Learning Objective

1. Peserta mampu memutuskan kriteria pemilihan data
2. Peserta mampu memutuskan teknik pemilihan data
3. Peserta mampu menentukan attributes (columns) data
4. Peserta mampu menentukan records (row) data

1. Memutuskan Kriteria dan Teknik pemilihan data

Konsep dan Definisi

- **Kriteria pemilihan data** mencakup kuantitas data (mencakup volume data yang menggambarkan ukuran data misalkan dalam *terabyte, petabyte atau jumlah record*) dan kualitas data (penilaian terhadap nilai mencurigakan, kosong, inkonsisten, duplikasi maupun ambigu). Kriteria bisa berbentuk ketentuan mengenai pencilan, korelasi antar atribut, data yang kosong dan sebagainya.
- **Teknik pemilihan data** adalah teknik dalam pengambilan sampel, namun secara garis besar dapat dibagi menjadi dua: *probability sampling atau random sampling* dan *non-probability sampling.*

Kriteria Pemilihan Data

- **Kualitas Data**

Kualitas data juga merupakan pertimbangan yang signifikan untuk sumber data eksternal (Azeroual, et al., 2018). Kualitas data merupakan serangkaian tindakan yang menentukan apakah data dapat dipahami secara independen untuk dapat digunakan kembali. Penggunaan kembali data berarti bahwa para peneliti asli atau peneliti lain dapat menggunakan data pada waktu mendatang tanpa menentukan apa yang mungkin digunakan secara spesifik (Peer,Green, & Stephenson, 2014).

Tabel 1. Persyaratan dan Kategori Kualitas Data

Persyaratan	Kategori
Accuracy	Sejauh mana nilai data sesuai dengan nilai aktual atau nilai sebenarnya
Relevancy	Sejauh mana data berlaku (terkait) dengan tugas pengguna data
Representation	Sejauh mana data disajikan dengan cara yang jelas dan jelas
Accessibility	Sejauh mana data tersedia

Sumber: Wang & Strong (1996b)

Kriteria Pemilihan Data

- **Dimensi Kualitas Data**

  Dimensi kualitas data merupakan kumpulan atribut kualitas data yang mewakili satu aspek kualitas data (Wang & Strong, 1996b).

