Peneliti Pertama
		Peneliti Muda
Peneliti Muda		Peneliti Muda
Peneliti Pertama		Peneliti Pertama

B. MEMBERSIHKAN DATA

Penanganan *Missing Value* – *IMPUTASI Nol/ Konstanta*

Mengganti nilai yang hilang dengan (dalam hal ini label data) dengan konstanta nol

Kelebihan

- Mudah diimplementasi
- cocok digunakan untuk data kategori

Kekurangan

- Tidak mempertimbangkan korelasi antar fitur
- Berpotensi menimbulkan bias

	col1	col2	col3	col4	col5		col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	df.fillna(0)	0	2	5.0	3.0	6	0.0
1	9	NaN	9.0	0	7.0		1	9	0.0	9.0	0	7.0
2	19	17.0	NaN	9	NaN		2	19	17.0	0.0	9	0.0

B. MEMBERSIHKAN DATA

Data Outlier – define dan jenisnya

Disebut sebagai nilai pencilan yang melekat pada sebuah variable. Setiap variable memiliki elemen nilai sejumlah tertentu dengan persebaran tertentu. Sebuah nilai α dikatakan outlier atau pencilan pada variable x , jika nilai tersebut berada jauh dari persebaran nilai lainnya dalam variable x .

#note: Data adalah kombinasi variable dan nilai

Nilai pencilan dapat mempengaruhi keakuratan hasil prediksi model

Beberapa penyebab terjadinya data pencilan

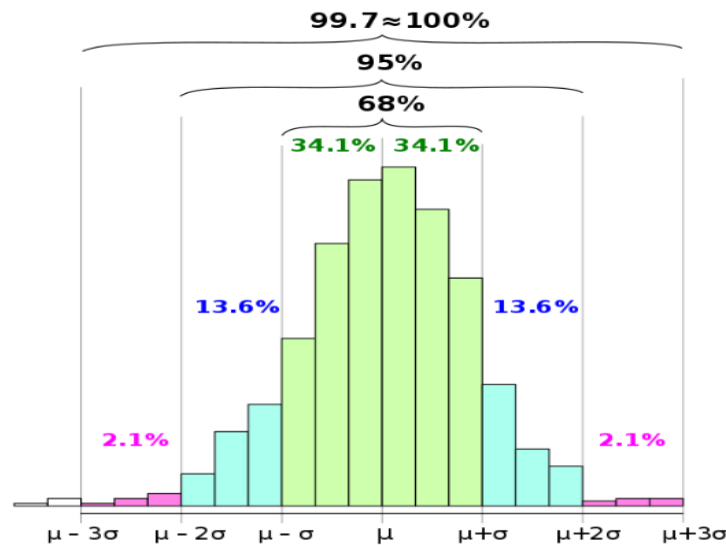
- Kesalahan pencatatan/pengukuran baik yang dilakukan secara manual maupun menggunakan alat.
- Kerusakan data.
- Data yang sesuai kenyataan/hasil observasi

B. MEMBERSIHKAN DATA

Data Outlier – Cara Mendeteksi

▪ Teknik *Standard Deviation*

- Dengan asumsi data memiliki sebaran normal dengan nilai rata-rata μ dan deviasi standar σ maka deviasi standar dari nilai rata-rata data dapat dipergunakan untuk memprediksi jumlah sampel dalam sebuah selang nilai.
- Sebuah nilai fitur diprediksi sebagai outlier jika nilainya diluar dari selang $[\mu - 3\sigma, \mu + 3\sigma]$



$$\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 68.27\%$$

$$\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$$

$$\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$$

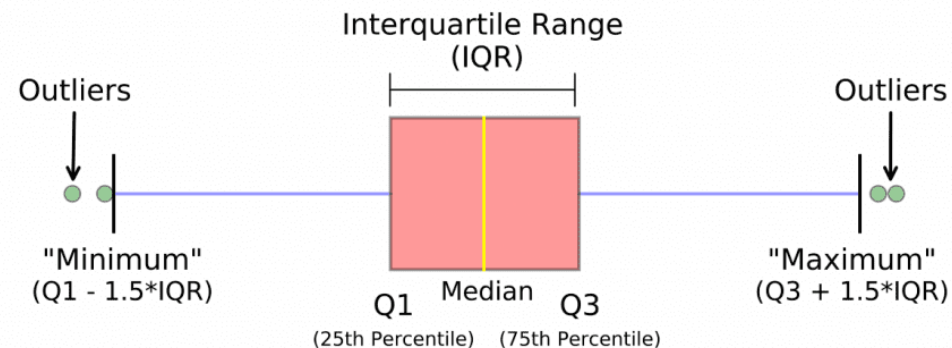
B. MEMBERSIHKAN DATA

Data Outlier – Cara Mendeteksi

▪ Teknik *Interquartil Range (IQR)*.3

- Metode IQR dipergunakan jika nilai fitur dipandang tidak menyebar normal.
- $IQR = Q_3 - Q_1$ dimana Q_3 adalah kuartil ke-3 dan Q_1 adalah kuartil ke-1.
- Sebuah nilai fitur diprediksi sebagai outlier jika nilainya diluar dari selang

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

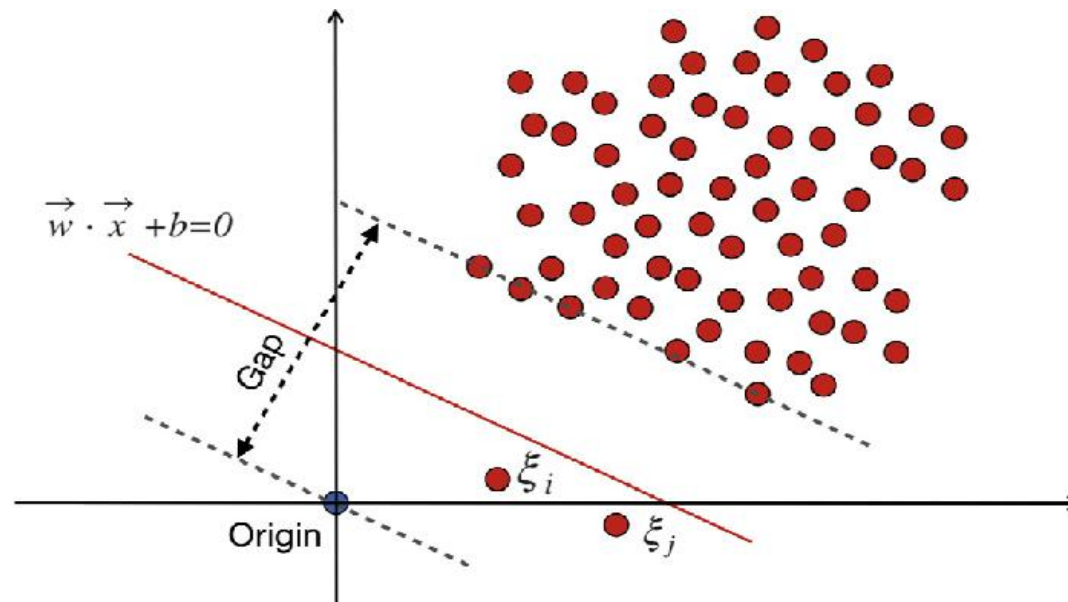


B. MEMBERSIHKAN DATA

Data Outlier – Cara Mendeteksi

■ Teknik *One-class Classification*

- Model Support Vector Machine (SVM) ditraining dengan dataset yang hanya memiliki satu kategori (kategori positif).
- Model hasil training dipergunakan untuk memprediksi data test apakah termasuk kedalam kategori positif atau tidak.



B. MEMBERSIHKAN DATA

Data Outlier – Mengatasi Outlier

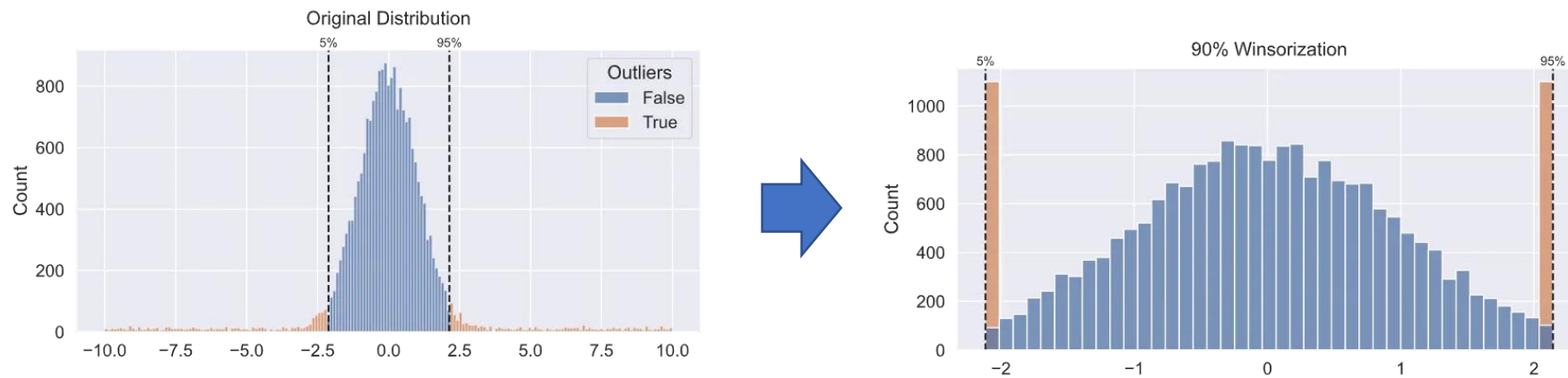
- Secara Umum menangani data outlier Mirip dengan penanganan *missing value*.
- Penanganan Data Outlier:
 - Dibuang, jika: 1. Terdapat kesalahan dalam pengambilan data, 2. Sedikit outlier pada dataset yang besar, 3. saat dimungkinkan mengambil data baru.
 - Ditangani, jika: keberadaan data outlier terlalu banyak sehingga berpotensi memunculkan bias pada analisis data.
 - Trimming, Winsorizing, Imputing , Z-score
 - Mengganti model Prediksi juga dapat membantu mengurangi potensi bias pada data yang mempertahankan nilai bias.

B. MEMBERSIHKAN DATA

Data Outlier – Mengatasi Outlier

Trimming Vs Winsorizing

- Trimming adalah metode penanganan outlier dengan menghilangkan dataset yang mengandung outlier.
- Winsorizing adalah metode penanganan outlier dengan mengganti nilai outlier tersebut dengan nilai inlier terdekat (batas atas atau batas bawah)

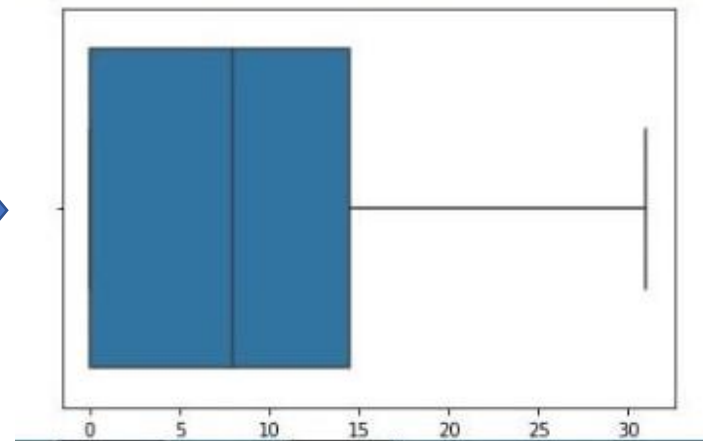
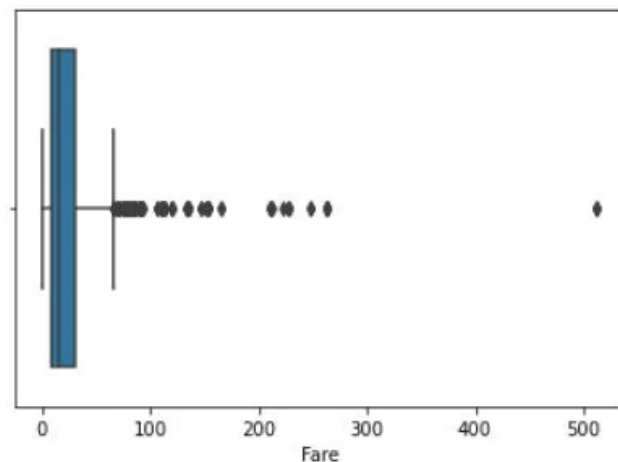


B. MEMBERSIHKAN DATA

Data Outlier – Mengatasi Outlier

Imputing-Imputation

- Mengganti nilai dari data outlier yang dibuang dengan sebuah nilai yang umumnya berasal dari *central tendency measurement*
- Pada umumnya pada data dengan outlier, mean menjadi tidak valid, sehingga proses imputasi untuk menangani data outlier seringkali menggunakan median.
- Teknik ini juga dapat digunakan untuk menangani kasus penggantian nilai pada data yang memuat *missing values*

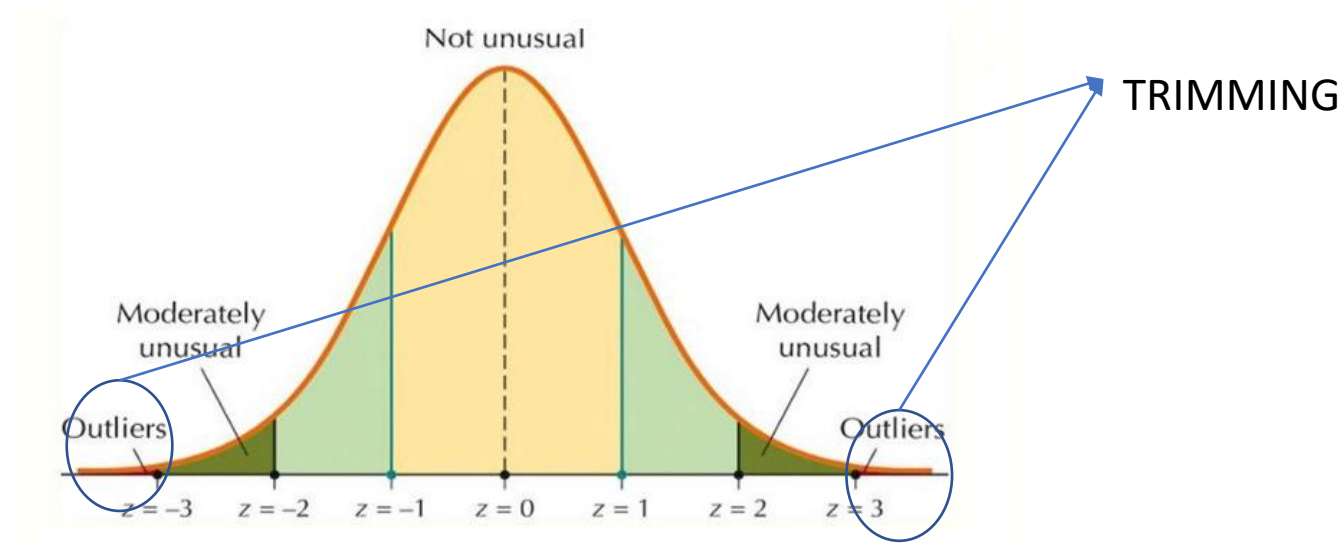


B. MEMBERSIHKAN DATA

Data Outlier – Mengatasi Outlier

Z-Score

- Salah satu dari metode mendeteksi outlier adalah melihat data yang berada diluar range ($\text{mean} - 3 \cdot \text{std}$: $\text{mean} + 3 \cdot \text{std}$).
- Metode penanganan outlier dengan z-score memiliki makna membuang data yang berada diluar range tersebut.



C. Mengkonstruksi data

Kode Unit : J. 62DMI00.009.1

Judul Unit : Mengkonstruksi Data

ELEMEN KOMPETENSI	KRITERIA UNJUK KERJA
1. Menganalisis teknik transformasi data	1.1 Analisis data untuk menentukan representasi fitur data awal . 1.2 Analisis representasi fitur data awal untuk menentukan teknik rekayasa fitur yang diperlukan untuk pembangunan model <i>data science</i> .
2. Melakukan transformasi data	2.1 Transformasi dilakukan untuk mendapatkan fitur data awal. 2.2 Rekayasa fitur data dilakukan untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model <i>data science</i> .
3. Membuat dokumentasi konstruksi data	2.3 Teknis transformasi data dijabarkan dalam bentuk tertulis. 2.4 Hasil transformasi data dan rekomendasi hasil transformasi dituangkan dalam bentuk tertulis.

C. Mengkonstruksi data

Berikut bagian dari Rekayasa Fitur dalam konteks Transformasi atau Kontruksi Data

- 1) Pemilihan fitur data (*feature selection*)
- 2) Transformasi fitur data (*feature transformation*)
 - 1) Imputation – sudah dijelaskan
 - 2) Handling Outlier –sudah dijelaskan
 - 3) Scaling
 - 4) Encoding data kategorikal.
- 3) Reduksi dimensi data (*dimensional reduction*)

C. Mengkonstruksi data - Pemilihan Fitur

Pemilihan fitur data (*feature selection*)

- Tujuan dari pemilihan fitur data adalah mengurangi jumlah fitur yang merepresentasikan data.
 - Mengurangi beban komputasi,
 - Meningkatkan kinerja model prediktif.
- Pendekatan pemilihan fitur data:
 - *Unsupervised* : tidak melibatkan fitur target.
 - Menghilangkan fitur yang memiliki deviasi standar yang rendah.
 - Menghilangkan sebuah fitur dari dua fitur yang memiliki korelasi tinggi.
 - *Supervised* : melibatkan fitur target.
 - Metode *wrapper*,
 - Metode *filter*,
 - Metode *intrinsic*

C. Mengkonstruksi data - Pemilihan Fitur

Pemilihan fitur data (*feature selection*) –metode supervised

- *Filter* :
 - Memilih fitur data secara statistik.
 - Memilih fitur yang memiliki hubungan tertinggi dengan fitur target, misalnya diukur dengan koefisien korelasi.
- *Wrapper* :
 - Mengembangkan beberapa model prediktif dengan beberapa subset fitur data.
 - Memilih subset fitur yang menghasilkan kinerja model tertinggi yang diukur dengan sebuah metrik kinerja.
- *Intrinsic* :
 - Menggunakan model machine learning untuk memilih fitur secara otomatis pada saat training model, misalnya: metode *Recursive Feature Elimination* (RFE)

C. Mengkonstruksi data – Transformasi Fitur *Scaling*

- Beberapa model prediksi akan mencapai kinerja tinggi apabila seluruh fitur data input memiliki skala yang sama.
- Metode umum untuk merubah skala data:
 - **Normalisasi:** teknik untuk merubah skala fitur sehingga nilai fitur berada pada selang $[0,1]$ dengan persamaan:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

dimana: x'_i adalah data hasil perubahan skala, x_i adalah data asli, x_{min} adalah nilai minimum x, dan x_{max} adalah nilai maksimum x.

- **Standardisasi:** teknik untuk merubah skala fitur sehingga nilai fitur memiliki rata-rata – dan deviasi standar 1 dengan persamaan:

$$x'_i = \frac{x_i - \bar{x}}{s_x}$$

dimana: x'_i adalah data hasil perubahan skala, x_i adalah data asli, \bar{x} adalah rata-rata x, dan s_x adalah deviasi standar x.

C. Mengkonstruksi data – Transformasi Fitur *Encoding*

- Model prediksi dibangun dengan mensyaratkan seluruh nilai data sudah dalam bentuk representasi numerik.
- Data kategorikal harus dirubah menjadi representasi numerik dengan teknik encoding.
- Beberapa teknik encoding data kategorikal adalah:
 - One-hot encoding
 - Label encoding
 - Dealing with High Cardinality Categorical Data:
 - Hashing Encoding
 - Target Encoding
 - Weight of Evidence (WOE) Encoding
 - Count Encoding

C. Mengkonstruksi data – Transformasi Fitur

Ordinal encoding

- Dipergunakan untuk jenis data kategorikal yang memiliki urutan nilai.
- Setiap kategori diberikan sebuah nilai bilangan bulat.
- Misalnya, data terdiri dari 3 kategori yaitu: Merah, Hijau, Biru. Maka: encoding masing-masing kategori adalah:
 - Merah 1
 - Hijau 2
 - Biru 3

C. Mengkonstruksi data – Transformasi Fitur

One-hot encoding

- Dipergunakan untuk jenis data kategorikal yang **tidak** memiliki urutan nilai.
- Setiap kategori diwakili dengan kode 1-bit.
- Jika data tidak masuk ke dalam lebih dari satu kategori maka hanya kategori tertentu yang diwakili dengan kode 1-bit.
- Misalnya, data terdiri dari 3 kategori yaitu: Merah, Hijau, Biru. Maka: encoding masing-masing kategori adalah:
 - Merah [001]
 - Hijau [010]
 - Biru [100]

