

# DATA PREPROCESSING

Pengajar:

Dian Ade Kurnia, M.Kom  
Rusnanda Farhan  
Rusnandi Fikri  
Rika Sahriana





## Capaian Pembelajaran

1. Mahasiswa mampu melakukan pemilihan dan pemilahan data sesuai kebutuhan dan sumberdaya yang dimiliki
2. Mahasiswa mampu melakukan pembersihan data
3. Mahasiswa mampu melakukan pemeriksaan kualitas dan kecukupan data

# Pengenalan Data Preprocessing



## Terminologi dan Definisi

Pre-Processing

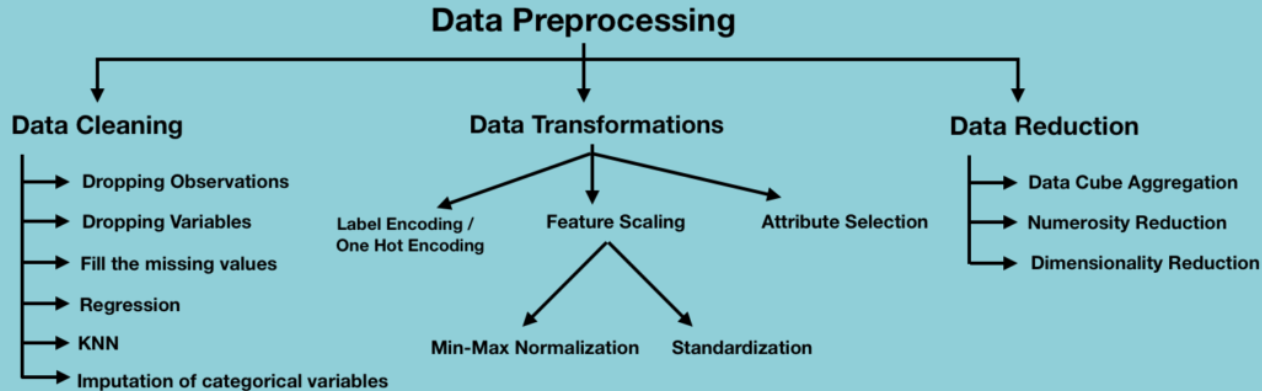
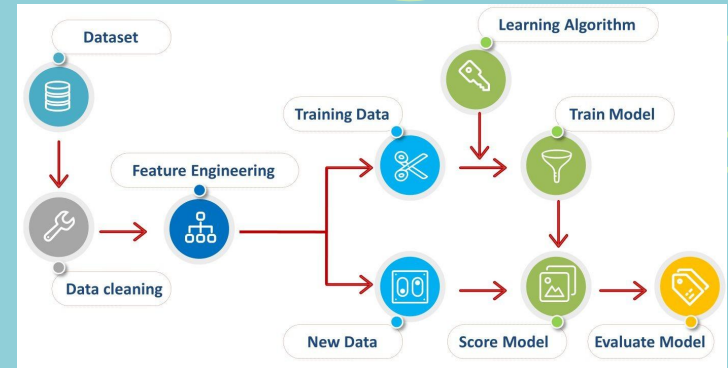
Data  
Manipulation

Data Cleansing

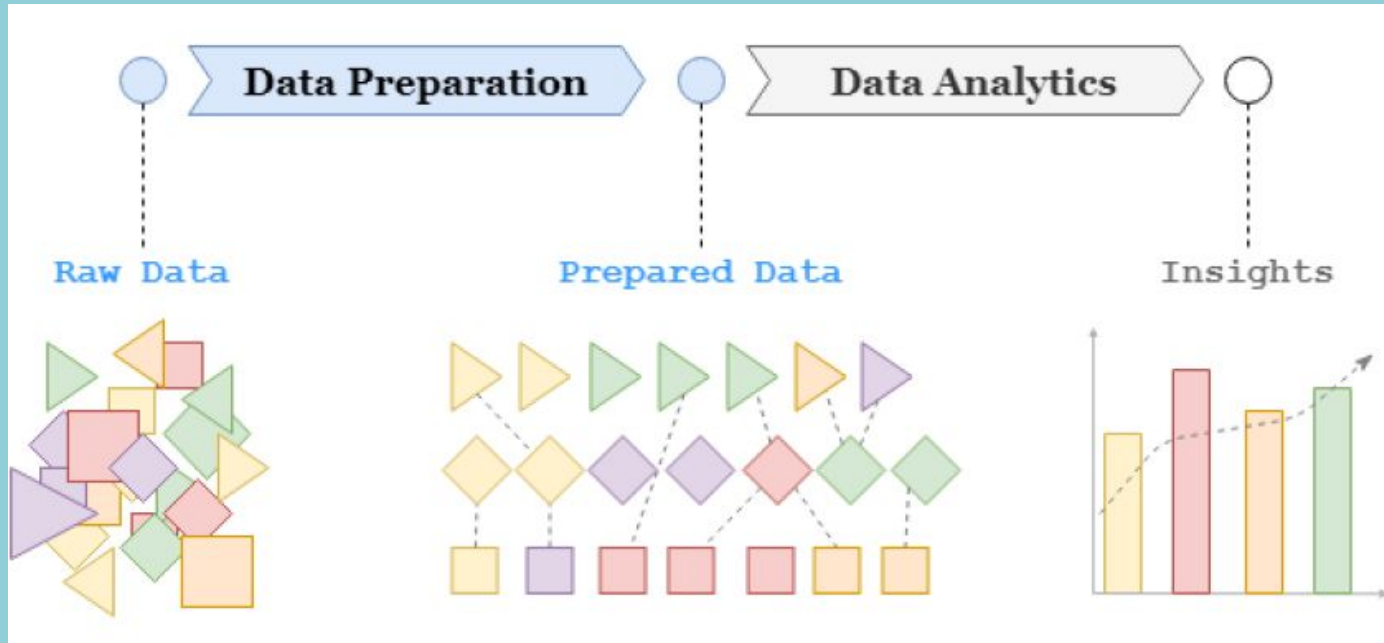
Normalization

## What is Data Preprocessing?

Data preprocessing adalah tahap menyiapkan/membersihkan data yang kotor untuk selanjutnya akan diproses menggunakan model machine learning. Dengan dilakukannya data preprocessing maka dapat meningkatkan efisiensi model dan meningkatkan performa model.