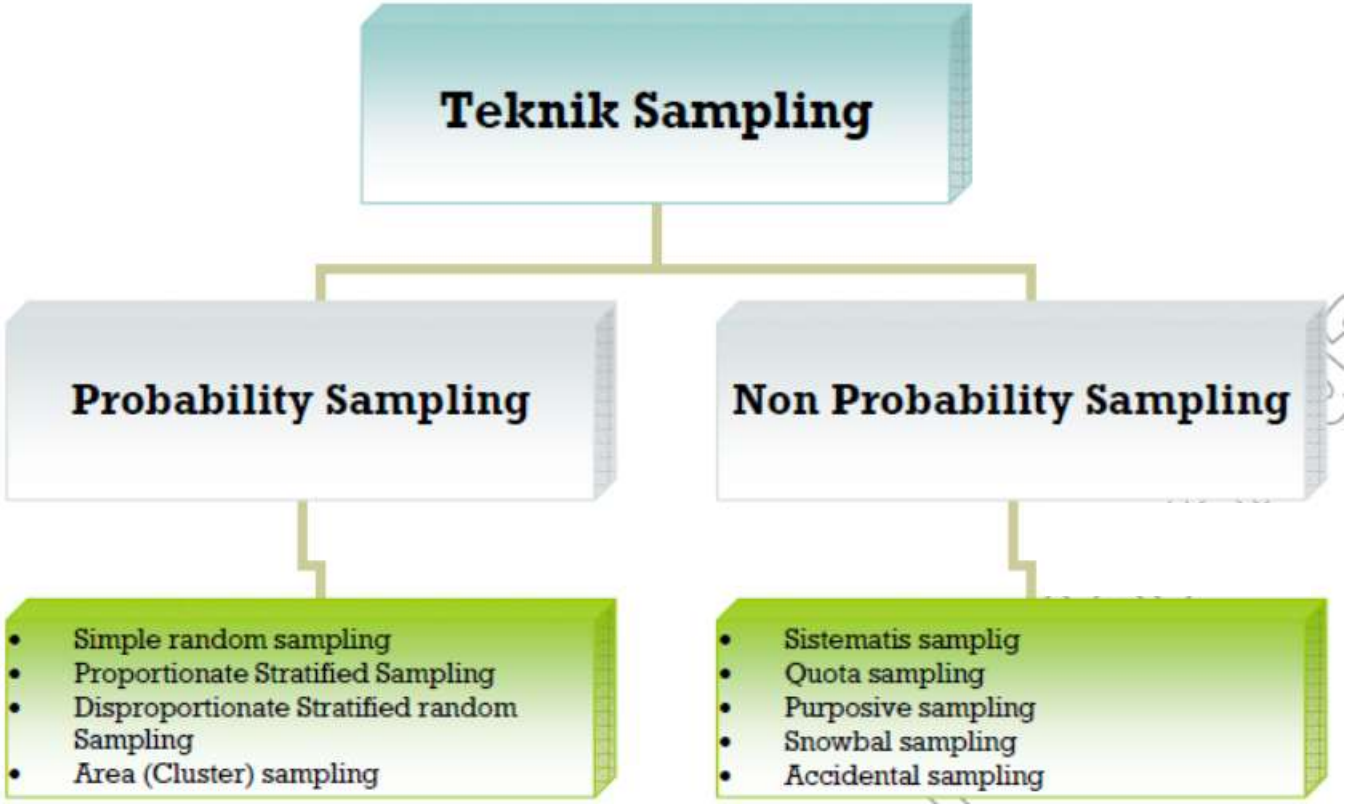
Tabel 2. Dimensi Kualitas Data

Dimensi	Keterangan
Completeness	Sejauh mana data cukup luas, mendalam, dan ruang lingkup untuk tugas yang dihadapi
Correctness/free of error	Sejauh mana data benar dan dapat diandalkan
Representation	Sejauh mana usia data tepat untuk tugas yang dihadapi
Consistency	Sejauh mana data selalu disajikan dalam format yang sama dan kompatibel dengan data sebelumnya

Sumber: Wang & Strong (1996b)

Teknik Pemilihan Data

- **Teknik random sampling (probability sampling)** atau pengambilan sampling secara acak adalah teknik pengambilan sampel dimana semua individu dalam populasi baik secara sendiri-sendiri atau bersama-sama memiliki kesempatan yang sama untuk dipilih menjadi anggota sampel
- **Teknik non random sampling (non probability sampling)** adalah cara pengambilan sampel dimana tidak semua anggota populasi memiliki kesempatan yang sama untuk dipilih menjadi sampel penelitian. Penggunaan teknik non probability sampling ini terkadang digunakan dengan mempertimbangkan factor-faktor


Teknik Sampling
Probability Sampling
Non Probability Sampling
Simple random sampling
Proportionate Stratified Sampling
Disproportionate Stratified random Sampling
Area (Cluster) sampling
Sistematis samplig
Quota sampling
Purposive sampling
Snowbal sampling
Accidental sampling

Data sampling

- **Target Population**:
  - The population to be studied/ to which the investigator wants to generalize his results.
  - The entire group of people or objects to which the researcher wishes to generalize the study findings
  - Example: all peoples infected with COVID19, All low-birth-weight infants, all Indonesian citizen.
- **Accessible Population**:
  - The portion of the population to which the researcher has reasonable access; may be a subset of the target population.
  - Example: all peoples infected with COVID19 in Jakarta Province, All low-birth-weight infants di East Java Province, all Indonesian citizen live in the Java island.
- **Sample:**
  - The selected elements (people or objects) chosen for participation in a study;
  - People are referred to as subjects or participants

Data sampling

- **Sampling Unit**: smallest unit from which sample can be selected.
  - Example: household,
- **Sampling frame**: list of all the sampling units from which sample is drawn.
  - Example: a list of peoples infected with COVID19 in Jakarta Province, a list low-birth-weight infants di East Java Province.
- **Sampling scheme**: method of selecting sampling units from sampling frame.
- **Sampling**: the process of selecting a group of people, events, behaviors, or other elements with which to conduct a study.

