

# Mastering Preprocessing Data

Data Engineering Week 4

25 Mei 2023



**Whisnumurty Galih Ananta**

Lead of Data Engineering, CODER TUS



**Whisnumurty Galih Ananta**

Lead of Data Engineering at **<coder>**  
TELKOM UNIVERSITY SURABAYA

“Passionate Data Science and Machine Learning enthusiast, dedicated to continuous learning and growth by actively contributing to challenging projects”

- Machine Learning Cohort at Bangkit , Feb 2024 - Now
- Lead of Data Engineering at CODER, Des 2022 - Now
- Core Team at GDSC Tel-U, Aug 2023 - Now

Find Me:



[github.com/whisnumurtyga](https://github.com/whisnumurtyga)



[linkedin.com/in/whisnumurtyga](https://linkedin.com/in/whisnumurtyga)

# What ~~you~~ we will learn

1. What is Data Preprocessing
2. Why Preprocess Data?
3. What is Data Quality?
4. Dimensions of Data Quality
5. Data Preprocessing Pillar
6. Data Preprocessing Steps
7. Data Preprocessing Best Practices
8. Importance of Data Preprocessing
9. Data Preprocessing Hands On
10. Data Preprocessing in Machine Learning Competition

# What is Data Preprocessing

Data preprocessing is a **critical step** in the data science **process**, and it often **determines** the **success or failure** of a **project**.

Real-world data are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources

**While providing raw data**

**While going through findings and output**



# Why Preprocess the Data?

The **goal** of preprocessing data is to **ensure data quality**. Low-quality data will **lead** to low-quality mining results.



# What is Data Quality

Data quality **measures** how well a dataset meets **criteria**. There are many factors comprising data quality.

These include:

- 1. Accuracy
- 2. Completeness
- 3. Consistency
- 4. Uniqueness
- 5. Timeliness
- 6. Validity
- 7. Integrity
- 8. Precision
- 9. Currency
- 10. Conformity



# Dimensions of Data Quality

- **Accuracy**

This dimension refers to the correctness of the data values based on the agreed upon “source of truth.” It’s important to designate a primary data source; other data sources can be used to confirm the accuracy of the primary one.

- **Completeness**

This represents the amount of data that is usable or complete. If there is a high percentage of missing values, it may lead to a biased or misleading analysis if the data is not representative of a typical data sample.

- **Consistency**

Consistent data can be explained as how close your data aligns or is in uniformity with another dataset. Using different sources to check for consistent data trends and behavior.

- **Uniqueness**

The occurrence of an object or an event gets recorded multiple times in a dataset. For example, when reviewing customer data, you should expect that each customer has a unique customer ID.

- **Timeliness**

This dimension refers to the readiness of the data within an expected time frame. For example, customers expect to receive an order number immediately after they have made a purchase, and that data needs to be generated in real-time.

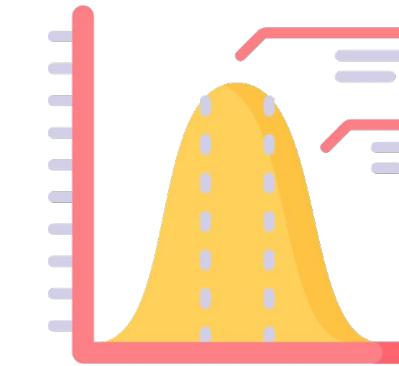
- **Validity**

How much data matches the required format for any business rules. Formatting usually includes metadata, such as valid data types, ranges, patterns, and more.

# Data Preprocessing Pillar



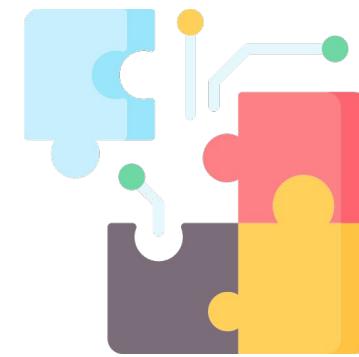
Data  
Cleaning



Data  
Transformation



Data  
Reduction



Data  
Integration



Data  
Augmentation

# Data Cleaning

This involves **detecting** and rectifying errors, **inconsistencies**, and **anomalies** within the dataset.

Common techniques include handling missing values, removing duplicates, and correcting inaccuracies/ outlier.

	Nama	Usia	Negara Asal	Kapal	Jumlah Harta Karun (USD)
0	Blackbeard	45.0	England	Queen Anne's Revenge	500000.0
1	Anne Bonny	32.0	Ireland	Unknown	300000.0
2	Jack Sparrow	NaN	Caribbean	Black Pearl	1000000.0
3	Barbarossa	55.0	Spain	Queen Anne's Revenge	700000.0
4	Ghost	100000.0	None	Unknown	NaN
5	Anne Bonny	32.0	Ireland	Unknown	300000.0



FB: @Statisticsss

When I started  
cleaning Data



When I finished  
cleaning Data

IG: @StatisticsForYou

# Handling Missing Values

## Type of Missing Values

- Missing Completely At Random (MCAR)

This type of missing data is purely random and lacks any discernible pattern.

ex: data yang hilang mungkin terjadi karena masalah teknis saat memasukkan data survei, seperti kegagalan sistem komputer.