# Langkan Pemrosesan Data



# Data Preprocessing = Feature Engineering

## What is Feature?

Feature adalah data dependen atau predictor yang digunakan untuk analisis.

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr.	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mr.	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M.	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs.	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. W.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. J.	male		0	0	330877	8.4583		Q

location	date_of_sale	property_size_sq_m	number of bedrooms	price	type
Slough	12/4/1993	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	terraced,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

## Feature and Target

Seperti yang dijelaskan sebelumnya Feature adalah data dependen atau predictor yang digunakan untuk analysis. Lalu Target adalah dependen variable atau label, ini adalah data yang akan diprediksi.

Input					Label	
BRAND	TYPE	CYLINDER	ENG-SIZE	STROKE	PRICE	RISK
Brand-A	sedan	four	109	3.4	13950	POS
Brand-A	sedan	five	136	3.4	17450	POS
Brand-B	sedan	four	108	2.8	16430	POS
Brand-B	sedan	four	108	2.8	16925	POS
Brand-C	hatchback	three	61	3.03	5151	NEG
Brand-C	hatchback	four	90	3.11	6295	NEG
Brand-D	hatchback	four	90	3.23	5572	NEG
Brand-D	hatchback	four	90	3.23	6377	NEG



$f(x)$





## DATA CLEANSING

**Missing Value**, adalah data yang tidak lengkap diperlihatkan dengan Na di dataframe. Untuk mengetahui apakah ada missing value dalam dataset dapat menggunakan : `Dataframe.isnull().sum()`

```
# Checking missing value for each feature
print('Checking missingg value for each feature:')
print(dataset.isnull().any(),'\n')
print(dataset.isnull().sum(),'\n')

# Counting total missing value
print('\nCounting total missing value:')
print(dataset.isnull().sum().sum())
```

executed in 29ms, finished 11:15:27 2021-10-30

Checking missingg value for each feature:

Administrative	True
Administrative_Duration	True
Informational	True
Informational_Duration	True
ProductRelated	True
ProductRelated_Duration	True
BounceRates	True
ExitRates	True
PageValues	False
SpecialDay	False
Month	False
OperatingSystems	False
Browser	False
Region	False
TrafficType	False
VisitorType	False
Weekend	False
Revenue	False

dtype: bool

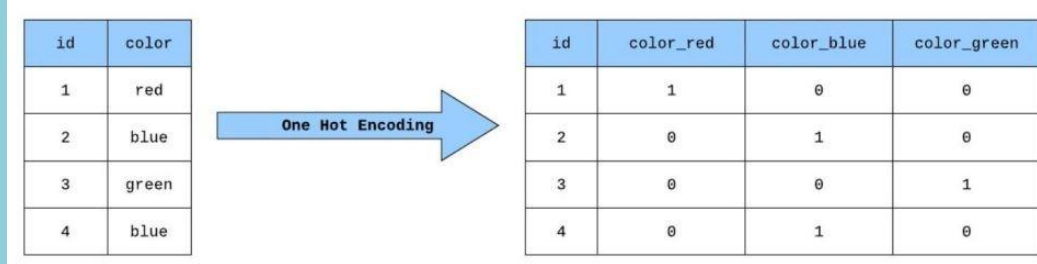
Administrative	14
Administrative_Duration	14
Informational	14
Informational_Duration	14
ProductRelated	14
ProductRelated_Duration	14
BounceRates	14
ExitRates	14
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0

dtype: int64

Counting total missing value:  
112

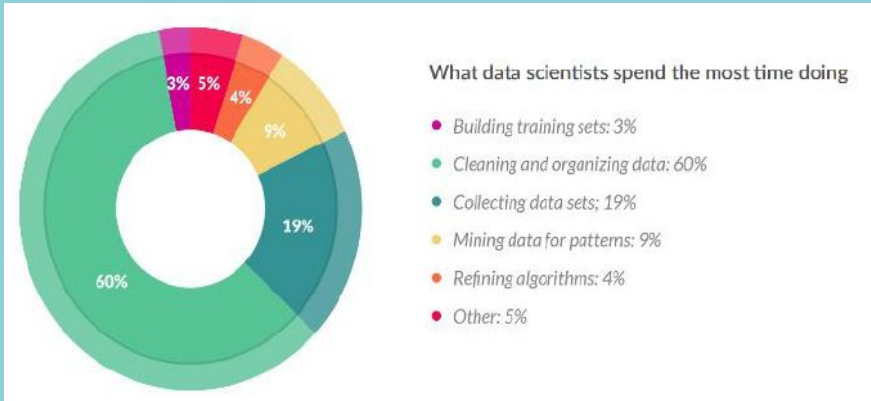
# Data Encoding

Dalam data kategorikal yang bertipe object tidak dapat diproses ke dalam model. Maka perlu dilakukan Encoding untuk merubah nilainya menjadi numeric.



# Fakta Terkait Data Preparation

## Porsi kegiatan data scientis



## Data Preparation

### Data Preparation Matters

**65%** of organizations said it is **very important to simplify making information available**. The most often required big data preparation activities are:



ensuring  
quality of  
data



extracting  
data from  
sources



establishing  
security



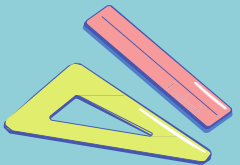
accessing  
data for  
integration



In the analytic process, the tasks in which organizations spend the most time are reviewing data for quality and consistency (**52%**) and preparing data for analysis (**46%**).