Probability and Non-probability sampling

Probability Sampling Techniques

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Multistage sampling
5. Cluster sampling

Non-probability Sampling Techniques

1. Convenience sampling (ease of access)
2. Snowball sampling (friend of friends)
3. Purposive sampling (judgemental)
4. Quota sampling
5.

2. Menentukan Attributes dan Record Data

Entities

Entity: A person, place, object, event, or concept in the user environment about which the organization wishes to maintain data.

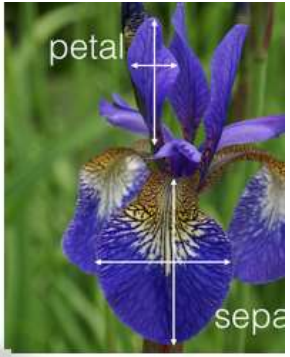
An **entity** is a person, place, object, event, or concept in the user environment about which the organization wishes to maintain data. Thus, an entity has a noun name. Some examples of each of these *kinds* of entities follow:

Person:	EMPLOYEE, STUDENT, PATIENT
Place:	STORE, WAREHOUSE, STATE
Object:	MACHINE, BUILDING, AUTOMOBILE
Event:	SALE, REGISTRATION, RENEWAL
Concept:	ACCOUNT, COURSE, WORK CENTER

Entity type: EMPLOYEE

Attributes	Attribute Data Type	Example Instance	Example Instance
Employee_Number	CHAR (10)	642-17-8360	534-10-1971
Name	CHAR (25)	Michelle Brady	David Johnson
Address	CHAR (30)	100 Pacific Avenue	450 Redwood Drive
City	CHAR (20)	San Francisco	Redwood City
State	CHAR (2)	CA	CA
Zip_Code	CHAR (9)	98173	97142
Date_Hired	DATE	03-21-1992	08-16-1994
Birth_Date	DATE	06-19-1968	09-04-1975

Dataset (Himpunan Data)



Attribute/Feature/Dimension

Class/Label/Target

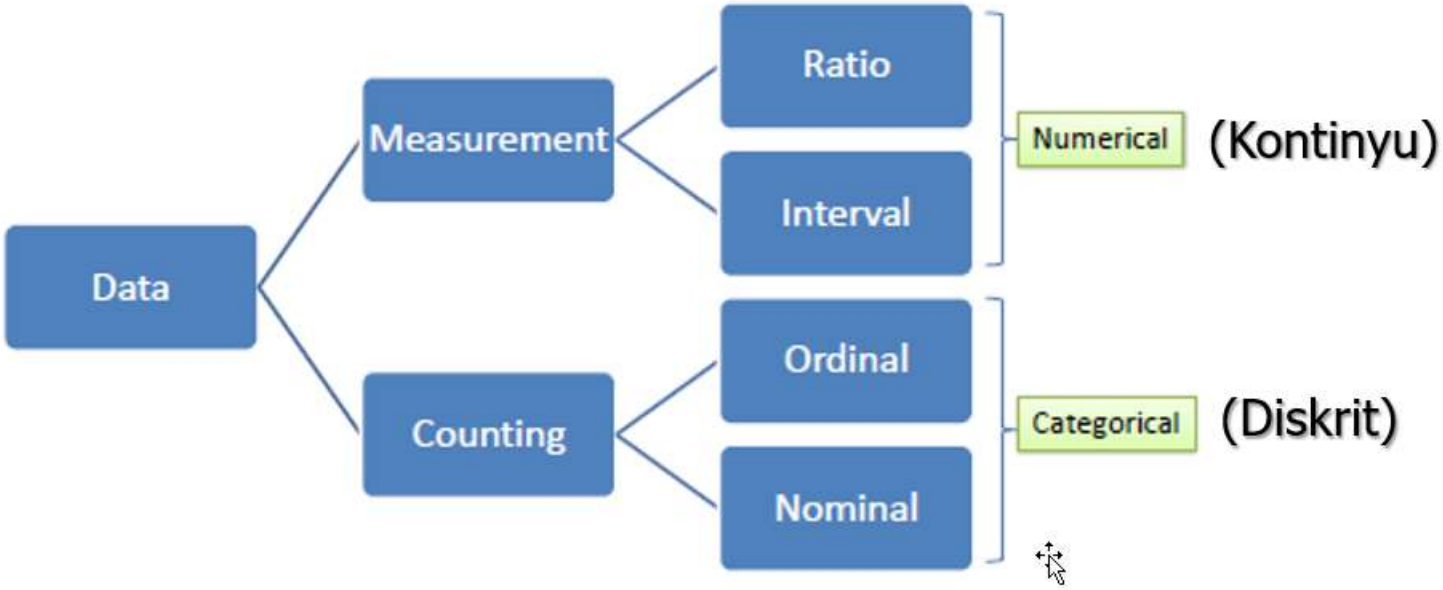
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
…					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
…					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*