- Missing at Random (MAR)

There is a pattern in the missing values, but this pattern can be explained by other observed variables.

ex: Data kesehatan mental yang hilang hanya berasal dari responden berusia lanjut tanpa hubungan langsung dengan status kesehatan mental.

- Missing Not At Random (MNAR)

This type of missing data has a specific pattern that cannot be explained by observed variables.

ex: responden dengan gaji tinggi enggan memberikan informasi gaji dalam survei, mempengaruhi keberadaan data yang hilang

Understanding the type of missing data is crucial because it determines the appropriate strategy for handling missing values and ensuring the integrity of statistical analyses.

## How to Handle

1. Deletion

This involves removing rows or columns with missing values. This is a straightforward method, but it can be problematic if a significant portion of your data is missing.

2. Imputation

- Mean/Median/Mode

values can be replaced with the mean, median, or mode of remaining values in the column.

- Last/Baseline Observation Carried Forward (L/B-OCF)  
the last/first known value from the same subject is used to fill in missing values for subsequent observations.

- Prediction of Missing Value

- KNN Imputer (Univariate)
- Iterative Imputer (Multi Variate)
- Deep Learning Library (Datawig)

- Constant Value (MCAR)

replace the missing values with the defined constant values  
ex: -9999, unknown

# Handling Outliers

## Type of Outlier

- **Global Outliers (Point Anomalies)**

A value is considered a global outlier if it is far outside the entire data set in which it is found.

ex: People with a height of 1 cm

- **Conditional Outliers**

A value is considered a contextual outlier if the value is within a certain context, but not as a whole.

ex: People with a height of 215 cm (possible, but rare in Asia)

- **Collective Outliers**

Subset of data points deviating significantly from the overall pattern without individual anomalies.

ex: Sales data for stores with stable sales patterns that suddenly have significantly different total sales on a particular day,

## How to Detect an Outlier

- IQR Method
- Z-Score Method
- Boxplot Analysis

If outliers have been identified, the next thing that needs to be done is to determine whether the identified data needs to be deleted, modified, or used as new insight in data analysis. Consider Dataset Size & Existence of Influential Observations (Model/ Analysis) Result

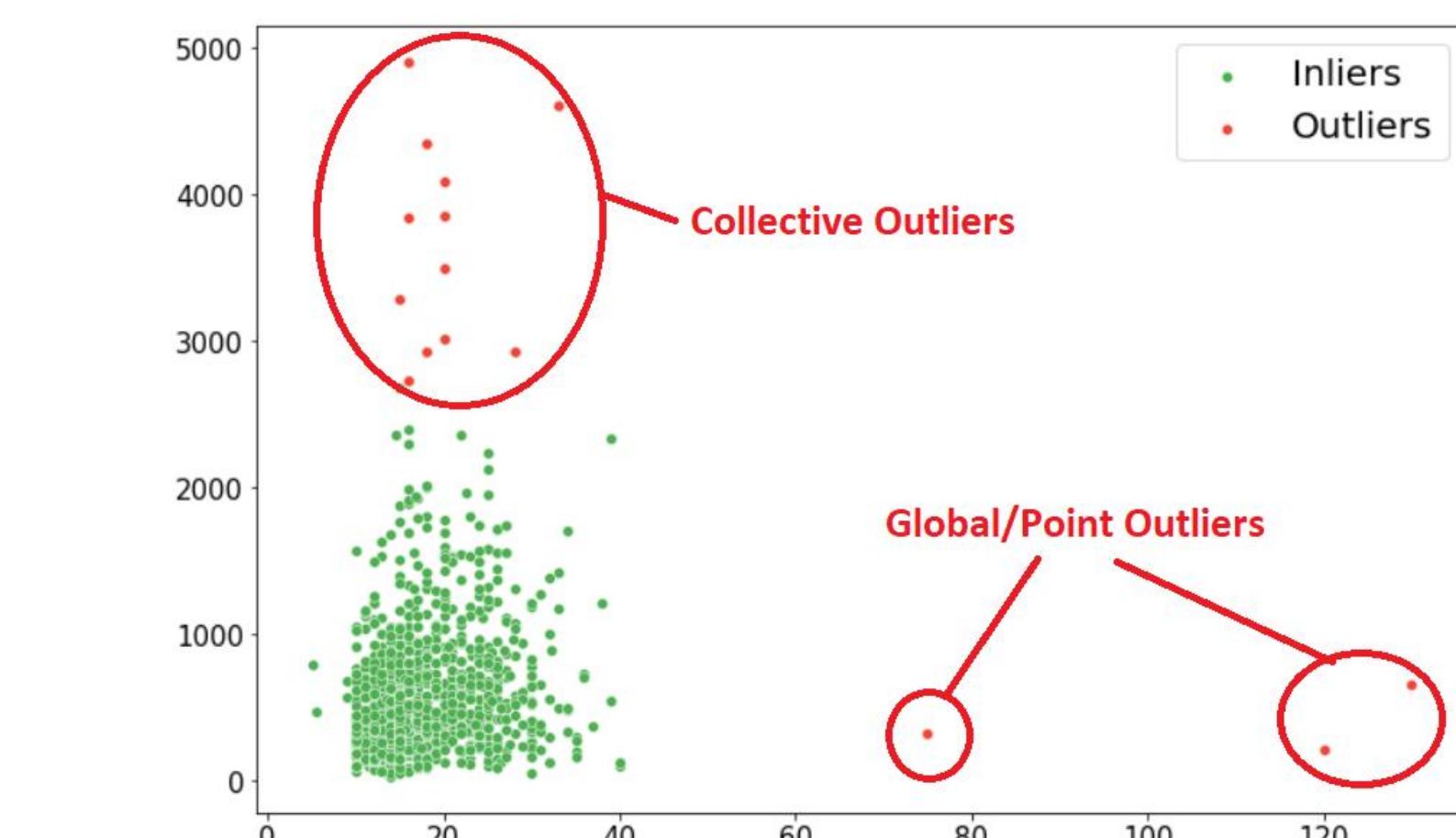
## How to Handle Outlier

- Remove Outlier

- Change the Scale (Normalization)

- Imputation

- Quantile Based Flooring and Capping (Define min/max threshold)
- Mean/Median/Mode

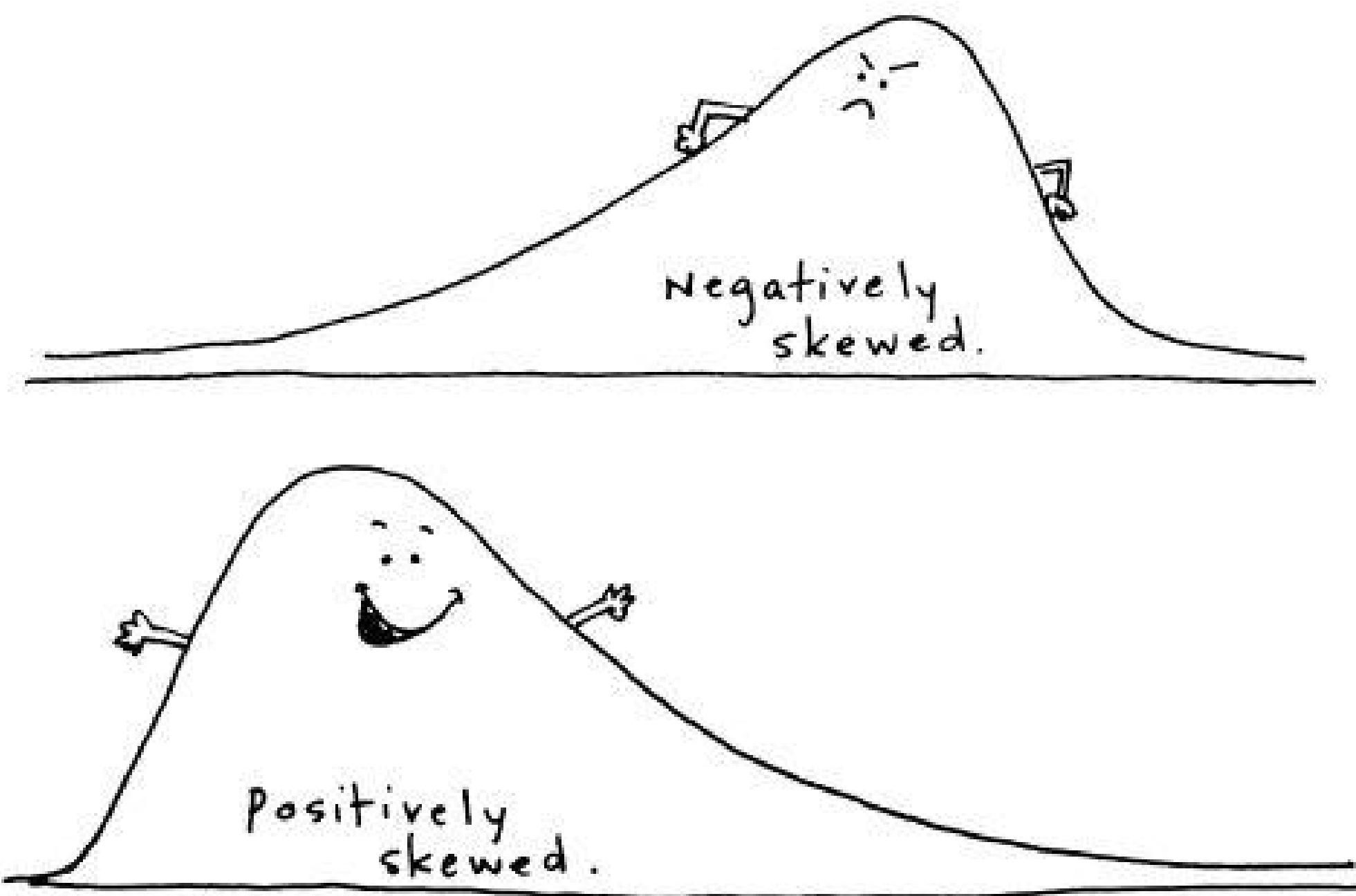


# Data Transformation

Data often comes in various formats and scales, necessitating standardisation or normalisation to ensure uniformity and comparability across features.

Transformation techniques such as **scaling**, **encoding categorical variables**, and **feature engineering** play a pivotal role in this stage.

	Nama	Umur	Tanggal Lahir	Gaji	Aktif	Penilaian	Review Date	Kode Pos
0	Andi	23	2000-05-17	5000000	True	3.9	2022-04-01 11:24:41.075660	12345
1	Budi	30	1993-08-25	7000000	False	4.2	2021-05-17 11:24:41.076756	67890
2	Citra	25	1998-03-14	6000000	True	4.5	2016-09-27 11:24:41.076756	54321
3	Dewi	22	2001-01-30	5500000	True	3.8	2015-12-26 11:24:41.076756	09876
4	Eka	28	1995-06-05	6500000	False	4.0	2011-06-09 11:24:41.076756	11223



# Data Transformation

## Numeric Transformation

- Normalisation

Scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.

change num values to maintain a consistent scale without distorting differences or losing information.

- Scaling

Scaling is a method used to normalise the range of independent variables or features of data within a particular range.

scaling doesn't necessarily preserve the original distribution or relative differences between values.

- Binning

Binning is used for the transformation of a continuous or numerical variable into a categorical feature.

It helpfull, but the bin sizes need to be chosen correctly.

## Encoding Categorical

- One-Hot Encoding

warna	biru	hijau	merah
0 merah	0.0	0.0	1.0
1 hijau	0.0	1.0	0.0
2 biru	1.0	0.0	0.0
3 merah	0.0	0.0	1.0
4 biru	1.0	0.0	0.0

- Label Encoding

```
['merah', 'hijau', 'biru', 'merah', 'biru']
[2 1 0 2 0]
```

- Ordinal Encoding

```
1 # Definisikan urutan kategori
2 urutan_kategori = ['merah', 'hijau', 'biru']
3 kategori = ['merah', 'hijau', 'biru', 'merah', 'biru']
4 kategori_array = np.array(kategori).reshape(-1, 1)
5
6 # Buat objek OrdinalEncoder dengan spesifikasi urutan
7 ordinal_encoder = OrdinalEncoder(categories=[urutan_kategori])
8
9 # Terapkan Ordinal Encoding pada data kategori
10 kategori_ordinal_encoded = ordinal_encoder.fit_transform(kategori_array)
```

## Feature Engineering

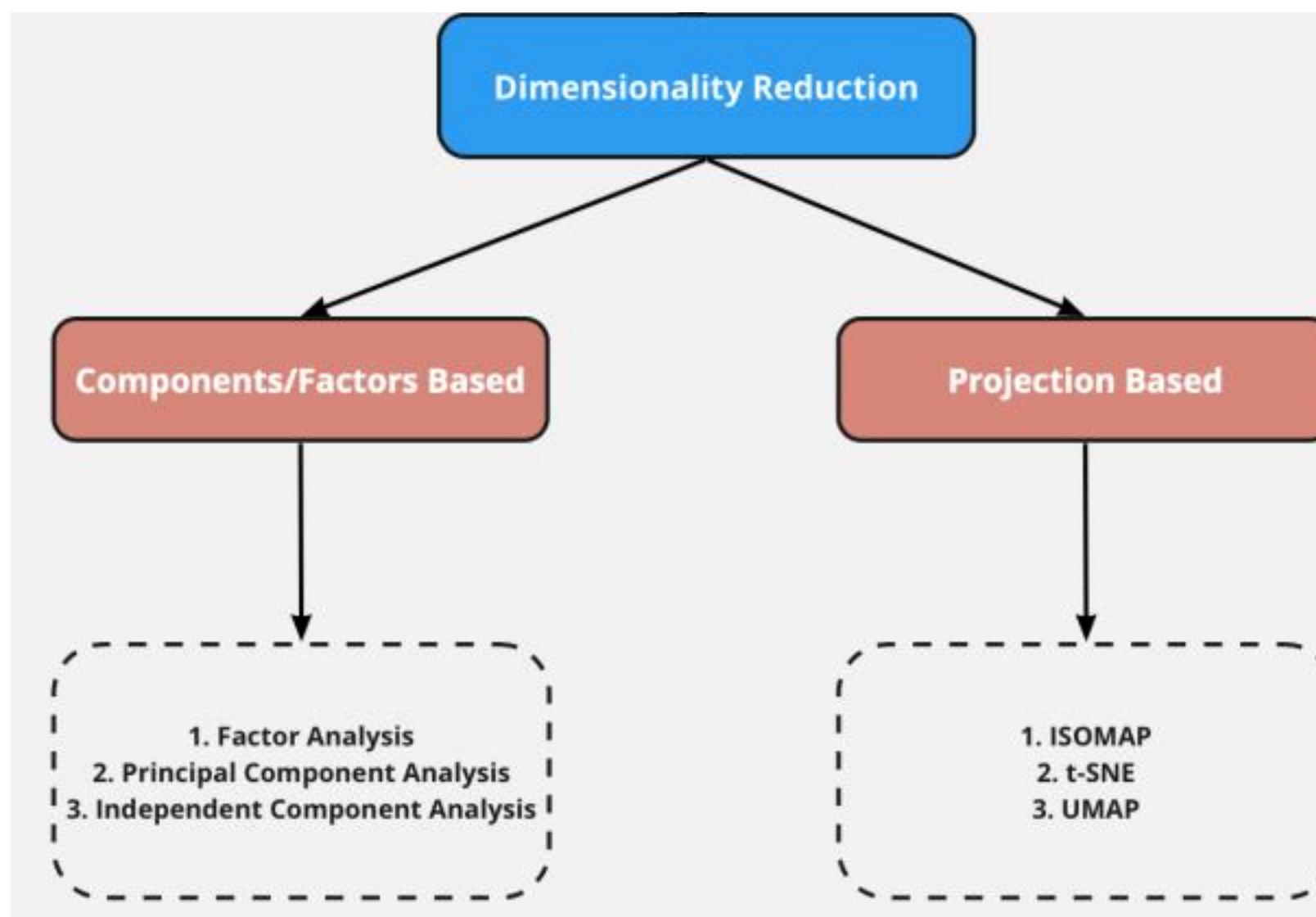
Process for developing and selecting features or attributes that will be used to carry out data analysis or create a machine learning model.