## How Important is Data Preprocessing

Data dunia nyata cenderung incomplete, noisy, dan inconsistent. Hal ini dapat menyebabkan rendahnya kualitas data yang dikumpulkan dan selanjutnya rendahnya kualitas model yang dibangun di atas data. Untuk mengatasi masalah ini, Data Preprocessing menyediakan operasi yang dapat mengatur data ke dalam bentuk yang tepat untuk pemahaman yang lebih baik. Kita tidak dapat memahami perilaku atau tren data. Oleh karena itu, kita perlu mengubah atau mengaturnya agar menjadi format yang tepat dengan menggunakan Data Preprocessing. Dengan dilakukannya data preprocessing maka dapat meningkatkan efisiensi model dan meningkatkan performa model.



# Pentingnya Data Preparation



# Manfaat Data Preparation

Kompilasi data menjadi efisien dan efektif  
Identifikasi dan Memperbaiki Error  
Mudah perubahan secara global  
Menghasilkan informasi yang akurat untuk pengambilan keputusan  
Nilai bisnis dan ROI (Return of Investment) akan meningkat

## Data preparation market size



Perkembangan Data Preparation

# Tantangan Data Preparation



Memakan waktu lama

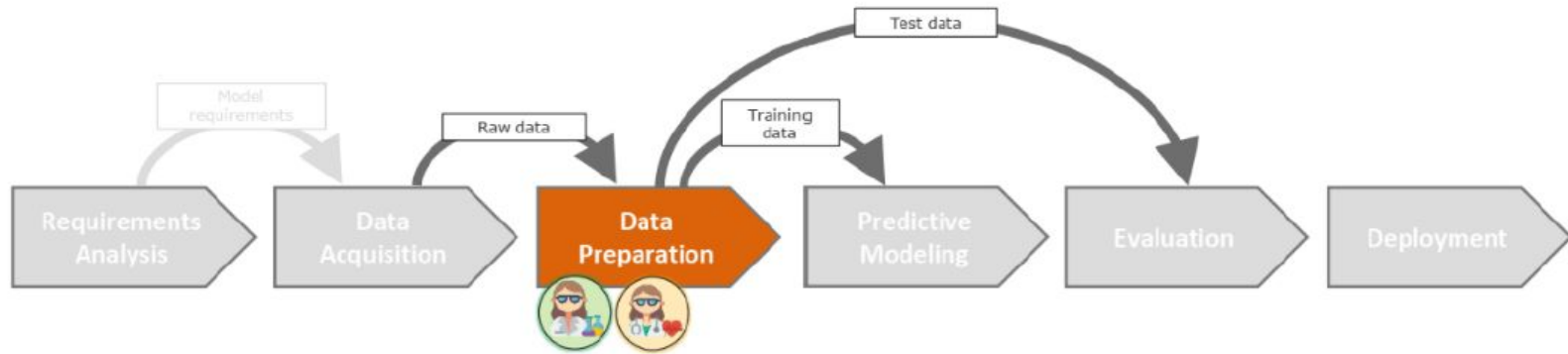
Porsi teknis yang dominan

Data yang tersedia tidak akurat / jelas/tidak langsung pakai

Data tidak balance saat pengambilan sampel

Rentan akan error

# Tahapan Data Preparation



## Roles



Data Scientist



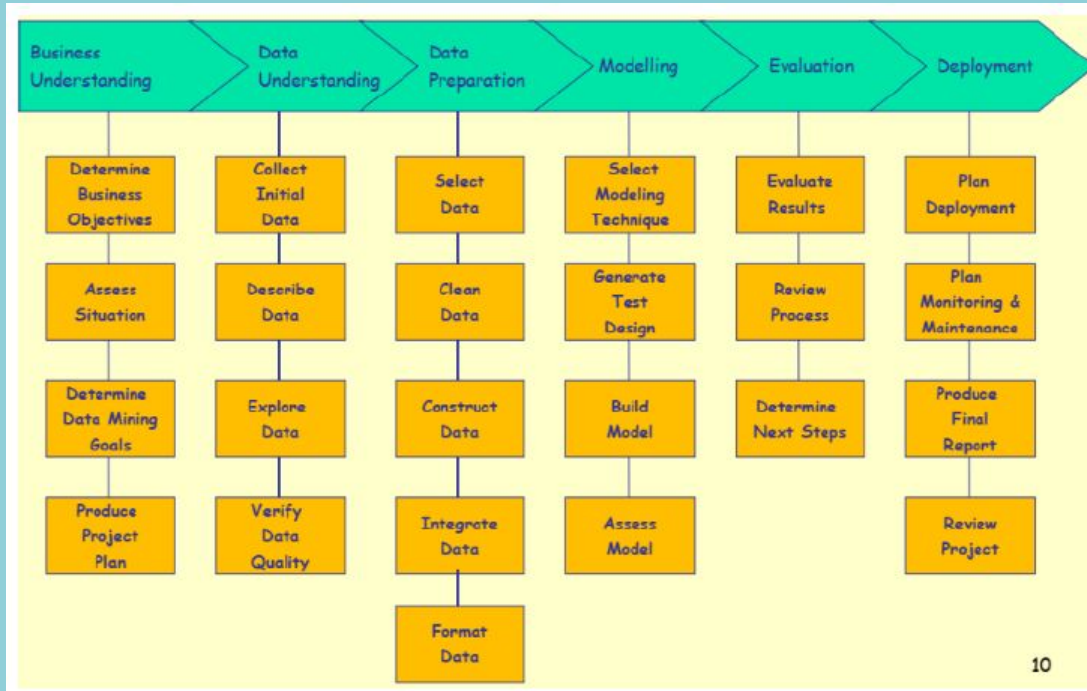
Domain Expert



(Data) Engineer

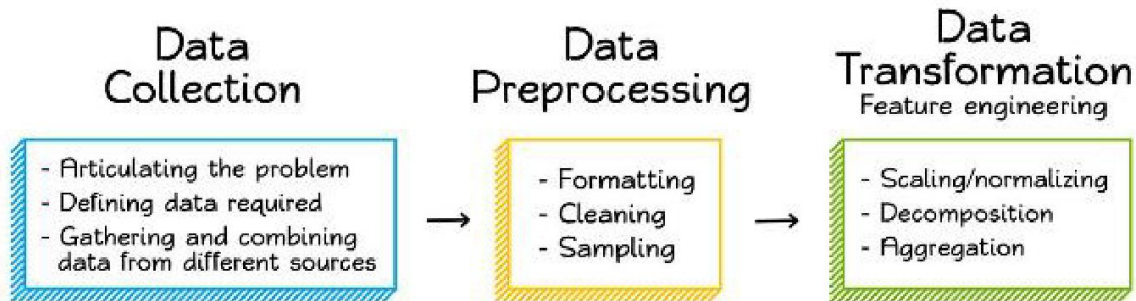


# Data Preparation dalam CRISP-DM



# Tahapan Data Preparation versi simpel

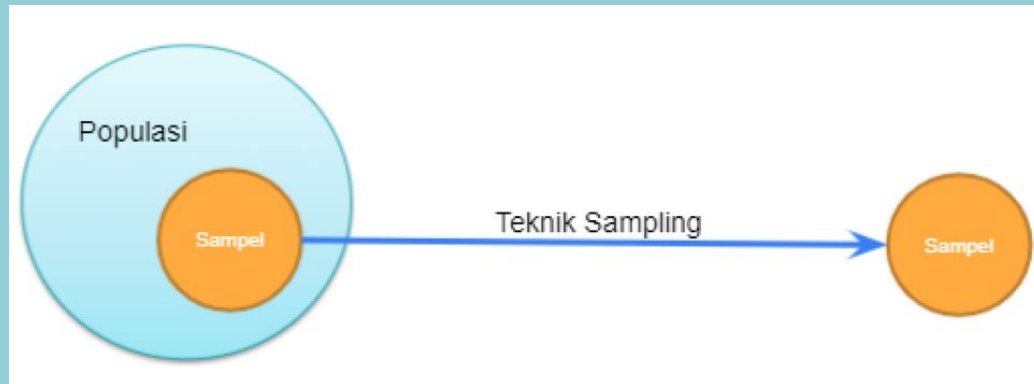
## Data Preparation Process



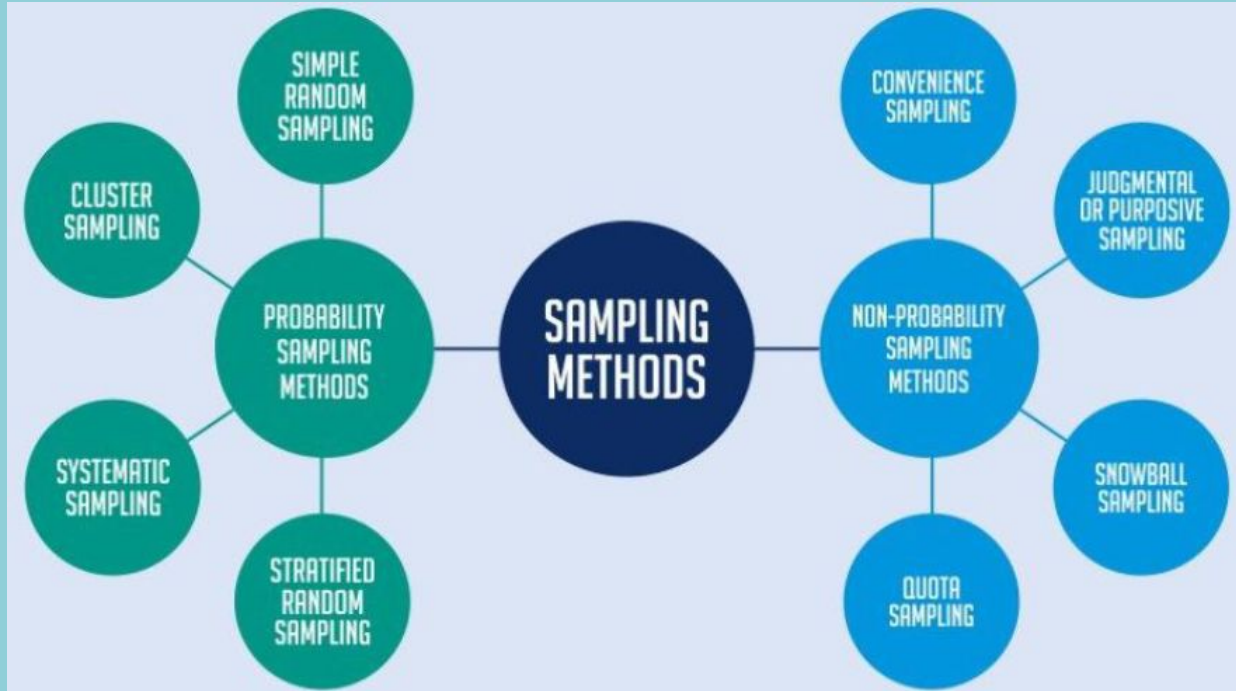
# Sampling Data

## Pengertian Sampling

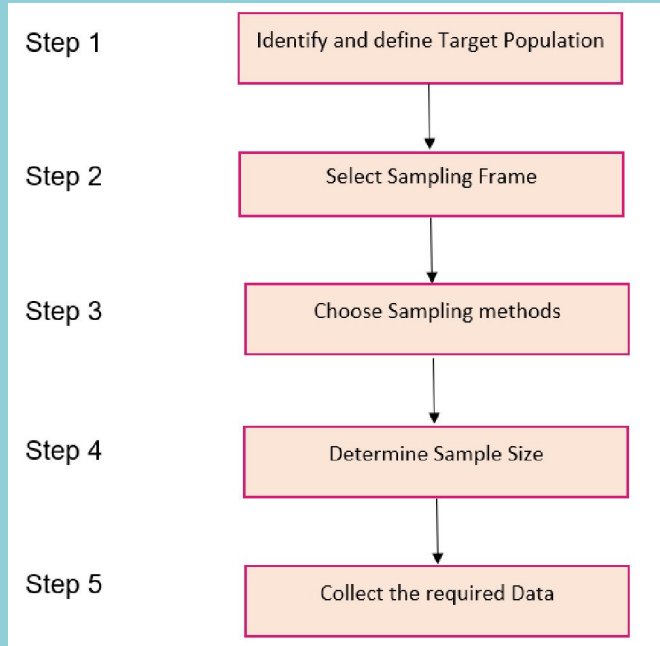
Sebelum melakukan tahapan dalam data preparation, terlebih dahulu adalah pemilihan/penentuan objek yang dapat dilakukan dengan menggunakan penentuan POPULASI dan SAMPEL



# Metode Sampling



# Tahapan Sampling



1. Identifikasi dan mendefinisikan target dari populasi yang akan digunakan
2. Memilih sampel frame dari populasi yang digunakan
3. Memilih metode sampling mana yang akan digunakan
4. Menentukan ukuran sample
5. Mengumpulkan Data yang dibutuhkan



# Imbalance Dataset

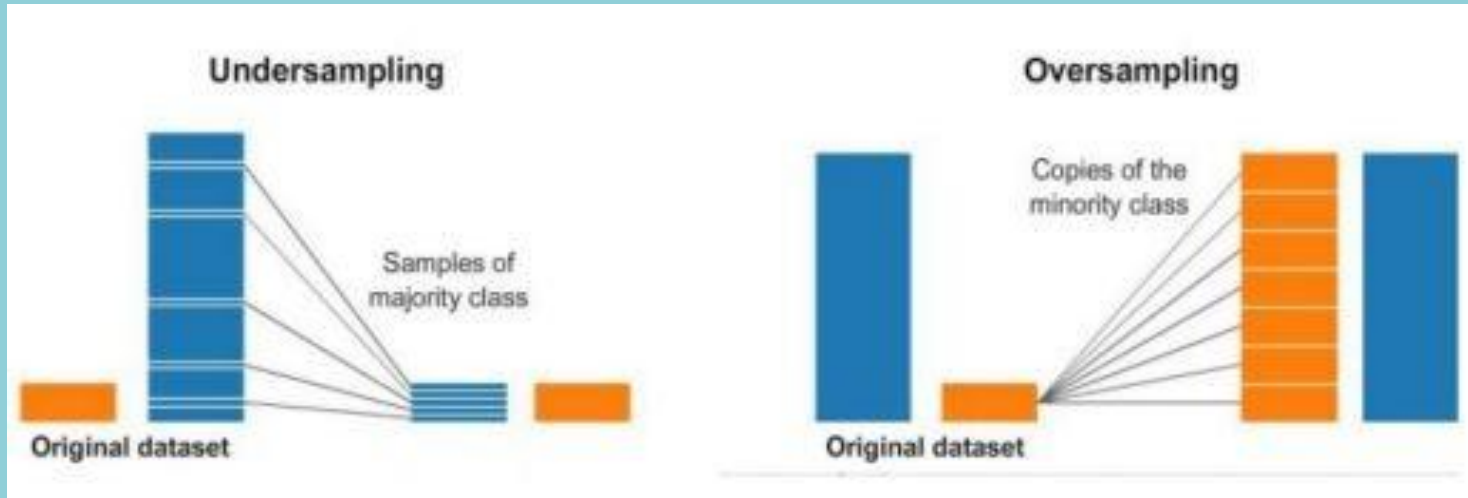
Imbalance Dataset merupakan data yang biasanya diolah secara klasifikasi dengan salah satu kelas/label pada datanya mempunyai nilai yang sangat jauh berbeda jumlahnya dari kelas lainnya.

Pada Imbalance dataset, biasanya memiliki data dengan kelas yang sedikit (rare class) dan data dengan kelas yang banyak (abundant class)

Salah satu untuk mengatasi imbalance dataset adalah RESAMPLING

# Imbalanced

Data yang memiliki rasio yang tidak berimbang antara data satu dengan data lainnya dapat dikatakan sebagai imbalanced. Dengan begitu dataset harus dibuat balance dengan hanya menggunakan **Training Dataset**.



# RESAMPLING



Gunakan pengukuran (metrik) yang tepat, misalnya dengan menggunakan :

Precision/Spesifikasi : berapa banyak instance yang relevan

Recall / Sensitivitas : berapa banyak instance yang dipilih

F1-Score : harmonisasi mean dari precision dan recall

MCC : koefisien korelasi antara klasifikasi biner antara observasi dan prediksi

AUC : relasi antar tingkat true-positive vs false-positive

Resampling Data Training, dengan dua metode

Undersampling

Oversampling



## Melakukan Resampling

### 1. Over-Sampling

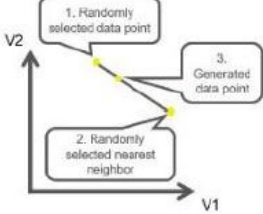
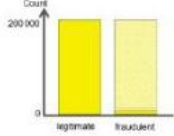
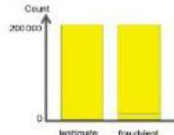
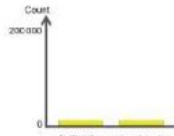
Melakukan generate pada rare class sehingga jumlah dari rare class sama dengan jumlah abundant class.

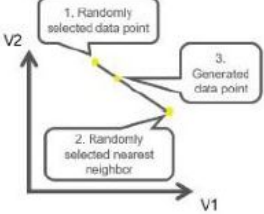
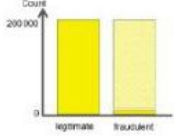
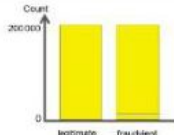
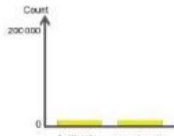
### 2. Under-Sampling

Melakukan seleksi pada abundant class secara acak/ random sehingga abundant class nilainya berkurang sampai dengan jumlahnya sama dengan rare class.



# TEKNIK RESAMPLING

Resampling method	Description	Target class distribution after resampling
Oversampling (SMOTE)	<p>Generate new synthetic fraudulent transactions until the number of fraudulent transactions is ca. equal to the number of legitimate transactions:</p> <ol style="list-style-type: none"> <li>1. Select one of the fraudulent transactions in the training data randomly</li> <li>2. Select one of its <math>n</math> nearest neighbors in the same fraudulent class randomly</li> <li>3. Select a random point between the existing fraudulent transaction and its nearest neighbor</li> </ol> 	<ul style="list-style-type: none"> <li>• Original data in yellow</li> <li>• New synthetic data in light patterned yellow</li> </ul> 
Oversampling (Bootstrap)	Randomly draw with replacement a sample of fraudulent transactions until the number of fraudulent transactions is ca equal to the number of legitimate transactions	
Undersampling (Bootstrap)	Randomly draw with replacement as many legitimate transactions as there are fraudulent transactions	

Resampling method	Description	Target class distribution after resampling
Oversampling (SMOTE)	<p>Generate new synthetic fraudulent transactions until the number of fraudulent transactions is ca. equal to the number of legitimate transactions:</p> <ol style="list-style-type: none"> <li>1. Select one of the fraudulent transactions in the training data randomly</li> <li>2. Select one of its <math>n</math> nearest neighbors in the same fraudulent class randomly</li> <li>3. Select a random point between the existing fraudulent transaction and its nearest neighbor</li> </ol> 	<ul style="list-style-type: none"> <li>• Original data in yellow</li> <li>• New synthetic data in light patterned yellow</li> </ul> 
Oversampling (Bootstrap)	Randomly draw with replacement a sample of fraudulent transactions until the number of fraudulent transactions is ca equal to the number of legitimate transactions	
Undersampling (Bootstrap)	Randomly draw with replacement as many legitimate transactions as there are fraudulent transactions	

# PEMILIHAN (SELEKSI FITUR) DATA

## Manfaat seleksi fitur Data

Reduksi Overfitting

Meningkatkan Akurasi

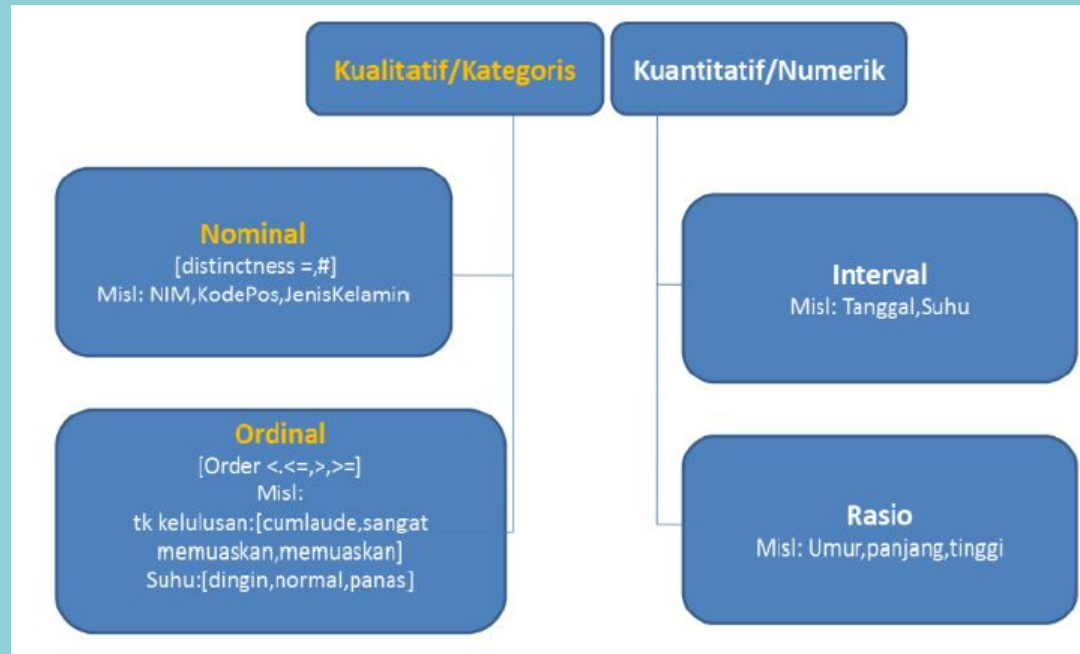
Reduksi Waktu Training

## Jenis seleksi fitur data

Unsupervised

Supervised

# Membedakan Jenis Data

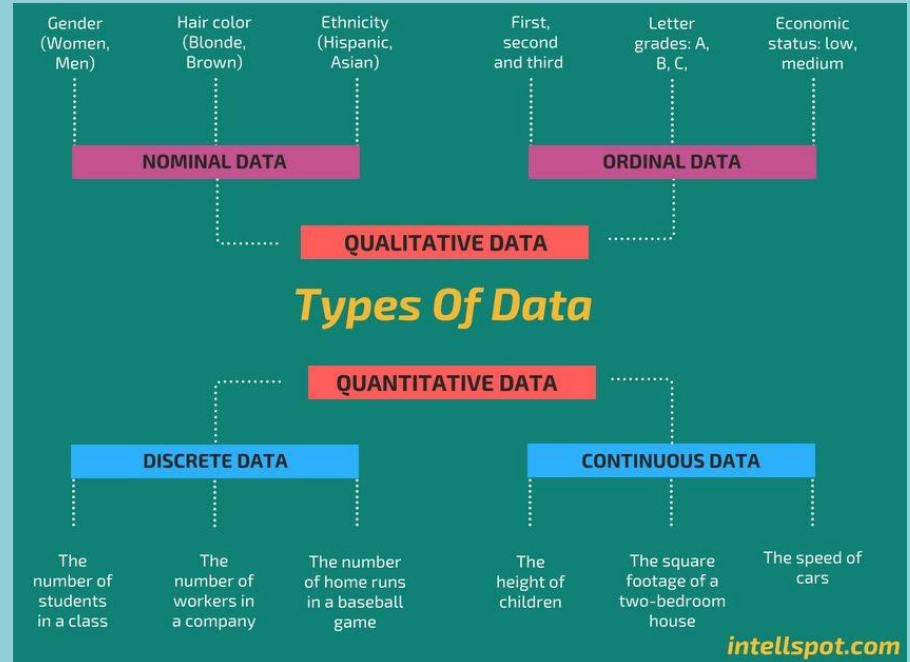
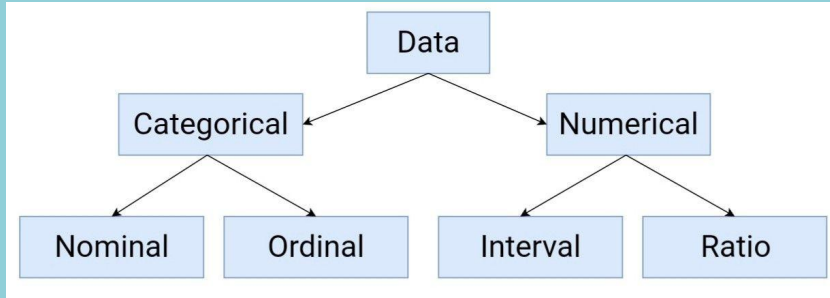


# Data Type

Tipe data adalah konsep yang sangat penting dalam machine learning. Dengan mengetahui tipe data akan memudahkan dalam proses data preprocessing.

Kenapa tipe data penting:

- Untuk mengaplikasikan pengukuran statistic ke data dengan benar
- Menyimpulkan dengan benar asumsi tertentu dari dataset



# Validasi Data



Harus dapat membedakan antara **VERIFIKASI** vs **VALIDASI**

**VERIFIKASI** bertujuan untuk membuktikan bahwa sesuatu ada atau benar , atau untuk memastikan bahwa sesuatu adalah benar atau salah.

**VALIDASI** bertujuan untuk membuat sesuatu yang resmi diterima atau disetujui, terutama setelah memeriksanya apakah Kuat atau Lemah

VALIDASI atau VALIDITAS adalah mengukur sejauh mana perbedaan skor yang mencerminkan perbedaan sebenarnya baik itu antar individu, kelompok atau juga situasi yang mengenai karakteristik yang akan diukur atau juga kesalahan sebenarnya pada individu ataupun juga kelompok yang sama dari satu situasi ke situasi yang lain .

# Apa yang di Validasi ?

Tipe Data  
Range Data  
Uniqueness  
Consisten Expression

Format Data  
Nilai Null/Missing Values  
Misspelling/Type  
Invalid Data



# Teknik Validasi

Akurasi atau ketepatan data yang ada  
Kelengkapan data  
Konsistensi data  
Ketepatan waktu validasi data

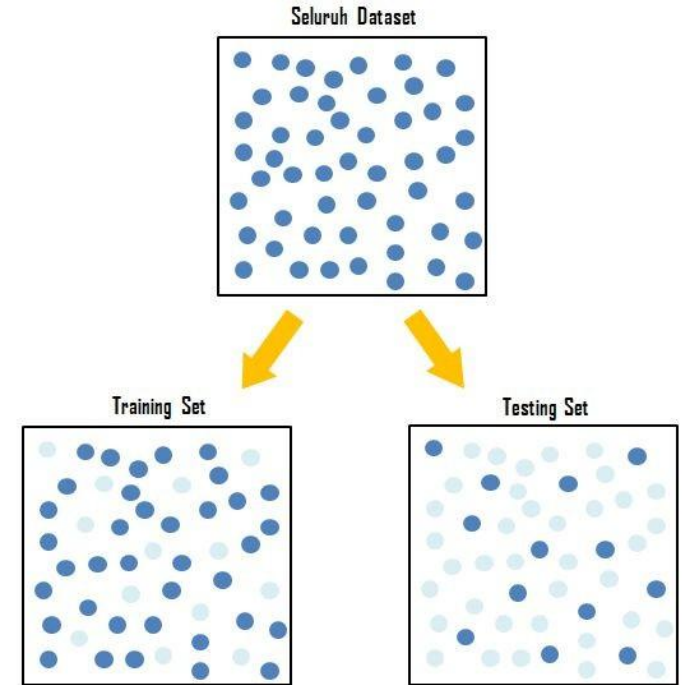
Kepercayaan  
Nilai Tambah  
Penafsiran data  
Kemudahan Akses






# Train Test Split

Train/test split adalah salah satu metode yang dapat digunakan untuk mengevaluasi performa model machine learning. Metode evaluasi model ini membagi dataset menjadi dua bagian yakni bagian yang digunakan untuk training data dan untuk testing data dengan proporsi tertentu.





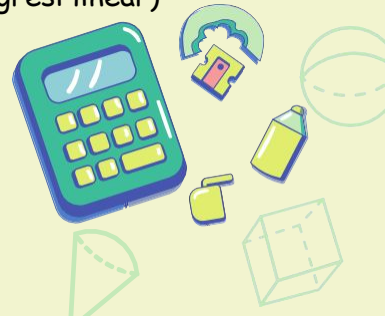
# Feature Scaling

Kenapa harus melakukan feature scaling:

- 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 menggunakan feature scaling:

- Gunakan feature scaling jika model ML yang digunakan terpengaruhi oleh skala data (KNN, Logistic Regression, SVM)
- Gunakan Standardization bila tahu bahwa data memiliki sebaran normal/Gaussian
- Gunakan Standardization bila model yang kita pakai punya asumsi tentang normalitas (e.g. regresi linear)
- Gunakan normalization apabila tidak memenuhi 2 kriteria di atas.



# Data Preprocessing Berdasarkan Tipe Data

# Preprocessing Data pada Data Teks

- Case Folding

Proses mengkonversi tulisan menjadi “uppercase” dan “lowercase”.

- Stopword Removal

Menghilangkan kata-kata yang dianggap tidak berpengaruh secara signifikan di dalam data teks.

- Stemming/ Lemmatization

Proses merubah kata menjadi bentuk akar kata.

Contoh : membuat -> buat, menulis -> tulis, dsb

- Slangword Handling

Mengatasi kata-kata non-formal, kata-kata sehari-hari, singkatan ataupun kata gaul yang ada di dalam teks dengan merubahnya ke bentuk kata formal.

Contoh : Yg -> yang, krn -> karena, dll.

- Feature Extraction

Proses merubah kata dan kalimat ke dalam bentuk vektor representatif. Beberapa metode yang cukup sering digunakan diantaranya TF-IDF, word embedding, skip gram, dll.

# Preprocessing Data pada Data Gambar

- Changing Color-space

Biasanya digunakan untuk merubah jenis channel seperti BGR  $\leftrightarrow$  Gray, BGR  $\leftrightarrow$  HSV, dll.

- Geometric Transformations

Beberapa transformasi yang dapat dilakukan seperti : translation, rotation, affine transformation, dll

- Morphological Transformations

Melakukan berbagai operasi pada gambar seperti Erosion, Dilation, Opening, Closing, dll untuk berbagai kebutuhan tertentu.

- Image cropping

- Resize image

- dll

Python image processing :

[https://docs.opencv.org/4.x/d6/d00/tutorial\\_py\\_root.html](https://docs.opencv.org/4.x/d6/d00/tutorial_py_root.html)

Rotation



Opening



Original Image



Erosion



Dilation



Closing



## Preprocessing Data pada Data Audio

- Spectral features

Dalam data audio, spectral features merepresentasikan energi frekuensi yang berubah-ubah nilainya pada setiap satuan waktu.

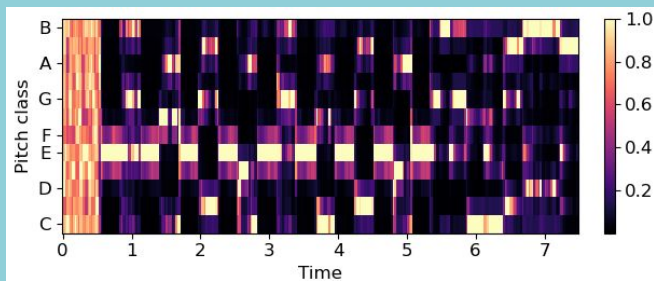
- Rhythm features

Preprocessing yang dilakukan untuk mengekstrak informasi terkait ritme dan juga tempo audio.

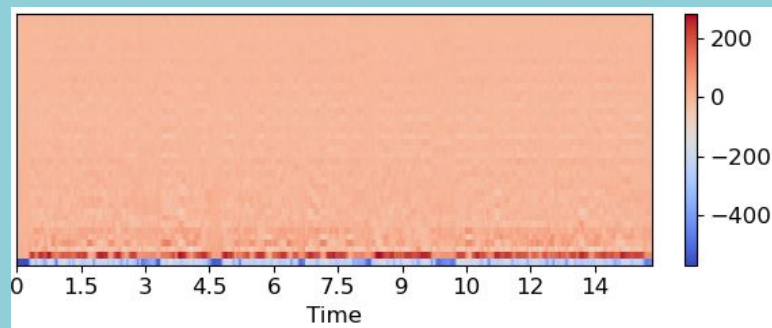
- Audio cropping

Python audio processing :  
<https://librosa.org/doc/latest/index.html>

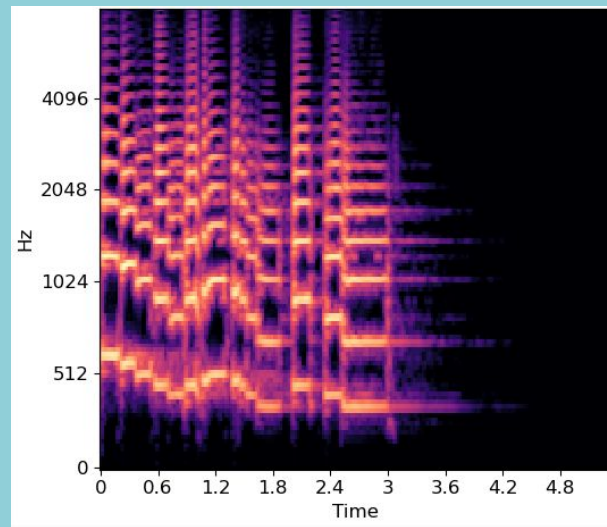
Chromagram



MFCC



Spectrogram





## Ringkasan

Data preparation memiliki sebutan lain, diantaranya pre-processing, data cleaning, data manipulation

Data preparation mengambil porsi kerja terbanyak dalam data science sebesar 60%-80%.

Data preparation membutuhkan ketelitian dan kesabaran/kerajinan dari peneliti Data Science, terutama pemula.

Data Validation merupakan tahapan kritis dari Data Science 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 Data Science terkait penggunaan tools dan coding.

Kebersihan Data merupakan syarat mutlak untuk model prediksi yang baik