Record/ Object/ Sample/ Tuple/ Data

Nominal

Numerik

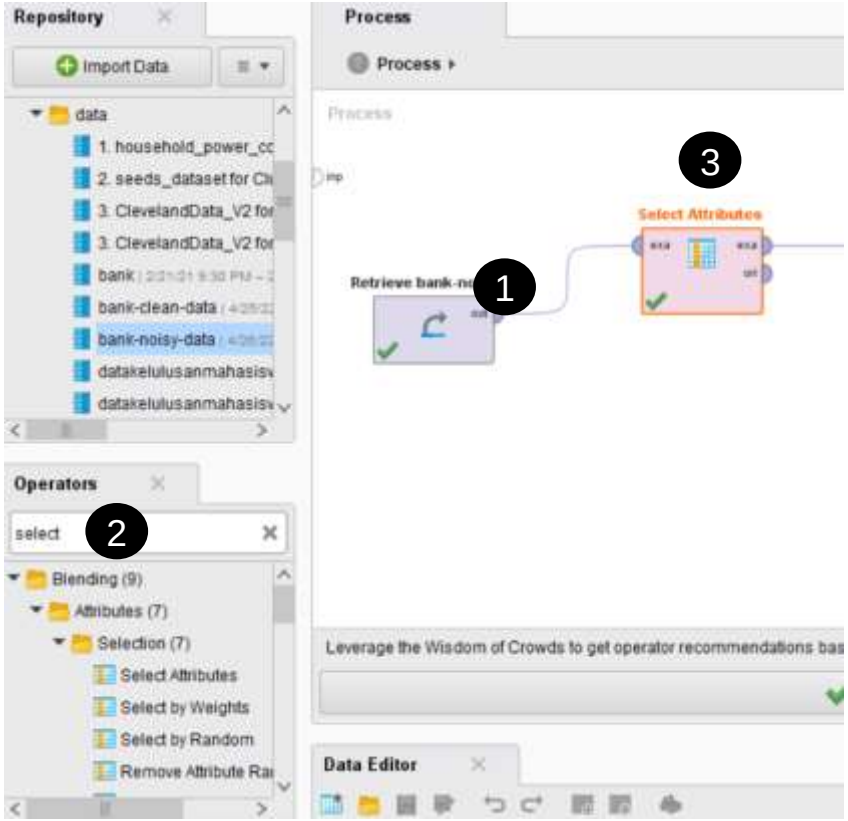
Tipe Data


Data
Measurement
Counting
Ratio
Interval
Ordinal
Nominal
Numerical
Categorical
(Kontinyu)
(Diskrit)