belanja_buah	belanja_daging	belanja_ikan	belanja_kue
50575.0	260967.0	50575.0	20230.0
6069.0	44506.0	80920.0	20230.0
117611.0	265460.0	96341.0	145573.0
206346.0	1613901.0	27725.0	125868.0
90563.0	311757.0	40358.0	33875.0

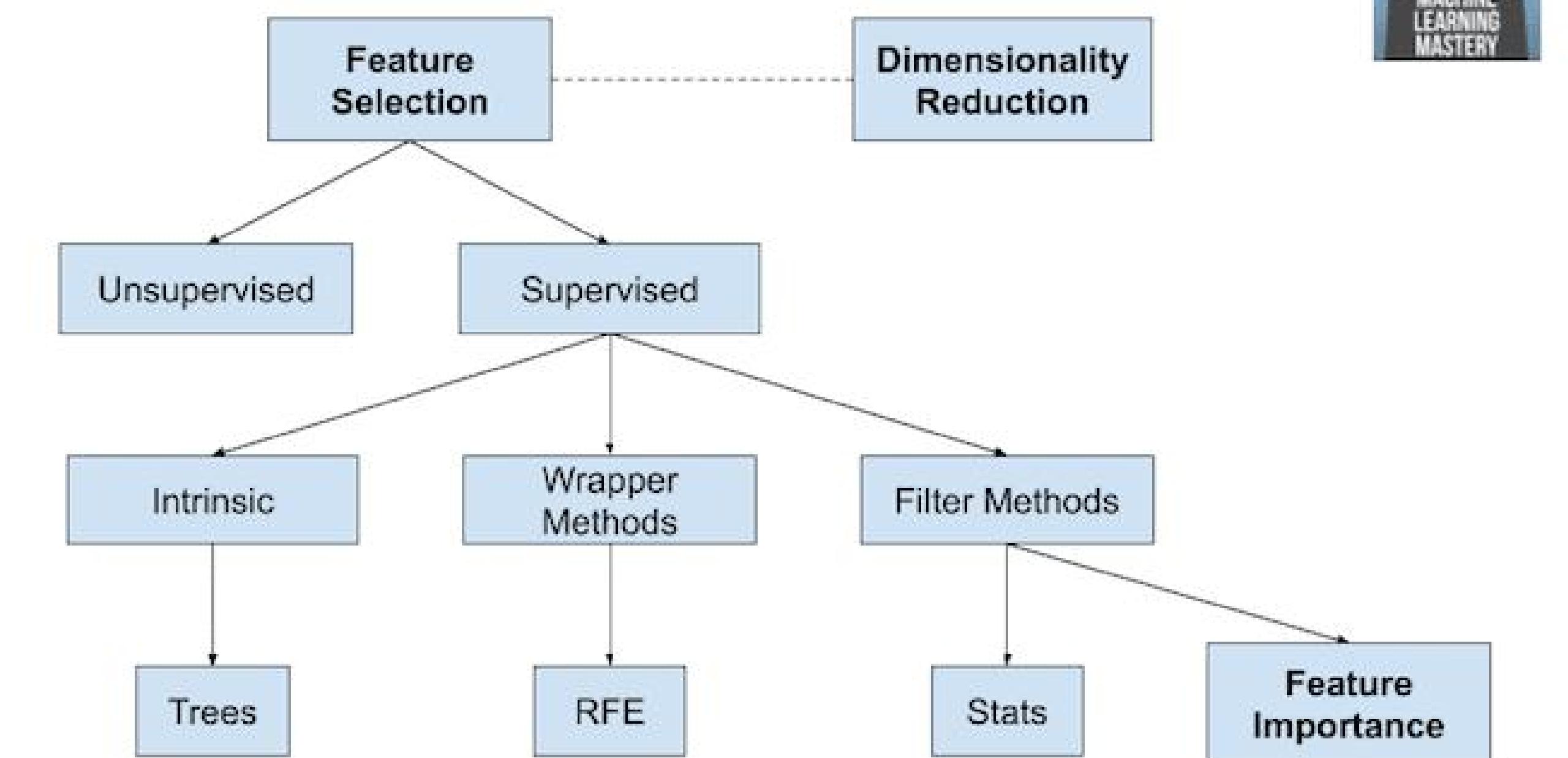
# Data Reduction

In instances where the dataset is excessively large or contains redundant features.

data reduction techniques like dimensionality reduction or feature selection can streamline the dataset without sacrificing critical information.



## Overview of Feature Selection Techniques

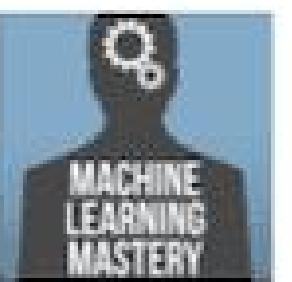


feature  
importance

recursive  
feature elimination

chi square  
fisher score

Copyright © MachineLearningMastery.com



MACHINE  
LEARNING  
MASTERY

# Data Integration

Integrate data from multiple sources or datasets by combining them based on common identifiers or keys, ensuring consistency and completeness in the merged dataset.

	tanggal	ot	eventID	datetime	latitude	longitude	location
0	2008-11-01	00:31:25.143	bmg2008vkye	2008-11-01 00:31:25.143741+00:00	-0.604440	98.895531	Southern Sumatra, Indonesia
	latitude	longitude	magnitude	mag_type	depth_y	phasecount	azimuth_gap
1	-6.611860	129.387220	5.507549	mb	30	62.0	45.46
2	-3.650586	127.990680	3.539674	MLv	5	4.0	331.97
3	-4.198925	128.097000	2.424314	MLv	5	5.0	326.37
4	-4.091891	128.200470	2.410045	MLv	10	5.0	314.65

Unnamed: 0	eventID	datetime	latitude	longitude	magnitude	mag_type	depth	phasecount	azimuth_			
0	bmg2008vkye	2008-11-01 00:31:25.143741+00:00	-0.604440	98.895531	2.989742	MLv	20	6.0	146.70			
1	bmg2008vlag	2008-11-01 01:34:29.660856+00:00	-6.611860	129.387220	5.507549	mb	30	62.0	45.46			
2	bmg2008vlaj	2008-11-01 01:38:14.802129+00:00	-3.650586	127.990680	3.539674	MLv	5	4.0	331.97			
3	bmg2008vlbt	2008-11-01 02:20:05.909515+00:00	-4.198925	128.097000	2.424314	MLv	5	5.0	326.37			
4	bmg2008vlcd	2008-11-01 02:32:18.756155+00:00	-4.091891	128.200470	2.410045	MLv	10	5.0	314.65			
tgI	ot	lat	lon	depth	mag	remark	strike1	dip1	rake1	strike2	dip2	rake2
0	2008/11/01	21:02:43.058	-9.18	119.06	10	4.9	Sumba Region - Indonesia	NaN	NaN	NaN	NaN	NaN
1	2008/11/01	20:58:50.248	-6.55	129.64	10	4.6	Banda Sea	NaN	NaN	NaN	NaN	NaN
2	2008/11/01	17:43:12.941	-7.01	106.63	121	3.7	Java - Indonesia	NaN	NaN	NaN	NaN	NaN
3	2008/11/01	16:24:14.755	-3.30	127.85	10	3.2	Seram - Indonesia	NaN	NaN	NaN	NaN	NaN
4	2008/11/01	16:20:37.327	-6.41	129.54	70	4.3	Banda Sea	NaN	NaN	NaN	NaN	NaN

# Data Augmentation

Generate synthetic samples or variations of existing data instances to augment the dataset, particularly in scenarios with limited data availability, to enhance model robustness and generalisation.

## Tabular

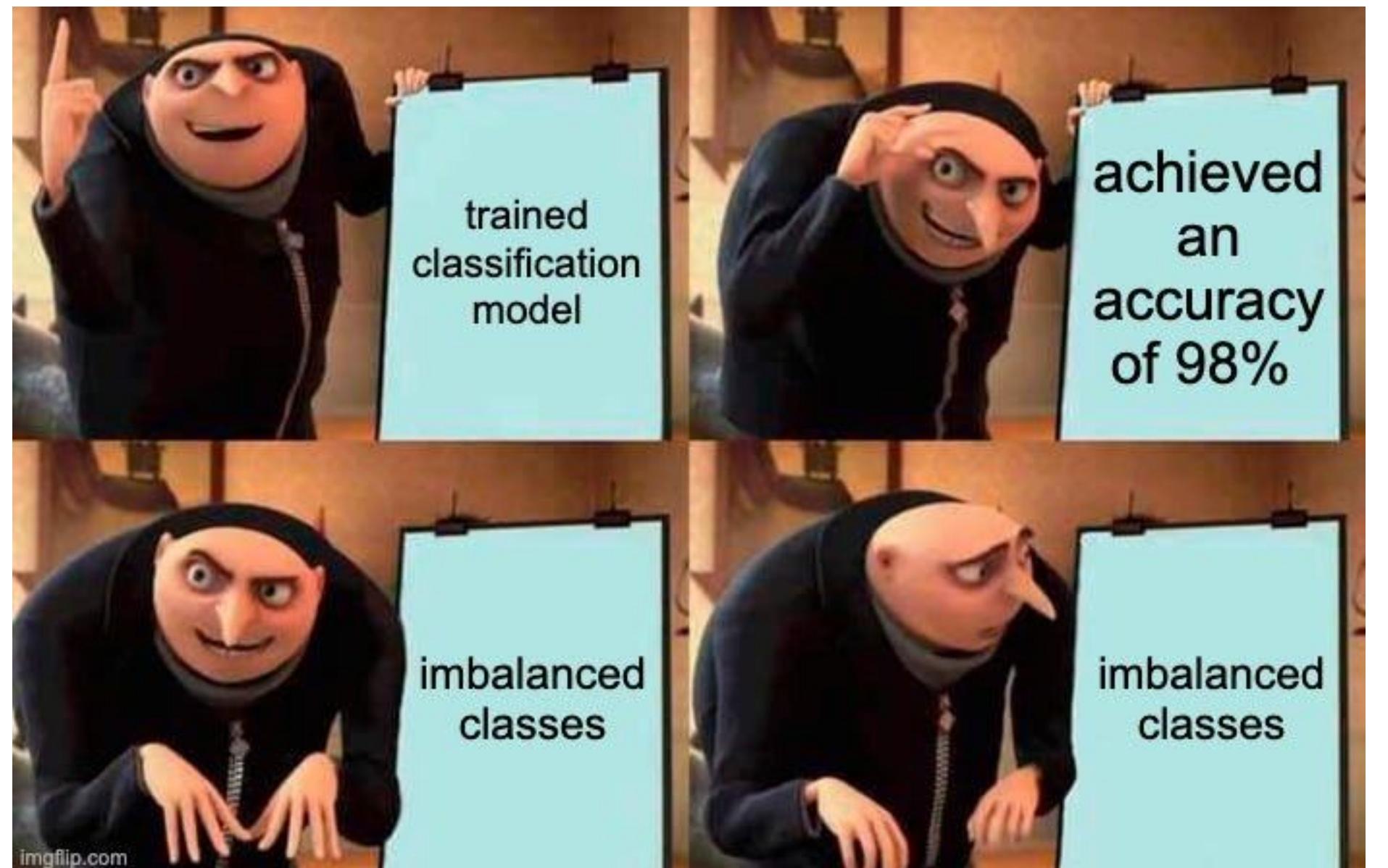
- Smote (Synthetic Minority Over-sampling Technique)
- Random Sampling

## Image

- Change Contras/ Rotation/ Brightness etc (modified image property)

## Text

- Retrieval Augmented Generation (RAG)



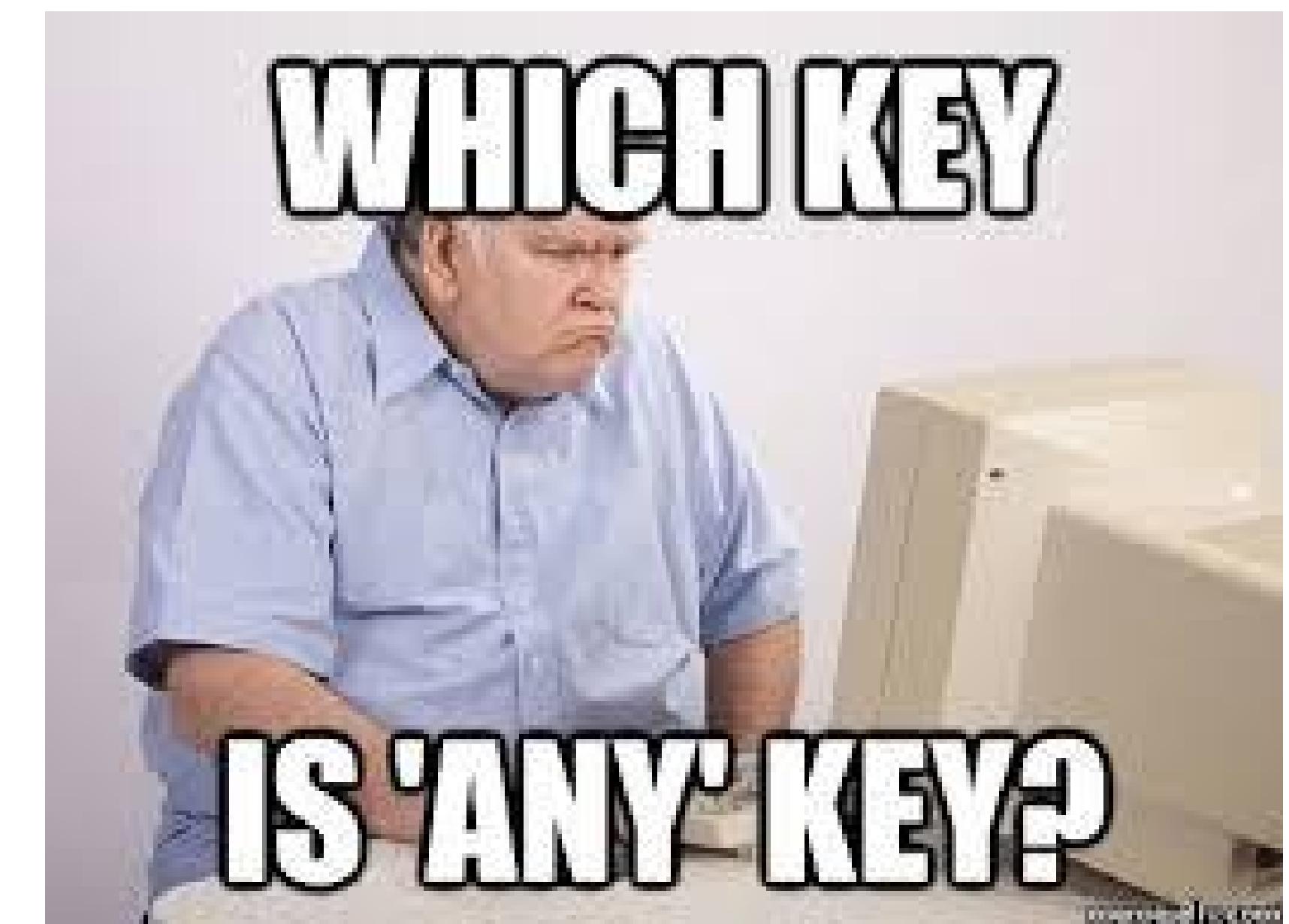
# Key of Data Processing

- **Data Understanding**
  - Type of Data
  - Structure
  - Features
  - Potential Issues

- **Trial and Errors**

Try any method you know and choose the best result

But when you understand the characteristics of each existing data preprocessing technique, you can save more time than trying repeatedly which wastes a lot of time



# Importance of Data Pre-processing

- **Enhanced Data Quality**

Raw data often contains errors, inconsistencies, and missing values that can distort analysis and lead to erroneous conclusions. Data preprocessing addresses these issues, ensuring the dataset is clean, accurate, and anomalies-free.

- **Improved Model Performance**

High-quality data is indispensable for training accurate and robust machine learning models. By preprocessing the data to remove noise, outliers, and irrelevant features, the model can focus on learning meaningful patterns

- **Facilitates Feature Engineering**

Feature engineering, the process of selecting, extracting, or creating relevant features from the raw data, is a crucial aspect of model development. Data preprocessing lays the groundwork for effective feature engineering by standardising, transforming, and encoding the features to make them suitable for analysis and model training.

- **Mitigates Overfitting and Underfitting**

Overfitting and underfitting in machine learning result from a mismatch between model complexity and available data. Data preprocessing helps balance this by ensuring models are trained on representative, well-structured datasets, reducing these risks.

- **Enables Data Exploration and Visualizations**

Preprocessed data is more amenable to exploratory data analysis and data visualisation, allowing data scientists to gain valuable insights into the underlying patterns, trends, and relationships within the data, visualising preprocessed data aids in identifying outliers, detecting correlations, and informing subsequent modelling decisions.

# Refference

<https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/>

<https://www.kaggle.com/discussions/general/451606>

<https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/>

<https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6d>

<https://medium.com/@vikakbary/outlier-definition-classification-handling-461037eb18be>

<https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e>

<https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/>

<https://medium.com/@myskill.id/feature-engineering-327d64277211>

<https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>

<https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/>

<https://timespro.com/blog/what-is-data-processing-know-the-importance-and-best-practices>

<https://medium.com/analytics-vidhya/data-augmentation-for-tabular-data-f75c94398c3e>

<https://www.blog.trainindata.com/mastering-data-preprocessing-techniques/>

<https://www.analyticsvidhya.com/blog/2021/05/feature-transformations-in-data-science-a-detailed-walkthrough/>

<https://medium.com/analytics-vidhya/a-guide-to-data-transformation-9e5fa9ae1ca3>

# Let's Hands On

# Penugasan

## Individu

What Awaits You 

 Series 01 · Data Science Pipeline

- [01-1 · What is Data Science?](#)
- [01-2 · Data Science Pipeline](#)

  Series 02 · Data Preprocessing

- [02-1 · Introduction](#)
- [02-2 · Missing Values](#)
- [02-3 · Categorical Encoding](#)
- [02-4 · Binning](#)
- [02-5 · Scaling](#)

1. Baca Section Series 01
2. Buat Notebook Duplicate Series 02,  
Jelaskan setiap code (Bisa pakai komen/  
markdown, gunakan bahasa yang singkat)  
menggunakan bahasa sendiri (Indonesia)

## Kelompok

1. Buka Churn Dataset yang dishare pada modul
2. Lakukan Preprocessing yang tidak terbatas pada (Handling Missing Values - Outlier, Data Transformation, Data Integration) dan cari minimal 5 Insight
3. Pengumpulan berupa folder drive yang berisi (notebook.ipynb, Project Report Docx (Data Team, Pembagian Tugas, Flow Preprocessing, Penjelasan Code, Insight/ Temuan)
4. Di Presentasikan Hasilnya di Next Meet (Boleh pakai PPT)

Silahkan buat kelompok secara mandiri yang beranggotakan 2-4 orang

# Absensi



<https://s.id/mdp-w4>

# Feedback Form



<https://s.id/ff-de23>

# Quote

*“Don't compare yourself with other people; compare yourself with who you were yesterday.”*

– Jordan Peterson



**Thank you**