C. Mengkonstruksi data – Transformasi Fitur

Dummy variabel encoding

- Jika terdapat C kategori maka setiap kategori direpresentasikan dengan C-1 bit
- Dummy variable encoding bertujuan untuk menghilangkan redundansi dari representasi kategori.
- Misalnya, data terdiri dari 3 kategori yaitu: Merah, Hijau, Biru. Maka: encoding masing-masing kategori direpresentasikan dengan 2 bit atau (3-1) bit yaitu:
 - Merah [01]
 - Hijau [10]
 - Biru [00]

C. Mengkonstruksi data – Transformasi Fitur

Reduksi dimensi data

- Dimensi data adalah jumlah fitur dari data tersebut.
- Curse of dimensionality adalah sejumlah masalah dibidang pemodelan prediktif yang disebabkan oleh dimensi data yang tinggi.
- Reduksi dimensi merupakan sebuah masalah yang bertujuan untuk mengurangi dimensi data.
- Beberapa metode untuk mereduksi dimensi data adalah:
 - Faktorisasi matriks
 - Pembelajaran Manifold
 - Metode Autoencoder
 - Linear Discriminant Analysis (LDA)
 - Principal Component Analysis (PCA)
 - Singular Value Decomposition (SVD)

D. MENENTUKAN LABEL DATA

KODE UNIT : J.62DMI00.010.1

JUDUL UNIT : Menentukan Label Data

DESKRIPSI UNIT: Unit kompetensi ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan untuk menentukan label data untuk pembangunan model *data science*

ELEMEN KOMPETENSI	KRITERIA UNJUK KERJA
1. Melakukan pelabelan data	1.1 Analisis hasil pelabelan data sejenis yang sudah ada diuraikan kesesuaiannya dengan <i>Standard Operating Procedure (SOP)</i> pelabelan. 1.2 Pelabelan data dilakukan sesuai dengan SOP pelabelan.
2. Membuat laporan hasil pelabelan data	2.1. Statistik hasil pelabelan diuraikan pada laporan. 2.2. Evaluasi proses pelabelan diuraikan pada laporan.

D. MENENTUKAN LABEL DATA

Data Labeling – Definisi dan Fenomena

- Proses ini sering dikenal dengan nama Anotasi data
- Dalam membangun model prediksi ataupun mempelajari pola sebab akibat sebuah fenomena, seringkali diperlukan data yang telah memiliki label.
- Namun demikian ada kalanya data tidak memiliki label dikarenakan objek yang diteliti tidak memiliki kompetensi untuk mendefinisikan label tersebut, ataupun data diambil tanpa pendampingan ahli pada domain terkait.
- Proses anotasi data dapat dilakukan secara:
 - Manual – pelabelan menggunakan pengetahuan *domain expert*
 - Otomatis – menggunakan pendekatan *unsupervised learning*
 - Semi otomatis – mengkombinasikan pendekatan manual dan otomatis

E. INTEGRASI DATA

KODE UNIT : J.62DMI00.011.1

JUDUL UNIT : Mengintegrasikan Data

DESKRIPSI UNIT: Unit kompetensi ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam integrasi data untuk pemodelan *data science*.

ELEMEN KOMPETENSI	KRITERIA UNJUK KERJA
1. Memeriksa <i>dataset</i> yang beragam	1.1 Data yang ada diperiksa kesesuaiannya dengan tujuan pemodelan <i>data science</i> . 1.2 Data yang ada dikumpulkan dengan data lainnya yang sesuai dengan tujuan pemodelan <i>data science</i> .
2. Menggabungkan <i>dataset</i>	2.1 Data yang sesuai pemodelan <i>data science</i> disatukan menjadi <i>dataset terintegrasi</i> untuk <i>data science</i> . 2.2 Data yang sudah disatukan diperiksa kualitas datanya terintegrasi. 2.3 Data yang sudah disatukan diformat sesuai dengan tujuan pemodelan <i>data science</i> .

E. INTEGRASI DATA

Data Integration– Definisi dan Fenomena

- Sebagaimana asas dalam kriteria memilah data, data yang digunakan harus representative sebagaimana tujuan teknis pada masalah bisnis yang hendak diselesaikan.
- Seringkali jika menggunakan data sekunder, membangun dataset utuh diperoleh dari beberapa sumber dengan format yang berbeda namun memiliki informasi yang saling berkesinambungan.
- Proses menyatukan beberapa sumber data tersebut menjadi sebuah kesatuan data tabular yang akan digunakan untuk Analisa data adalah merupakan proses integrasi data.