Tipe Nilai Atribut pada Rapidminer

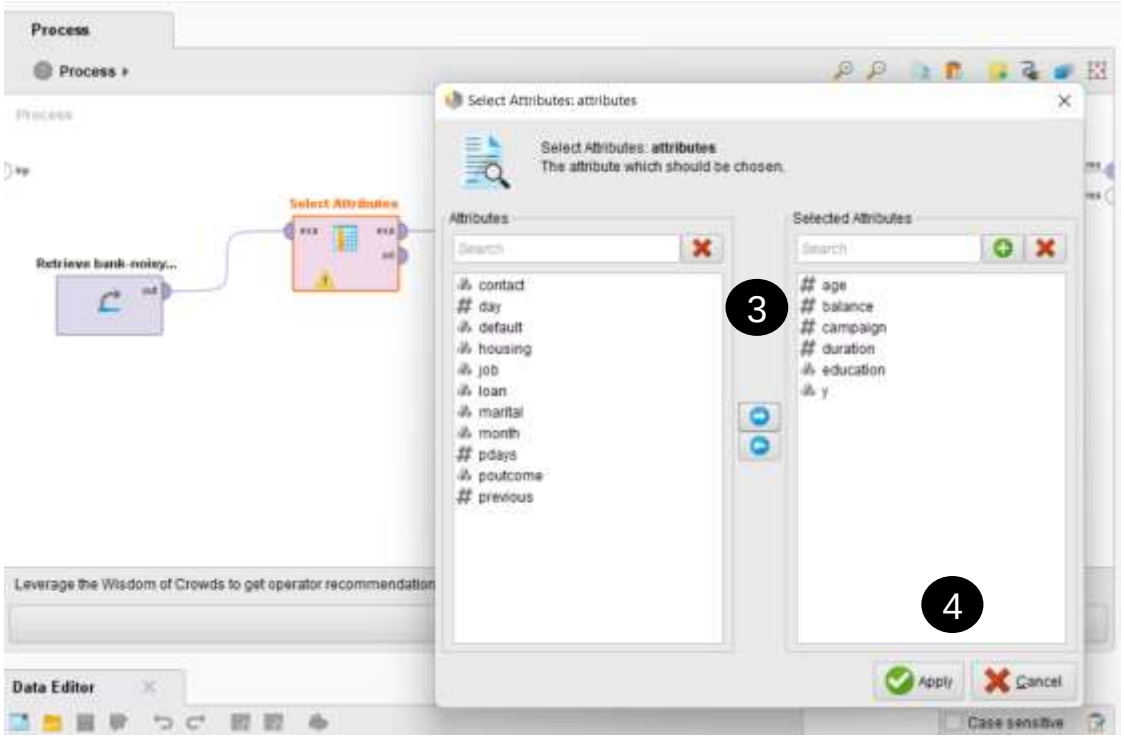
1. nominal: nilai secara kategori
2. binominal: nominal dua nilai
3. polynominal: nominal lebih dari dua nilai
4. numeric: nilai numerik secara umum
5. integer: bilangan bulat
6. real: bilangan nyata
7. text: teks bebas tanpa struktur
8. date_time: tanggal dan waktu
9. date: hanya tanggal
10. time: hanya waktu

