# *Advance Data Preprocessing for Machine Learning*

# *Introduction to Machine Learning*

*Learning Progress Review Week 12*

By:
MARVEL TEAM
Fikrie | Natalia | Satria

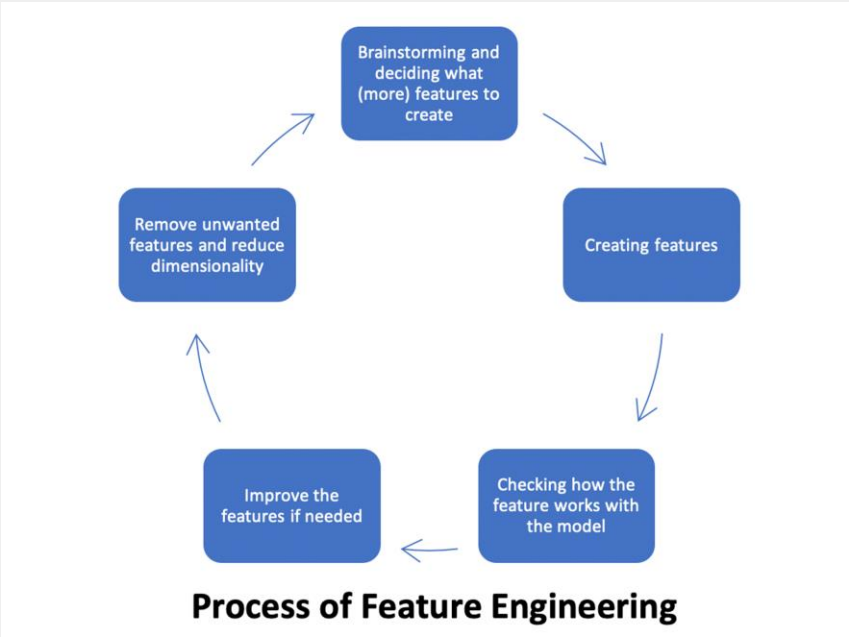
# 1. *Advance Data Preprocessing for Machine Learning*

# Feature Engineering

**Feature Engineering** adalah bagaimana kita menggunakan pengetahuan kita dalam memilih features atau membuat features baru agar model Machine Learning dapat bekerja lebih akurat dalam memecahkan masalah.

**Feature Engineering** akan memakan sebagian besar waktu kita saat membuat model Machine Learning. Jika kita melakukannya dengan baik, model yang di hasilkan akan mampu memprediksi atau memecahkan masalah lebih akurat.

# Feature Engineering




**Process of Feature Engineering**

Contoh:

- Menghitung BMI menggunakan tinggi dan berat.
- Rasio harga dan jumlah barang yang di jual.
- Selisih harga jual dan harga beli (keuntungan).

Handling Text Data

DigitalSkola
Uncover The World of Digital Skills with Us

# Cleaning

- Mengubah huruf capital menjadi huruf kecil (*Lowecase*).
- Menghapus karakter symbol atau tanda baca.
- Menghapus atau menerjemahkan angka menjadi sebuah text ( 1 -> One, 15 -> fifteen).
- Menghapus mentions (@person), hastag(#hastag) dan Url.
- Menghapus whitespace atau karakter kosong.

DigitalSkola
Uncover The World of Digital Skills with Us

# Tokenization

**Tokenizing** adalah proses *pemisahan* teks menjadi potongan-potongan yang idsebut sebagai token untuk kemudian di Analisa.

# Count Vectorizer

Untuk menggunakan data tekstual untuk pemodelan prediktif, teks harus di urai untuk menghapus kata-kata tertentu, proses ini disebut **tokenisasi**.

**Count Vectorizer** digunakan untuk mengonversi kumpulan dokumen teks menjadi vector jumlah istilah/token.

# TF-IDF Vectorization

**TF-IDF** adalah salah satu metode untuk memberikan *bobot term* dalam dokumen teks. Term disini bisa berupa kata. Bobot setiap term diperoleh dengan menghitung TF dan IDF

| | tf | | | | | | tf-idf | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| term | D1 | D2 | D3 | D4 | D5 | idf | D1 | D2 | D3 | D4 | D5 |
| phishing | 0 | 2 | 0 | 5 | 0 | 0,40 | 0,00 | 0,80 | 0,00 | 1,99 | 0,00 |
| attack | 1 | 5 | 3 | 2 | 0 | 0,10 | 0,10 | 0,48 | 0,29 | 0,19 | 0,00 |
| defense | 2 | 0 | 0 | 0 | 2 | 0,40 | 0,80 | 0,00 | 0,00 | 0,00 | 0,80 |
| method | 0 | 0 | 1 | 1 | 1 | 0,22 | 0,00 | 0,00 | 0,22 | 0,22 | 0,22 |
| trick | 0 | 1 | 2 | 3 | 1 | 0,10 | 0,00 | 0,10 | 0,19 | 0,29 | 0,10 |
| user | 2 | 0 | 0 | 0 | 0 | 0,70 | 1,40 | 0,00 | 0,00 | 0,00 | 0,00 |
| unknown | 3 | 1 | 1 | 2 | 1 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| Class | | | | | | | probing | phishing | probing | phishing | probing |
| Jumlah | | | | | | | 2,29 | 1,38 | 0,71 | 2,70 | 1,11 |

Handling Text Data

DigitalSkola
Uncover The World of Digital Skills with Us

# TF-IDF Vectorization

Rumus TF-IDF

$$W_{t,d} = TF_{t,d} * IDF_t \quad (1)$$

Keterangan :

$W_{t,d}$ = bobot dari $t$ *(term)* dalam satu dokumen

$TF_{t,d}$ = frekuensi kemunculan $t$ *(term)* dalam dokumen $d$

$IDF_t$ = *Inverse document frequency*, dimana

$$IDF_t = \log\left(\frac{N}{n_t}\right) \quad (2)$$

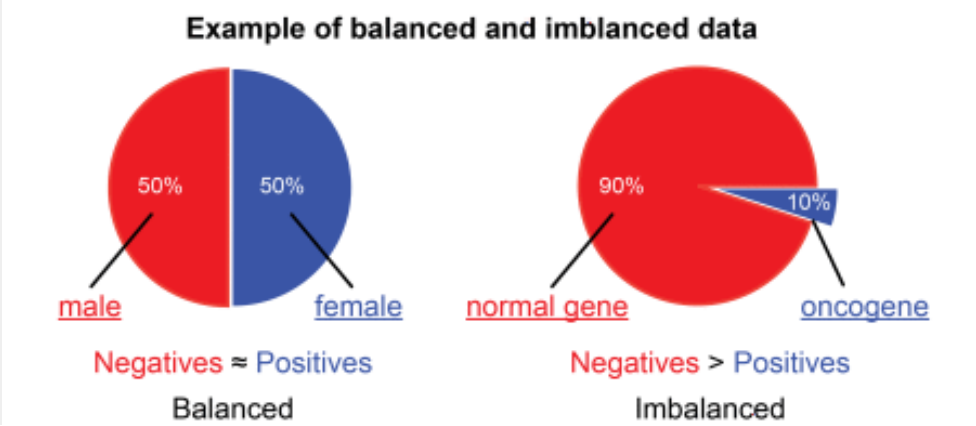
Keterangan :

$N$ = jumlah semua dokumen

$n_t$ = jumlah dokumen yang mengandung *term* $t$

# Imbalanced Dataset

Kelas yang tidak seimbang adalah masalah umum dalam klasifikasi pembelajaran mesin dimana terdapat rasio yang tidak proporsional di setiap kelas. ketidak seimbangan kelas dapat ditemukan di berbagai bidang termasuk diagnosa medis, penyaringan spam, dan deteksi penipuan.

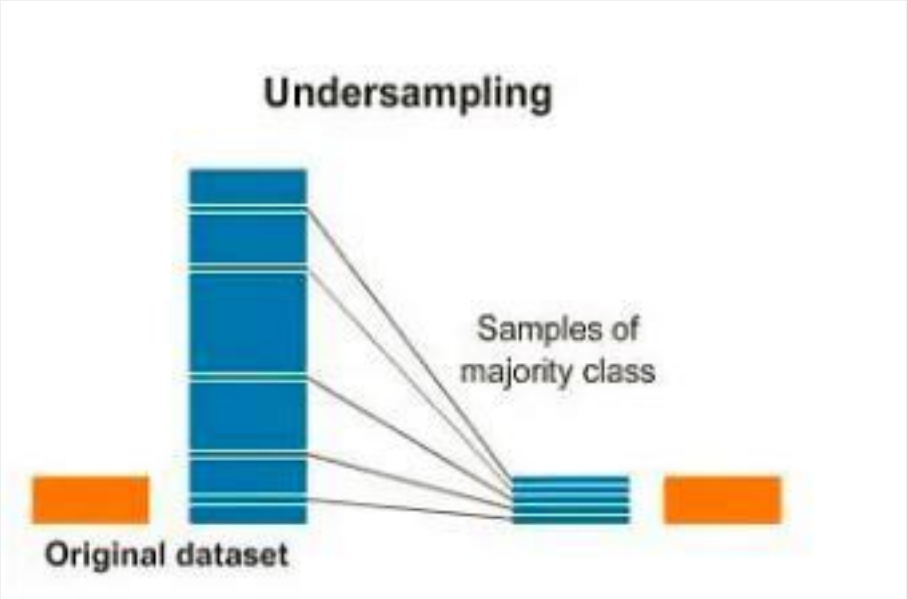


Imbalanced Dataset

DigitalSkola
Uncover The World of Digital Skills with Us

# Undersampling

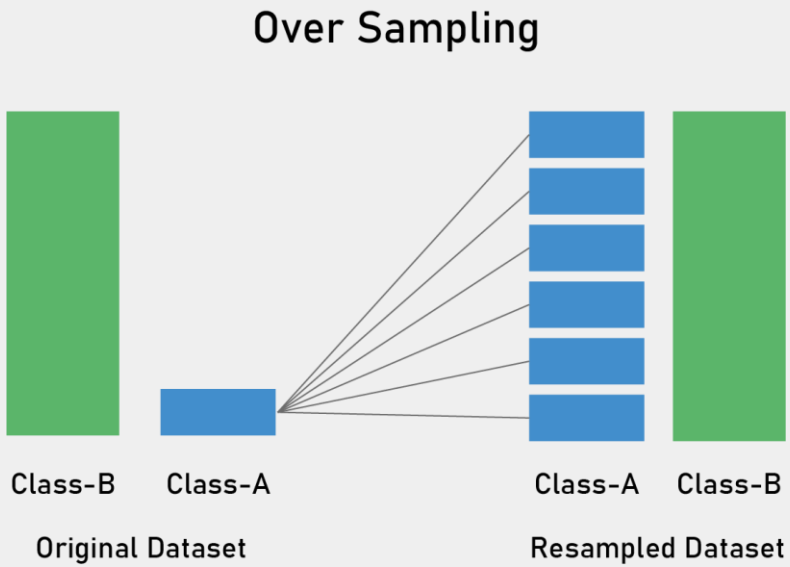
Menyeimbangkan dataset dengan mengurangi ukuran kelas yang berlimpah. Metode ini digunakan Ketika jumlah data mencukupi.



Imbalanced Dataset

# Oversampling

Digunakan Ketika jumlah data tidak mencukupi. Mencoba menyeimbangkan dataset dengan meningkatkan ukuran sample langka.



DigitalSkola
Uncover The World of Digital Skills with Us

# SMOTE

Metode **SMOTE** (*Synthetic Minority Over Sampling*) merupakan metode yang popular di terapkan dalam rangka menangani *ketidak seimbangan kelas*.

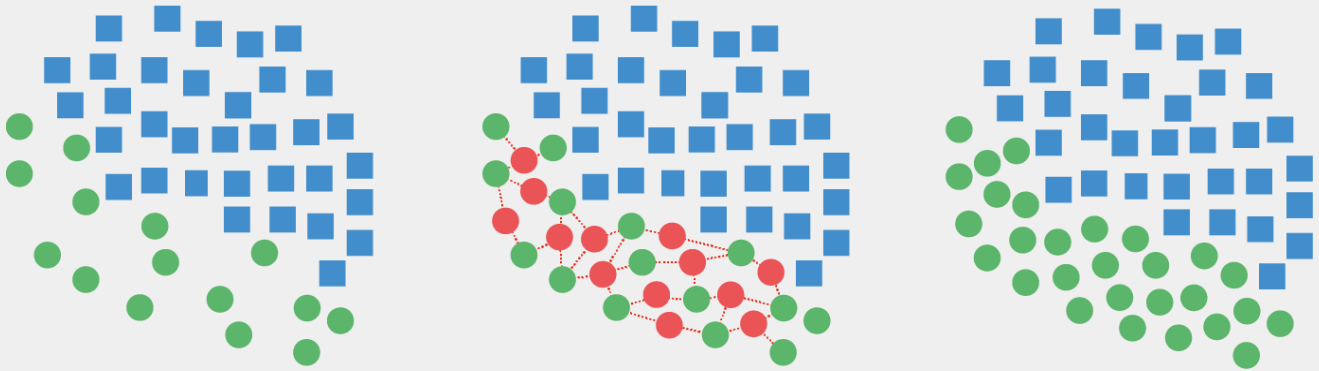
Teknik ini mensintesis sample baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara membuat instance baru dari minority class dengan pembentukan convex kombinasi dari instance yang saling berdekatan.

# Imbalanced Dataset

DigitalSkola
Uncover The World of Digital Skills with Us

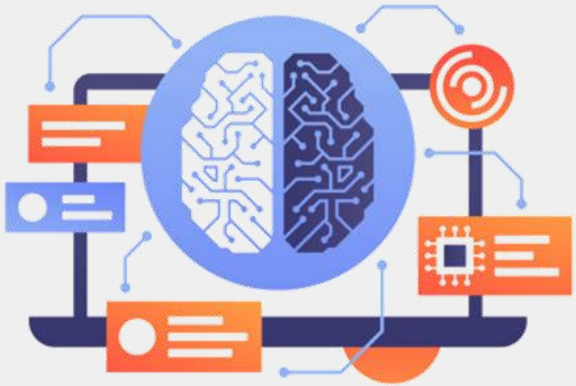
## SMOTE

### Synthetic Minority Oversampling Technique



# **2.** *Introduction to Machine Learning*

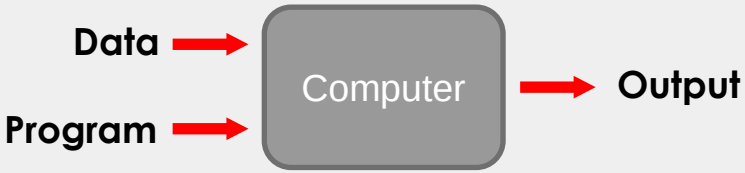
# Apa itu *Machine Learning*?



***Machine learning*** adalah proses yang memungkinkan sistem atau komputer untuk 'belajar' secara mandiri dan meningkatkan kemampuannya secara otomatis tanpa perlu instruksi pemrograman yang dituliskan secara eksplisit.

# Traditional Programming vs Machine Learning

## *Traditional Programming*



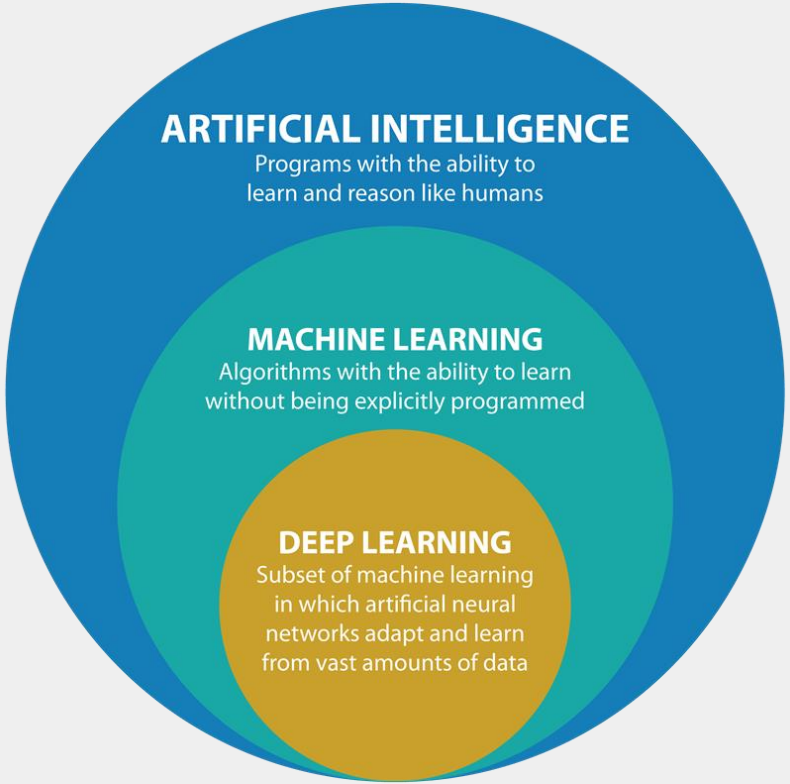

Pada pemrograman tradisional, komputer memerlukan data dan program untuk menghasilkan *output*.

## *Machine Learning*


Pada *machine learning*, data dan *output* dibutuhkan komputer untuk menghasilkan program.

# *Artificial Intelligence vs Machine Learning vs Deep Learning*

*Artificial Intelligence (AI)* diibaratkan sebagai payung yang lebih luas dimana *Machine Learning (ML)* dan *Deep Learning (DL)* berada dalam lingkupnya. Jadi, ML adalah bagian dari AI, dan DL adalah bagian dari ML.



# Contoh aplikasi *Machine Learning*

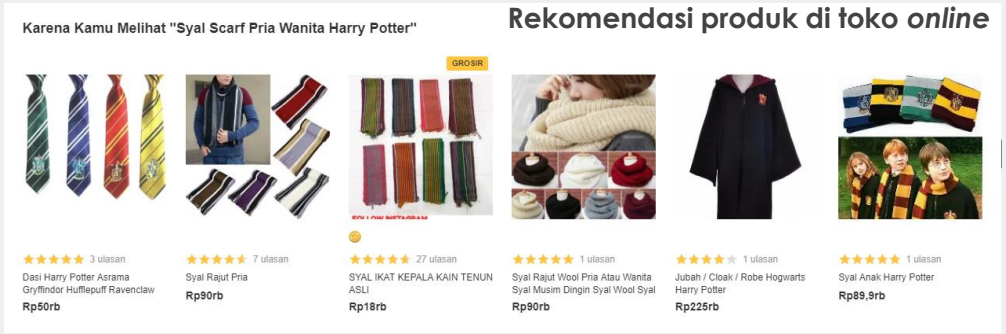
DigitalSkola
Uncover The World of Digital Skills with Us

**Virtual personal assistant**


**Rekomendasi *search***


**Rekomendasi titik penjemputan pada aplikasi Gojek**

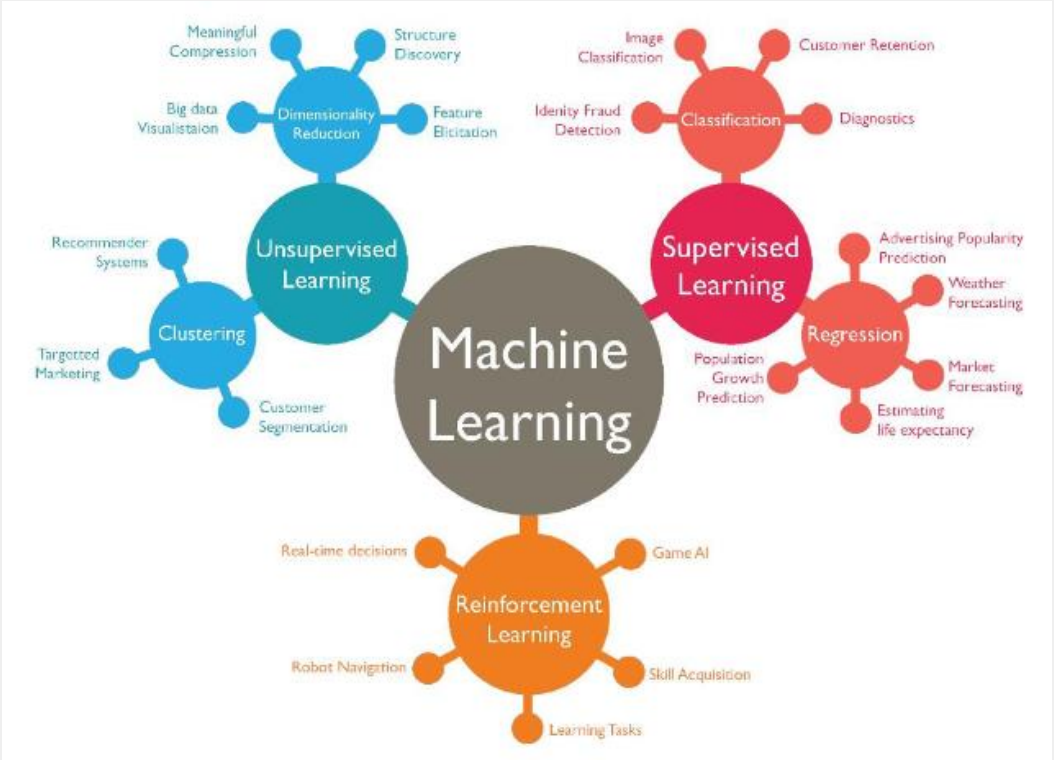

**Rekomendasi produk di toko *online***



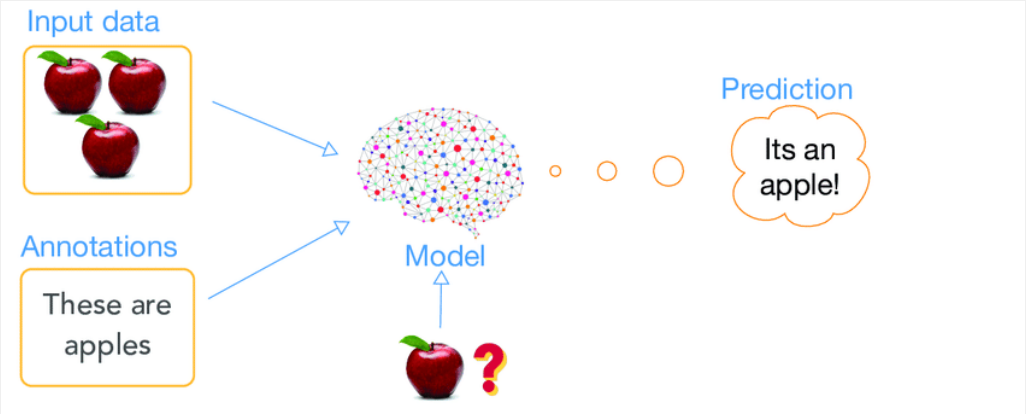
# Tipe-tipe *Machine Learning*

Secara garis besar, *machine learning* dikategorikan ke dalam 3 tipe utama, yaitu:

1. *Supervised learning*
2. *Unsupervised learning*
3. *Reinforcement learning*

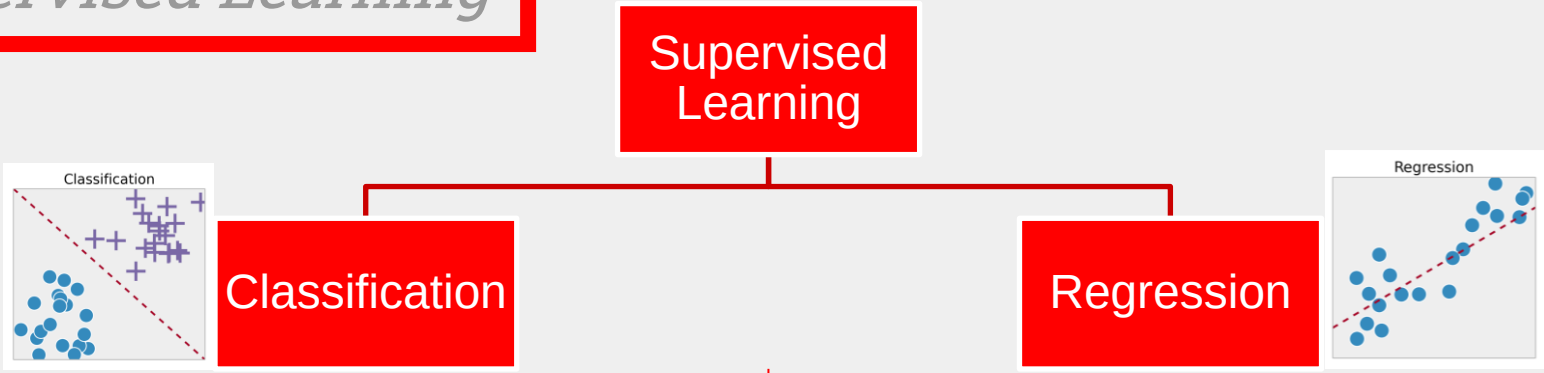


# Supervised Learning




***Supervised learning*** adalah suatu metode pembelajaran dimana sudah tersedianya data latih dan terdapat variable yang **memiliki label** sehingga tujuan akhirnya adalah mengelompokan suatu data ke data yang sudah ada.

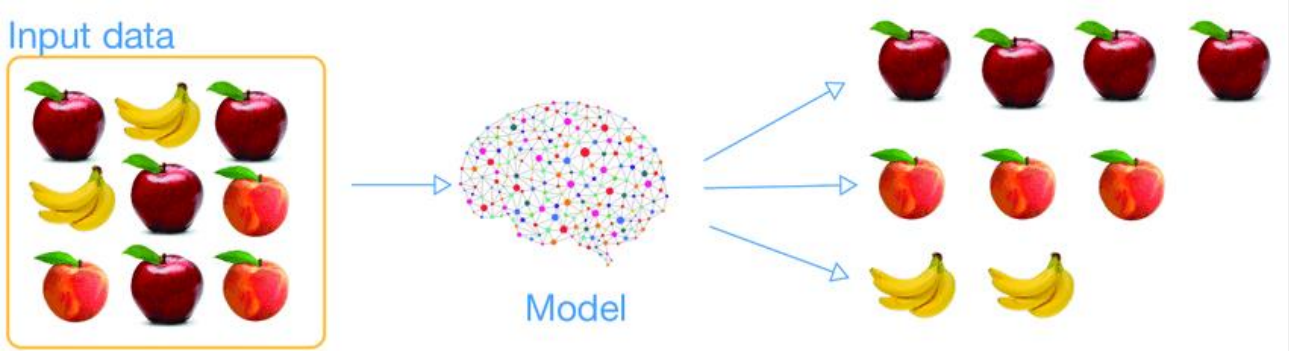
# Supervised Learning




***Classification*** digunakan ketika *output* yang ingin dihasilkan adalah kategorial atau berupa pengelompokan data berdasarkan label yang telah kita tentukan.

***Regression*** digunakan jika output yang diinginkan berupa nilai, misalnya untuk memprediksi harga rumah berdasarkan ukuran tanah dan bangunannya.

# *Unsupervised Learning*




***Unsupervised learning*** adalah suatu metode pembelajaran **tanpa label** data input. Pada algoritma unsupervised learning, data tidak secara eksplisit diberi label ke dalam kelas yang berbeda. Model mampu belajar dari data dengan menemukan pola implisit.

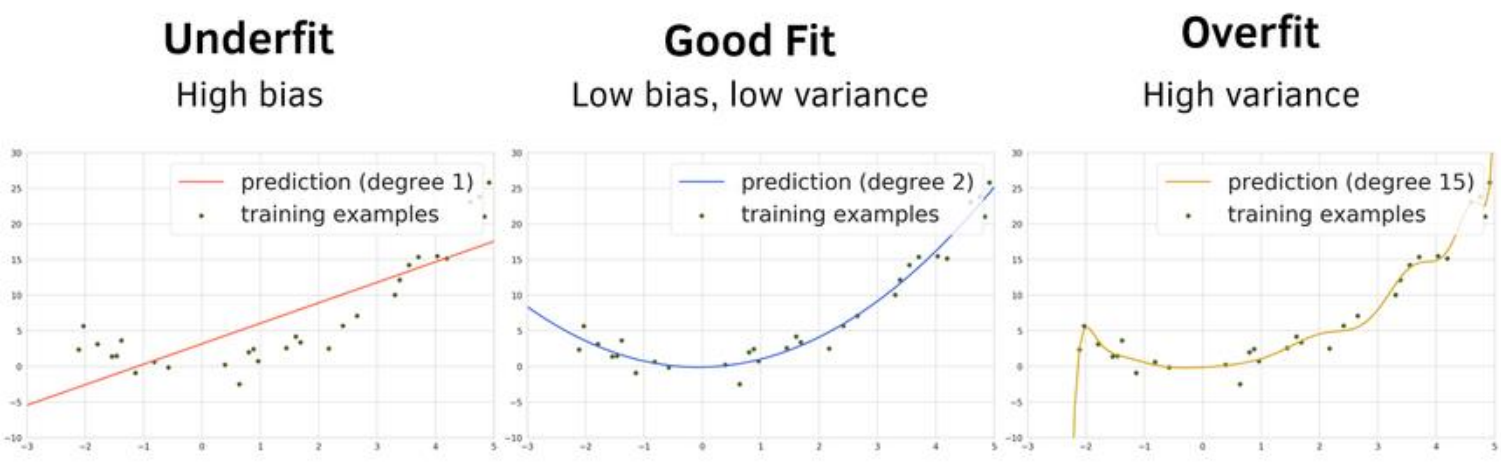
# *Unsupervised Learning : Clustering*

Metode paling umum dari unsupervised learning adalah *clustering*. ***Clustering*** merupakan metode pengelompokan yang secara otomatis membagi kumpulan data menjadi beberapa kelompok sesuai kesamaan.

Beberapa teknik clustering adalah:

1. K-means
2. DBSCAN
3. Hierarchical Clustering

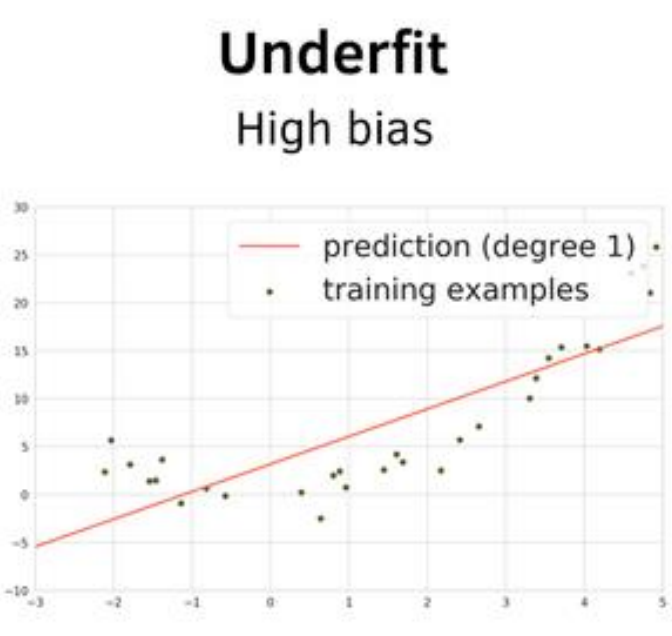
# Model Validation




Term: - Bias : tingkat kesalahan dari data pelatihan (data training)

- Variance : tingkat kesalahan dari data pengujian (data testing)

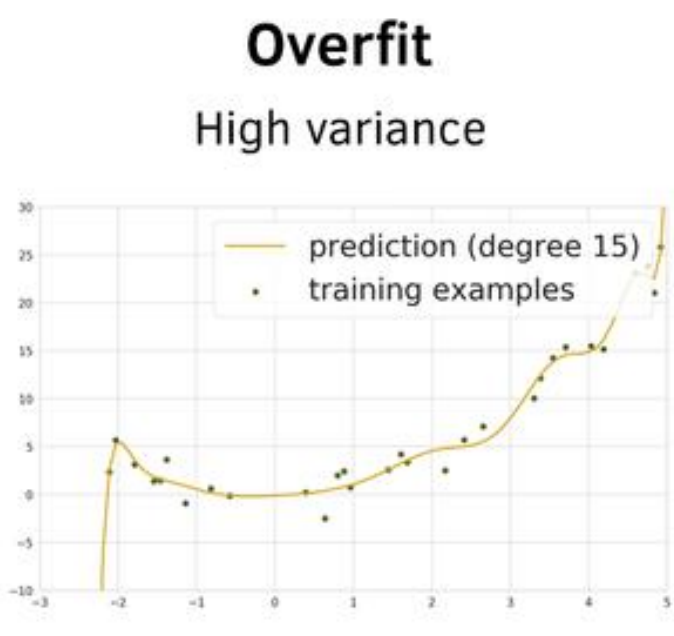
# Model Validation




***Underfitting* = bias tinggi dan varians rendah**

***Underfitting***, ketika model terlalu simpel dan tidak dapat menaganalisis data training secara menyeluruh sehingga hasil tidak optimal.

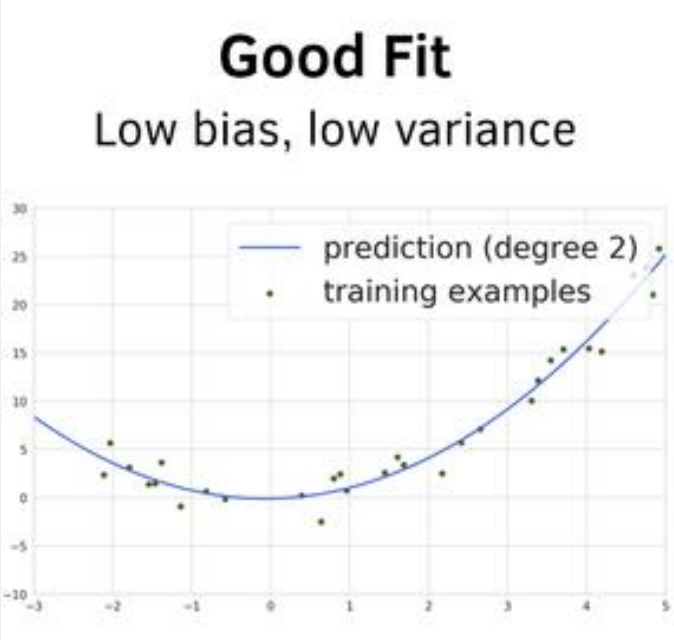
# Model Validation




***Overfitting* = varians tinggi dan bias rendah**

***Overfitting***, ketika model menghapal semua data training, tanpa menganalisis pola trend dari data sehingga tidak akan memberikan prediksi yang akurat.

# Model Validation




***Good fit* (model optimal) = bias rendah dan varians rendah**

***Good fit* (model optimal)**, ketika model teroptimasi dan juga tergeneralisasi dengan baik.

Special Thanks to :


Slide template by SlideCarnival