Praktek: Memilih Atribut dengan Tools Rapidminer

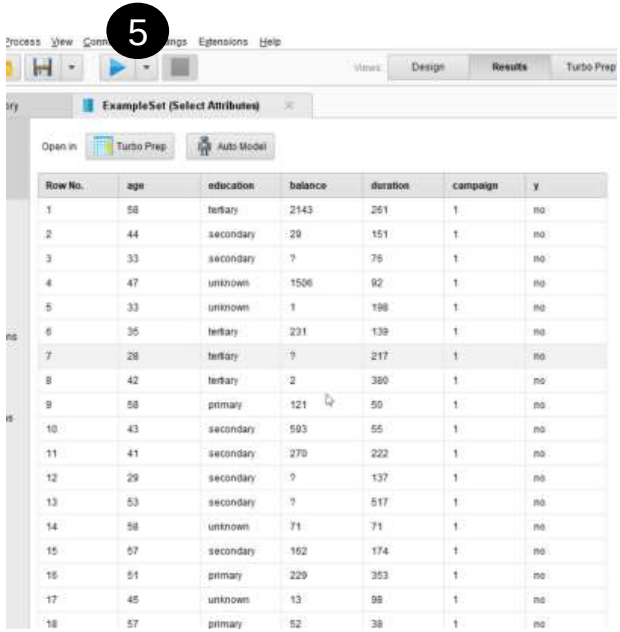


1. Pilih Dataset
2. Ketik Select atribut pada menu operator
3. Drag ke area process

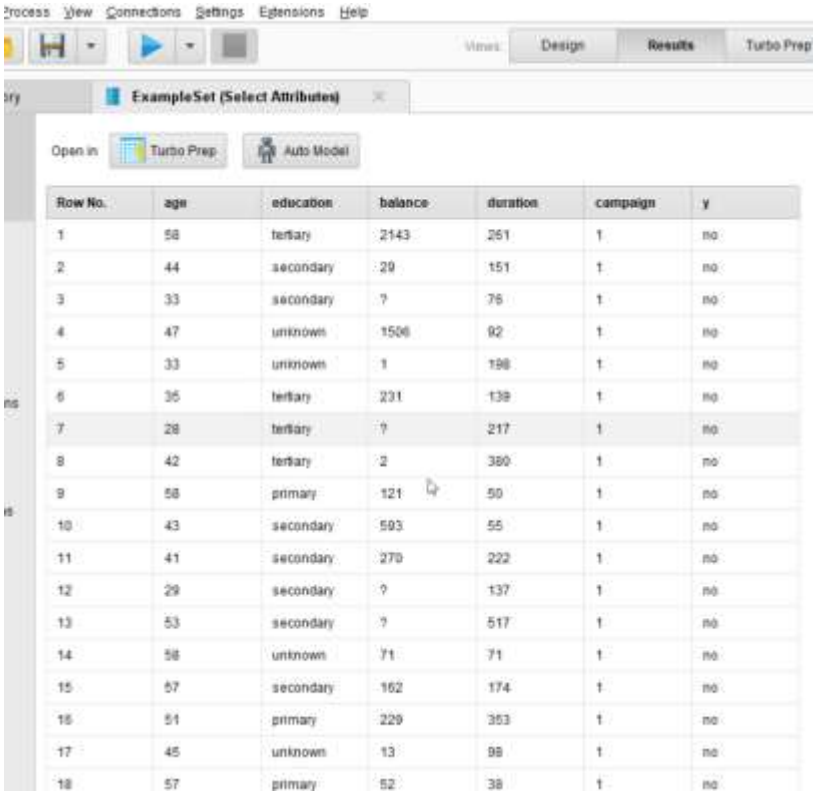
Memilih Atribut dengan Tools Rapidminer



3. Pilih Atribut sesuai tujuan teknis data science
4. Apply, lalu klik run
5. Hasilnya akan tampil sebagai berikut



Memilih Atribut dengan Tools Rapidminer



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

Latihan Tugas

1. Amati dataset pada studi kasus
2. Tentukan Atribut sesuai tujuan teknis data science

Referensi

- Krensky P. Data Pre Tools: Goals, Benefits, and The Advantage of Hadoop. Aberdeen Group Report. July 2015
- SAS. Data Preparation Challenges Facing Every Enterprise. ebook. December 2017
- https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=6e9aa0e36f63
- https://improvado.io/blog/what-is-data-preparation
- https://www.youtube.com/watch?v=VBn9fhaz-J8

Tools / Lab Online

- spreadsheet
- rapidminer

Summary

- Data preparation memiliki sebutan lain, di antaranya data pre-processing, data cleaning, data manipulation,
- Data preparation mengambil porsi kerja terbanyak dalam data science 60-80%
- Data preparation membutuhkan ketelitian dan kesabaran/kerajinan dari peneliti DS, terutama pemula
- Data Validation merupakan tahapan kritikal dari DS namun sering diabaikan para peneliti
- Seleksi Fitur harus dilakukan di awal tahapan data preparation setelah melakukan penentuan metode/teknik sampling
- Data cleaning merupakan pekerjaan yang sangat memerlukan keahlian teknik DS terkait penggunaan tools dan coding
- Kebersihan data merupakan sarat mutlak untuk Model Prediksi yang Baik.

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih