MACHINE LEARNING WITH PYTHON

GREGORINO AL JOSAN

OUTLINE

Introduction
- Background
- Contoh Aplikasi Machine Learning
- Jenis2 Problem Machine Learning
- Terminologi dalam Machine Learning
- Google Colab
- Python Packages for Machine Learning

Model-Model Machine Learning Berdasarkan Problem

Alur Kerja Machine Learning
- Pengumpulan Data
- Explorasi Data
- Pembersihan Data
- Pengolahan Data
- Seleksi Fitur
- Modeling

Deep Learning

Studi Kasus

Penutup

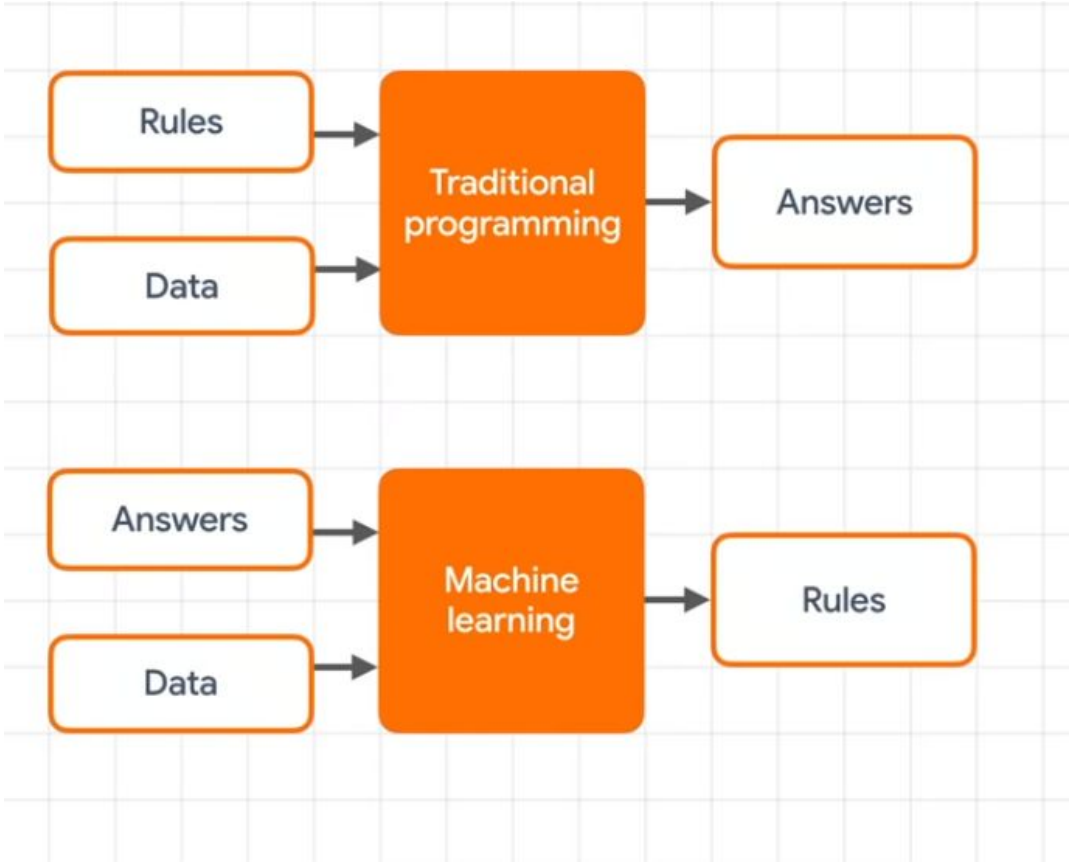
BACKGROUND

- Artificial Intelligence (AI) untuk membantu mempermudah pekerjaan
- Perkembangan Teknologi -> Data Banyak -> era Big Data
- Dibutuhkan teknik untuk mengolah data dan mengambil manfaat atau *insight* dari data
- *Machine Learning* adalah salah satu teknik yang digunakan untuk mengambil manfaat dari data, salah satunya untuk melakukan *predictive analysis*.

BACKGROUND

Machine learning is the study of algorithms that can learn from experience. As a machine learning algorithm accumulates more experience, typically in the form of observational data or interactions with an environment, its performance improves.

BACKGROUND



https://www.youtube.com/watch?v=_Z9TRANg4c0&list=PLOU2XLYxmsII9mzQ-Xxug4l2o04JBrkLV

CONTOH APLIKASI MACHINE LEARNING

- Sistem Rekomendasi
- Prediksi/Koreksi Kalimat
- Analisis sentimen
- Deteksi penyakit
- Forecasting

Hot Use Cases

- Image Generator (AI Art)
- ChatGPT

JENIS-JENIS *PROBLEM* MACHINE LEARNING

Supervised Learning -> output terdefinisi

- Regression -> output prediksi bernilai kontinu
- Classification -> output prediksi bernilai diskrit

Unsupervised Learning -> output tidak terdefinisi

- Clustering -> pengelompokkan data

TERMINOLOGI DALAM MACHINE LEARNING

TERMINOLOGI	PENGERTIAN
Data Point/Observasi/Datum	Satu baris informasi dalam data
Fitur	Variabel input yang akan dimasukkan ke dalam model
Target	Variabel yang ingin diprediksi
Model	Algoritma yang dipakai untuk dilatih agar dapat memberikan prediksi/*insight* dari data yang diberikan
Hyperparameter	Parameter yang mengontrol proses pembelajaran model
Inference	Melakukan prediksi dengan model yang dimiliki
Evaluation Metric	Rumus pengukur performa model

https://developers.google.com/machine-learning/crash-course/framing/ml-terminology

GOOGLE COLAB

Fitur Google Colab

- Storage 80 GB
- RAM 12 GB
- Free GPU
- Connection to Google Drive

Link: https://colab.research.google.com

GOOGLE COLAB

Advantage Google Colab

- Gak perlu instalasi kayak Jupyter Notebook Anaconda
- Kolaborasi

Disadvantage Google Colab

- Limited time GPU
- Gak bisa deployment

PYTHON PACKAGES FOR MACHINE LEARNING

Data Stuffs

- *Numpy*
- *Pandas*
- Scipy
- *Scikit-Learn*
- Tensorflow
- Keras
- PyTorch

Visualization

- *Matplotlib*
- *Seaborn*
- Plotly
- *Pandas*

Modeling

- *Sci-kit Learn*
- Tensorflow
- Keras
- PyTorch

MODEL-MODEL MACHINE LEARNING

MODEL-MODEL MACHINE LEARNING

Terdapat banyak model *machine learning* yang telah dikembangkan. Model-model tersebut dapat terbagi untuk beberapa permasalahan dalam *machine learning*, yaitu:

- Regresi
- Klasifikasi
- *Clustering*

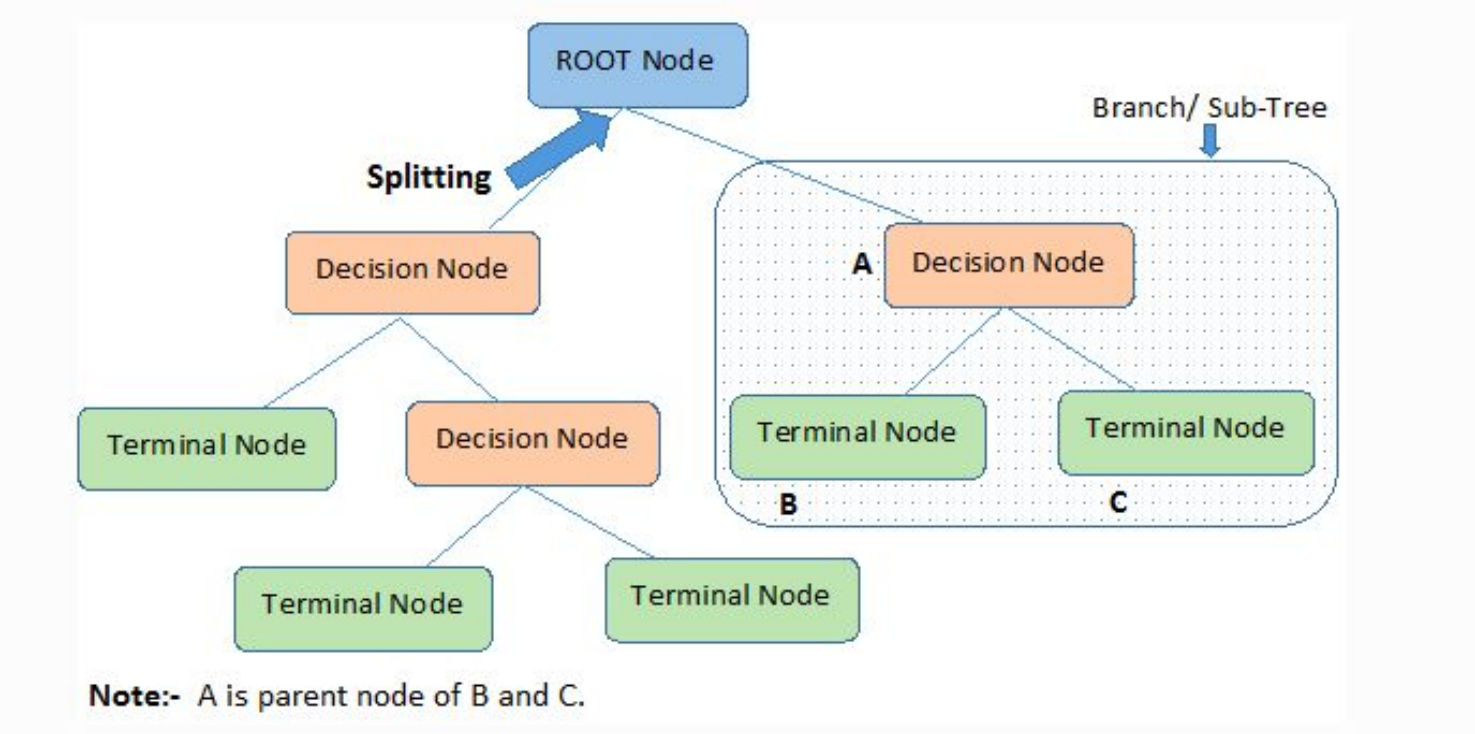
MODEL-MODEL MACHINE LEARNING BERDASARKAN *PROBLEM* - REGRESI

- Linear Regression
- Ridge Regression
- LASSO Regression
- Naïve Bayes
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- XGBoost

DECISION TREE

Decision Tree adalah model yang berbentuk pohon percabangan

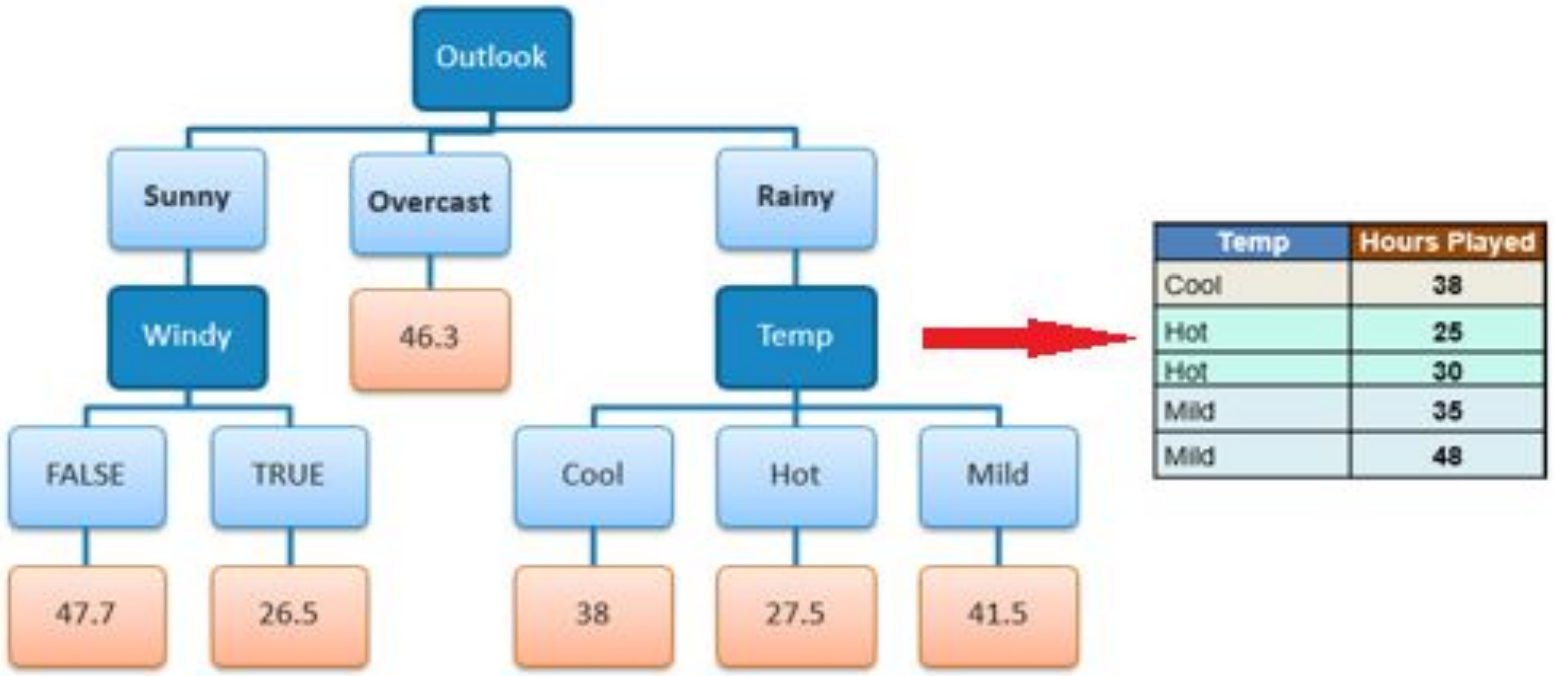
Decision Tree adalah model yang dapat memprediksi kelas atau nilai dari variable target dengan mempelajari aturan keputusan yang simpel yang dipelajari dari (training) data yang diberikan



https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

DECISION TREE REGRESSOR

Decision Tree Regressor adalah Decision Tree yang dimana output-nya berupa nilai kontinu



Temp	Hours Played
Cool	38
Hot	25
Hot	30
Mild	35
Mild	48

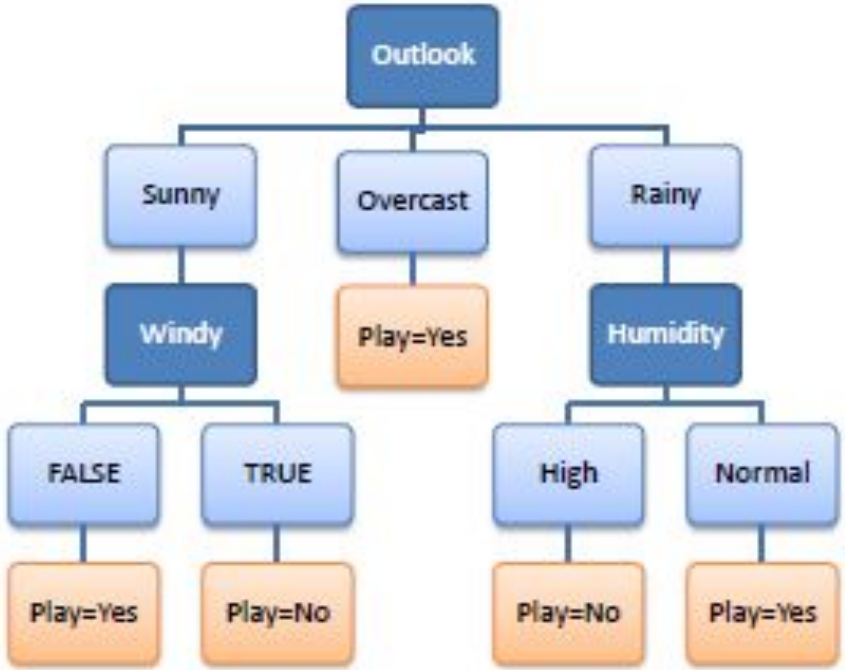
https://www.saedsayad.com/decision_tree_reg.htm

MODEL-MODEL MACHINE LEARNING BERDASARKAN *PROBLEM* - KLASIFIKASI

- Decision Tree
- Random Forest
- Logistic Regression
- Naïve Bayes
- K-Nearest Neighbour (K-NN)
- Support Vector Machine (SVM)
- XGBoost

DECISION TREE CLASSIFIER

Decision Tree Regressor adalah Decision Tree yang dimana output-nya berupa nilai diskrit



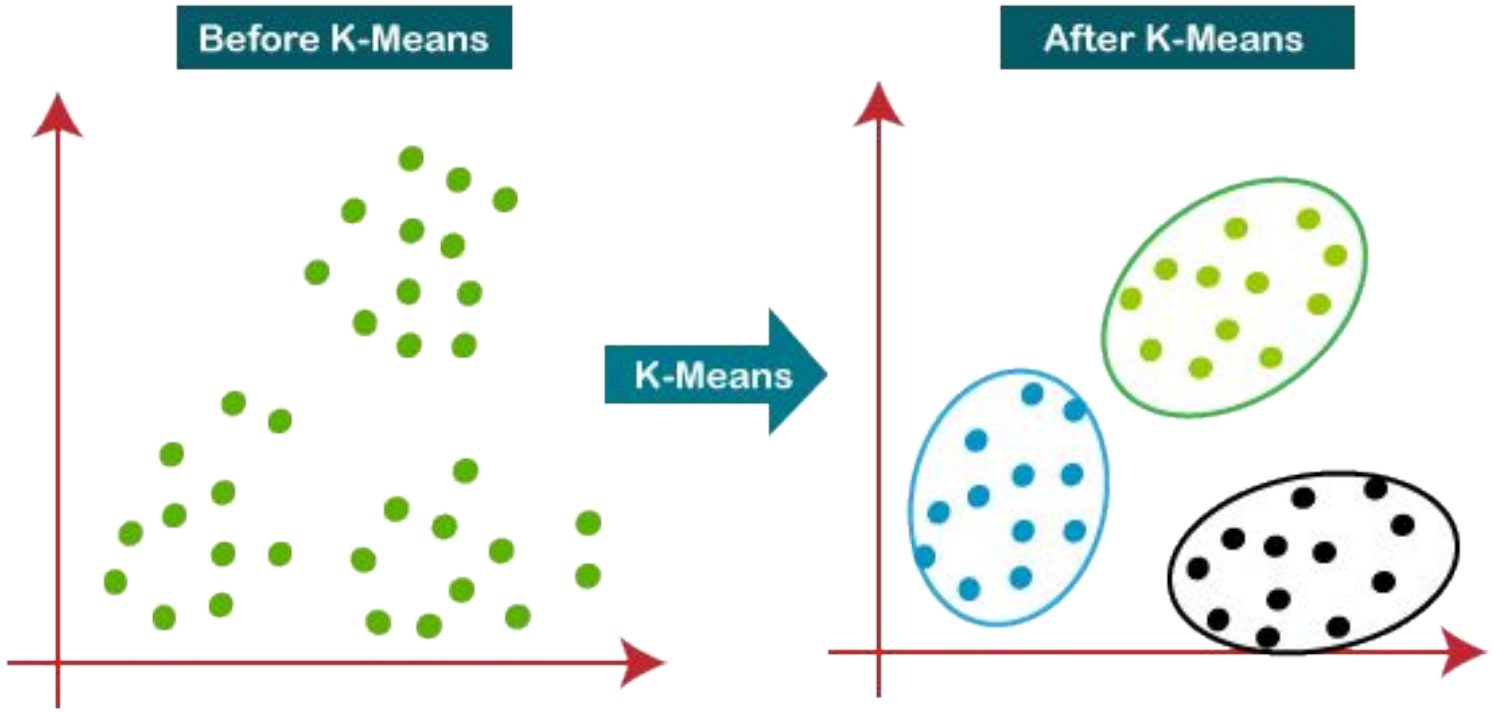
https://www.saedsayad.com/decision_tree.htm

MODEL-MODEL MACHINE LEARNING BERDASARKAN *PROBLEM - CLUSTERING*

- K-Means Clustering
- Partition Around Medoid (PAM)
- DBSCAN

K-MEANS CLUSTERING

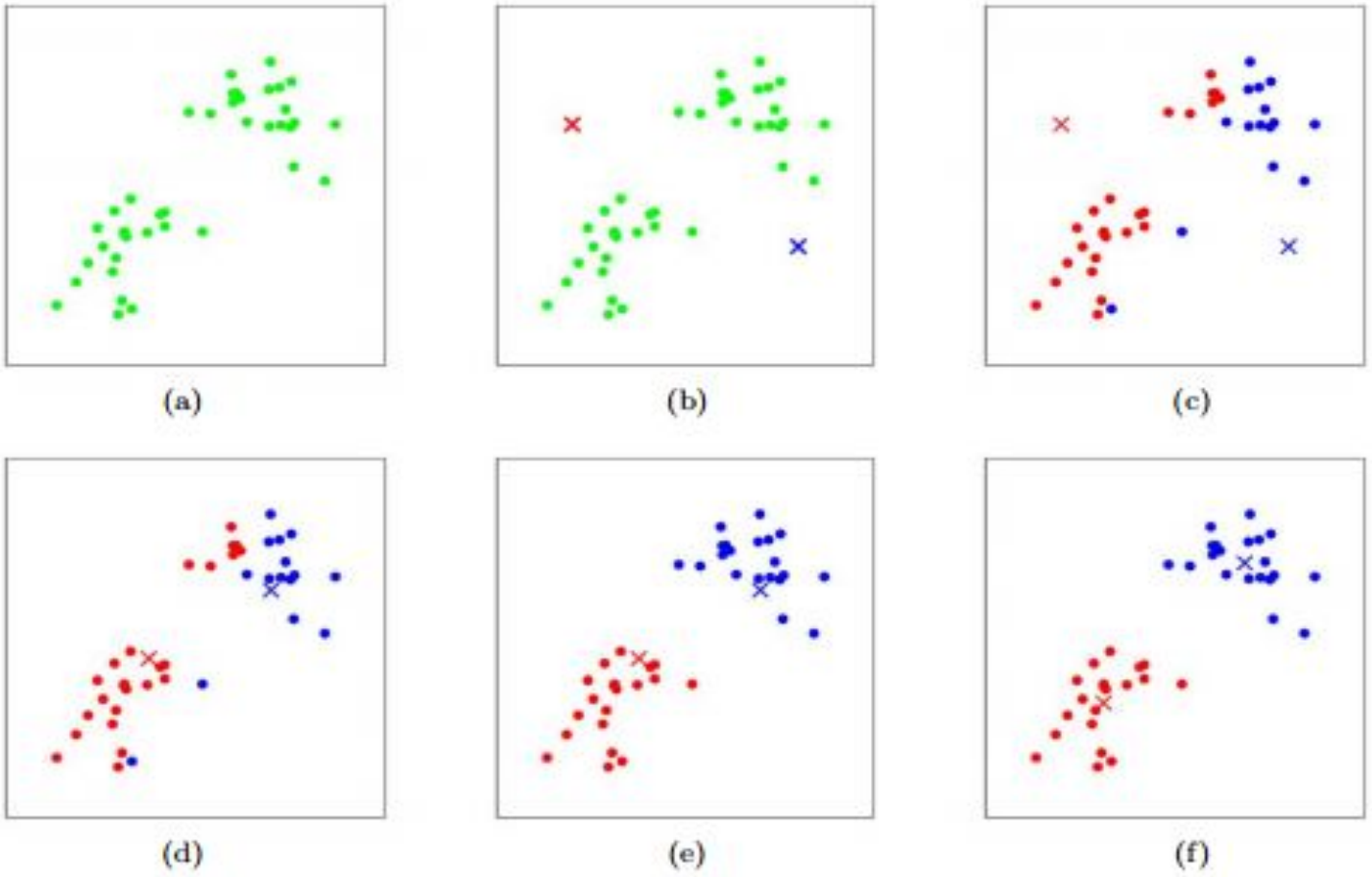
Pengelompokan data berdasarkan rata-rata dari masing-masing K *centroid*



https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

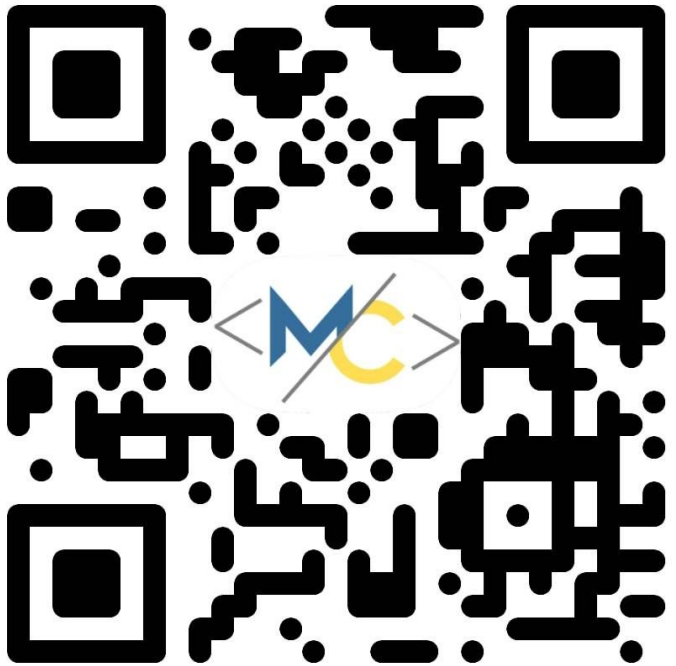
K-MEANS CLUSTERING

1. *Centroid* diinisialisasi secara random.
2. Masukkan data terdekat dari *centroid* ke dalam *centroid* tersebut.
3. Data-data dalam satu *centroid*/*cluster* dihitung nilai rata-ratanya.
4. *Centroid* bergerak ke titik rata-rata dari data dalam *centroid* tersebut.
5. Ulangi langkah 2-4 sampai konvergensi tercapai



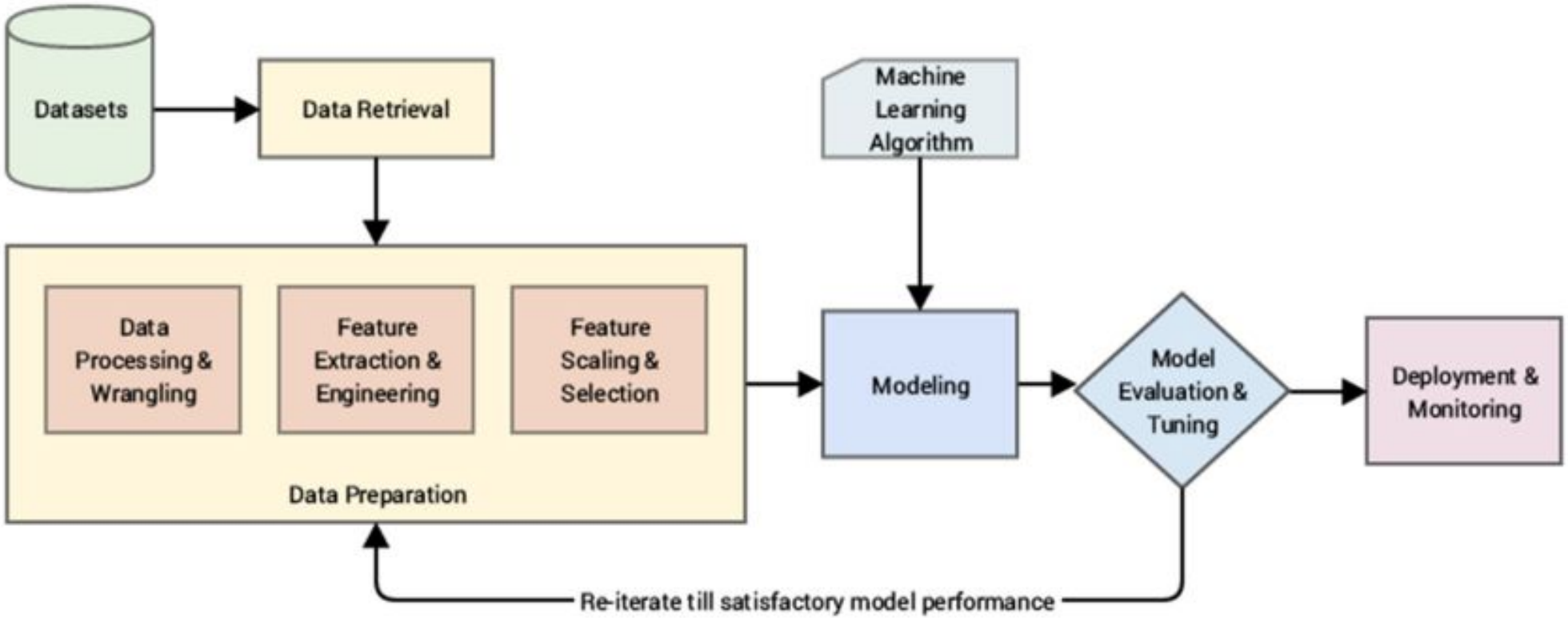
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

ALUR KERJA PROJECT MACHINE LEARNING



https://s.id/PembahasanMateriMC2_2022

WORKFLOW OF MACHINE LEARNING PROJECT



https://subscription.packtpub.com/book/big-data-and-business-intelligence/9781788831307/1/ch01lvl1sec05/standard-ml-workflow

ALUR KERJA MACHINE LEARNING PENGUMPULAN DATA

PENGUMPULAN DATA

Online Databases

- Kaggle
- Website Scraping
- Sports Reference (Olahraga)
- NCBI (Bioinformatika)

Private Sources

PENGUMPULAN DATA – KAGGLE

Kaggle is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges

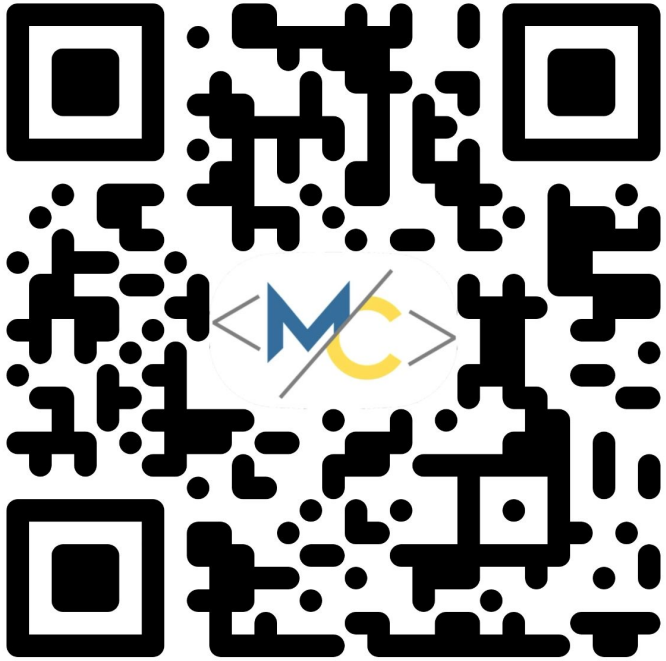
Dataset dari Kaggle bisa didownload lewat Google Colab

https://www.kaggle.com
https://en.wikipedia.org/wiki/Kaggle

DATASET MATERI

https://s.id/DatasetMateriMC2_2022

DEMONSTRASI DOWNLOAD DATASET KAGGLE KE GOOGLE COLAB/DRIVE



https://s.id/TutorialDownloadDataMC2_2022

ALUR KERJA MACHINE LEARNING EXPLORASI DATA

EXPLORASI DATA

Dilakukan untuk mengenali data yang dimiliki demi menemukan pola yang dimiliki data.

Beberapa hal yang dapat dicek:

- Tipe data
- Statistik dari data
- Distribusi data
- Anomali pada data (typo, nilai outlier yang ekstrim)
- Korelasi antara masing-masing kolom pada data
- Nilai-nilai yang dimiliki data (untuk kategorik)

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

EXPLORASI DATA

Explorasi data yang dilakukan di awal dapat menentukan hal yang dapat dilakukan berikutnya

Tahapan explorasi data dapat dilakukan berulang-ulang setelah berbagai tahapan pengolahan data dilakukan

ALUR KERJA MACHINE LEARNING PEMBERSIHAN DATA

PEMBERSIHAN DATA

Data Cleaning means the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is considered a foundational element of the basic data science.

Yang perlu diperhatikan…

- Informasi penting
- Typo
- Data Type
- Missing value
- Duplicate rows
- Outlier

https://monkeylearn.com/blog/data-cleaning-steps/
https://towardsdatascience.com/what-is-data-cleaning-how-to-process-data-for-analytics-and-machine-learning-modeling-c2afcf4fbf45

MISSING VALUE

Missing value adalah dimana suatu observasi memiliki kolom dengan ketidakberadaan nilai.

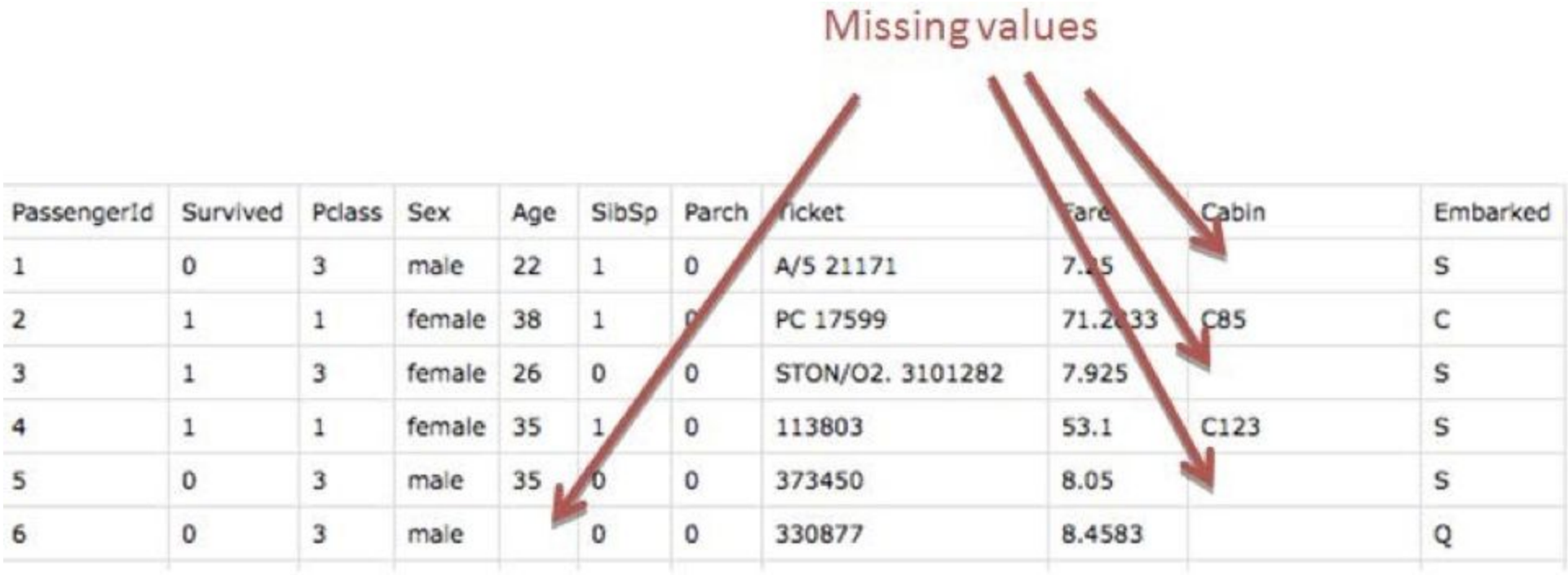
Missing value bisa terjadi karena kesalahan teknis, *human-error*, atau memang tidak ada info yang diberikan

Keberadaan *missing value* dapat merusak analisis dan proses modeling

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

MISSING VALUE

Contoh data dengan *missing value*



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

MISSING VALUE

Dalam library pandas, *missing value* direpresentasikan dengan NaN.

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

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

MISSING VALUE

Checking missing values

Handling missing values

- Menghapus baris yang memiliki *missing value*
- Imputasi. Mengisi *missing value* dengan beberapa opsi:
  - Untuk data numerik, menggunakan rata-rata atau median kolom atau nilai 0
  - Untuk data kategorik, bikin kategori baru yaitu ‘missing’ atau menggunakan nilai modus kolom

DUPLICATE ROWS

Duplicate rows adalah kondisi dimana ada 2 atau lebih baris/observasi yang memiliki kesamaan nilai pada seluruh kolom yang ada

Duplicate rows dapat memberikan hasil yang *bias* terhadap nilai yang direpresentasikan baris yang terduplikasi

DUPLICATE ROWS

Checking duplicate rows

How to handle duplicate rows

- Hapus barisan yang merupakan duplikasi

OUTLIERS

Outliers adalah kondisi dimana nilai dari suatu kolom dalam suatu observasi berada jauh dari distribusi nilai-nilai dalam kolom tersebut

Outliers dapat mengganggu analisis dan model

https://www.apexon.com/blog/why-being-aware-of-outliers-is-vital-for-machine-learning-datasets/

OUTLIERS

Checking outliers

How to handle outliers

- Hapus baris
- Bikin batasan
- Didiamkan

https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/

DEMONSTRASI PEMBERSIHAN DATA....

ALUR KERJA MACHINE LEARNING PENGOLAHAN FITUR

PENGOLAHAN FITUR

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

PENGOLAHAN FITUR – DATA NUMERIK

Mempengaruhi konvergensi beberapa model ML

Mempengaruhi model yang menggunakan jarak dalam algoritmanya

Normalisasi

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardisasi

$$X_{new} = \frac{X - \mu}{\sigma}$$

https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

PENGOLAHAN FITUR – DATA NUMERIK

Mengubah distribusi data yang berantakan menjadi lebih baik

Log Transform

$$X_{new} = Log(X)$$

https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/

PENGOLAHAN FITUR – DATA KATEGORIK

Memberikan representasi numerik dari data kategorik kepada model

Kategorik – Ordinal

- Ordinal Encoding

Kategorik – Nominal

- One-Hot Encoding
- Dummy Variable Encoding

https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/

PENGOLAHAN FITUR – DATA KATEGORIK

Ordinal Encoding

Kalimat	Sentimen
Wah bagus sekali	Positif
Barangnya jelek	Negatif
Agak kurang	Semi-negatif

Kalimat	Sentimen
Wah bagus sekali	5
Barangnya jelek	1
Agak kurang	2

PENGOLAHAN FITUR – DATA KATEGORIK

One-Hot Encoding

Nama	Jurusan
Mahasiswa A	Matematika
Mahasiswa B	Statistika
Mahasiswa C	Aktuaria

Nama	Matematika	Statistika	Aktuaria
Mahasiswa A	1	0	0
Mahasiswa B	0	1	0
Mahasiswa C	0	0	1

Dummy Variable Encoding

Nama	Jurusan
Mahasiswa A	Matematika
Mahasiswa B	Statistika
Mahasiswa C	Aktuaria

Nama	Matematika	Statistika
Mahasiswa A	1	0
Mahasiswa B	0	1
Mahasiswa C	0	0

PENGOLAHAN FITUR

Pembuatan fitur baru secara manual

Menciptakan fitur yang dapat merepresentasikan beberapa fitur yang sudah ada

DEMONSTRASI PENGOLAHAN FITUR....

ALUR KERJA MACHINE LEARNING SELEKSI FITUR

SELEKSI FITUR

Seleksi atau pemilihan fitur dilakukan untuk mengurangi jumlah input variable/fitur menjadi beberapa fitur yang berguna saja terhadap model.

Seleksi fitur utamanya dilakukan untuk menghapus fitur yang tidak informatif atau *redundant* untuk model.

- Korelasi antara fitur dengan target
- Korelasi antar fitur
- Domain knowledge
- PCA

https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

SELEKSI FITUR

Mengetahui apakah ada korelasi antara fitur dengan target.

Jika tidak ada korelasi antara fitur dengan target, maka fitur tersebut bisa dihapus/tidak digunakan

- Korelasi antara fitur dengan target
- Korelasi antar fitur
- Domain knowledge
- PCA

SELEKSI FITUR

Mengetahui apakah ada korelasi antara masing-masing fitur.

Jika ada 2 fitur yang berkorelasi kuat maka salah satu fitur dapat dipertimbangkan untuk tidak digunakan.

- Korelasi antara fitur dengan target
- Korelasi antar fitur
- Domain knowledge
- PCA

SELEKSI FITUR

Mengetahui fitur-fitur yang berguna untuk model dengan ilmu (*subject-matter expert*)

- Korelasi antara fitur dengan target
- Korelasi antar fitur
- Domain knowledge
- PCA

SELEKSI FITUR

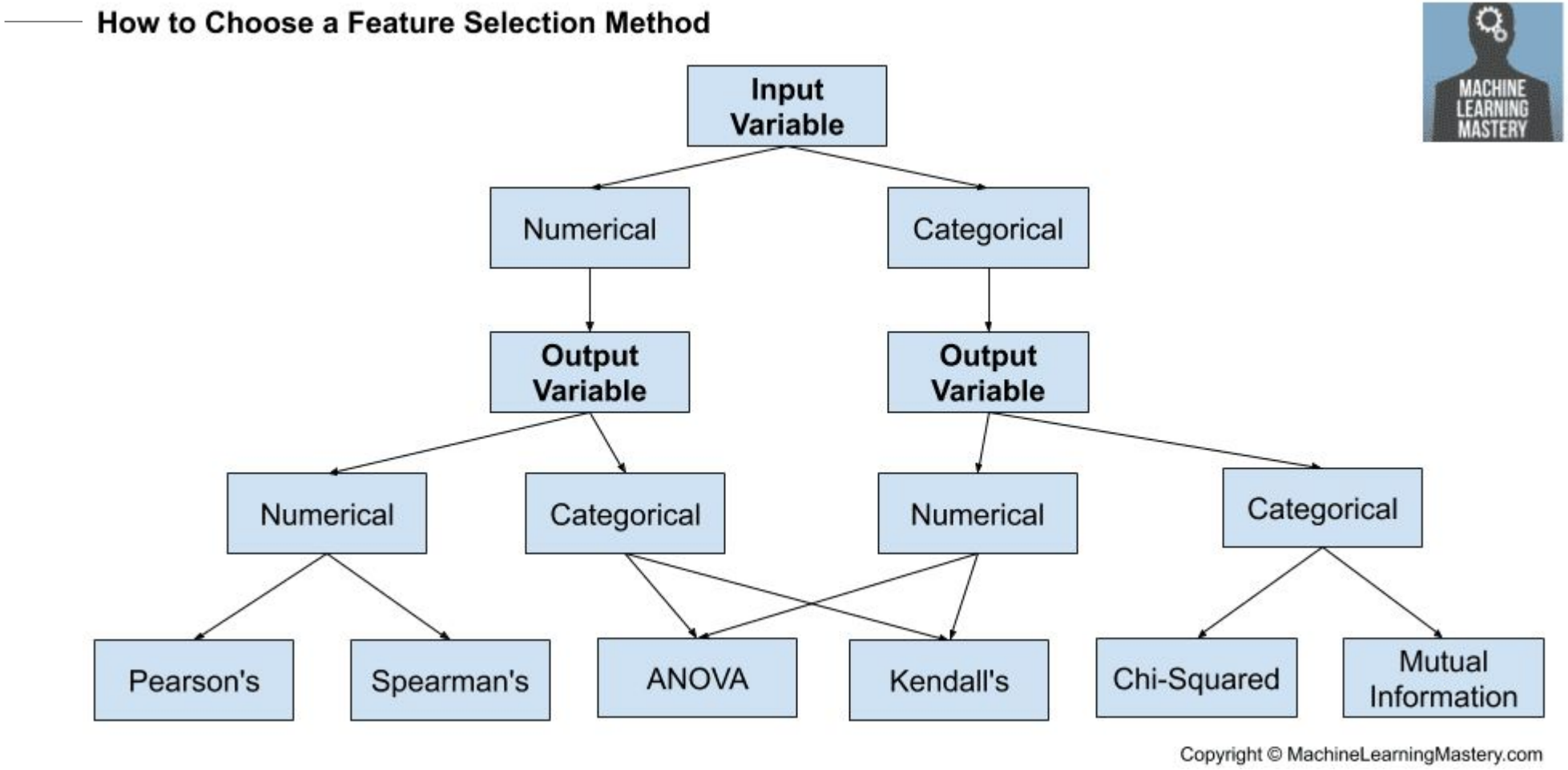
Mencari representasi data N-dimensi dengan dimensi yang lebih rendah

Reduksi fitur berdasarkan eigenvector

- Korelasi antara fitur dengan target
- Korelasi antar fitur
- Domain knowledge
- PCA

https://medium.com/geekculture/principal-component-analysis-3d2b3a0bb93e

PEMILIHAN FITUR



https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

DEMONSTRASI SELEKSI FITUR....

ALUR KERJA MACHINE LEARNING MODELING

FITTING MODEL

Membuat model mengenali *pattern* dari data yang diberikan

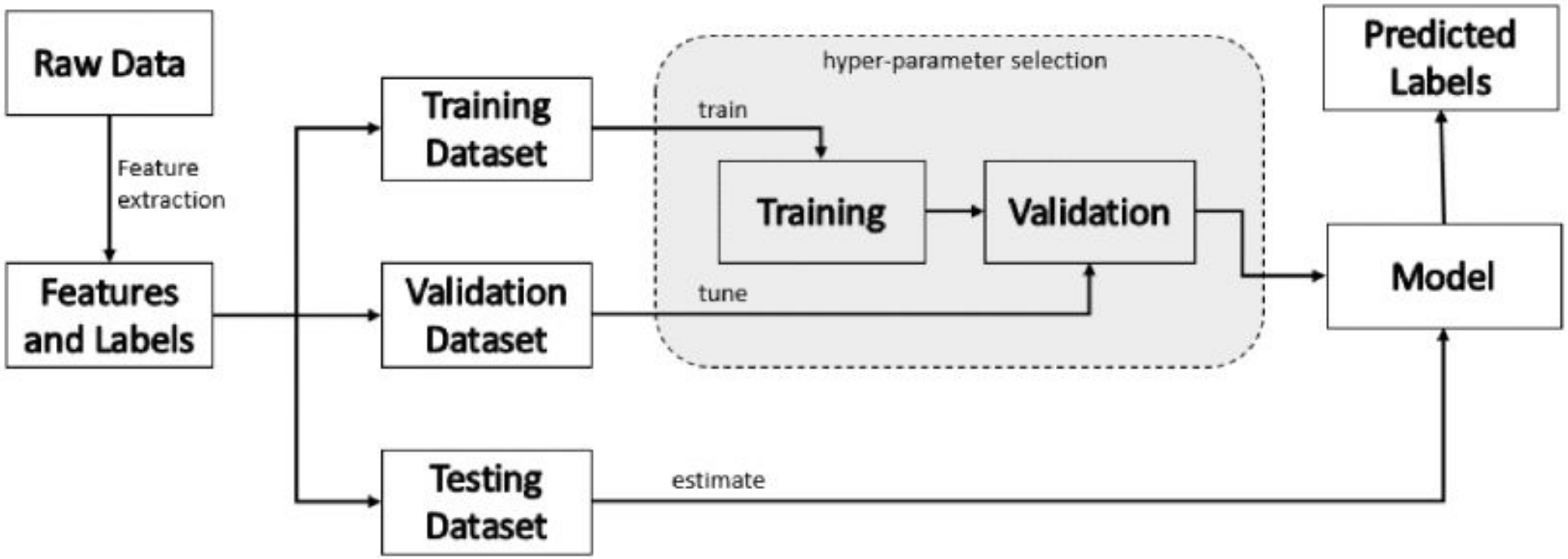
Dataset splitting strategy

- Training, Testing
- Training, Validation, Testing

Catatan mengenai splitting

- Distribusi data harus sama

SKENARIO FITTING MODEL JIKA MENGGUNAKAN 3 SPLIT



A. Basavaraju, J. Du, F. Zhou and J. Ji, "A Machine Learning Approach to Road Surface Anomaly Assessment Using Smartphone Sensors," in *IEEE Sensors Journal*, vol. 20, no. 5, pp. 2635-2647, 1 March1, 2020, doi: 10.1109/JSEN.2019.2952857

EVALUASI MODEL

Mengukur performa model yang telah dilakukan *fitting* terhadap data yang diberikan

Umum:

- Running Time

Berdasarkan Permasalahan:

- Regresi
- Klasifikasi

EVALUASI MODEL – REGRESI

Mean Squared Error (MSE)

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

Rooted Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

EVALUASI MODEL – REGRESI

R2-Score

Untuk mengukur kebergunaan model. Jika nilainya 1 berarti model sangat berguna.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

EVALUASI MODEL – KLASIFIKASI

Confusion Matrix

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

Akurasi

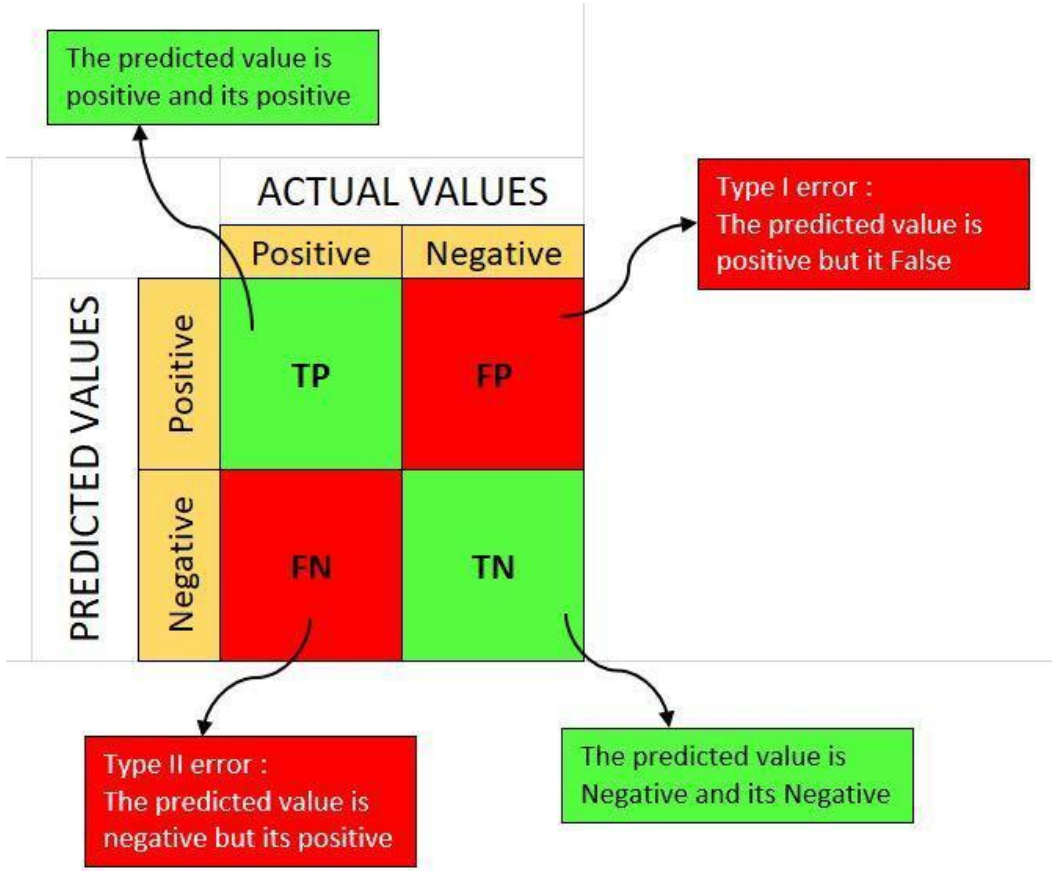
Presisi

Recall

https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

EVALUASI MODEL – KLASIFIKASI

Confusion Matrix adalah tabel representasi prediksi terhadap nilai asli

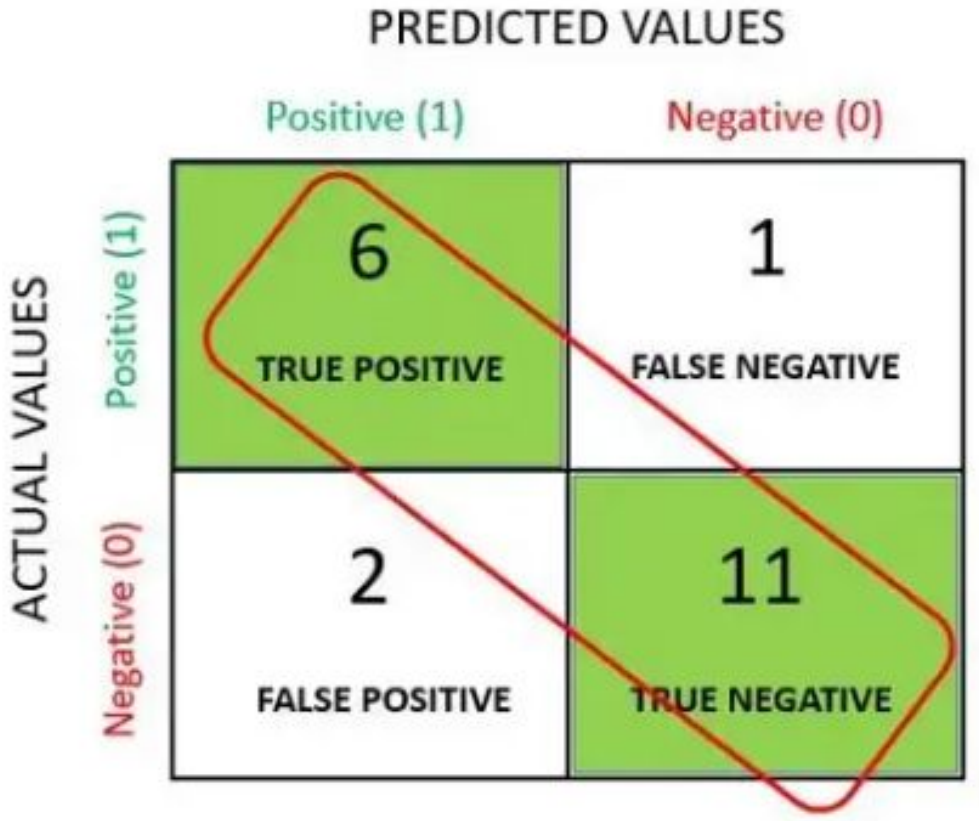


https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

EVALUASI MODEL – KLASIFIKASI

Akurasi

Mengukur seberapa akurat prediksi yang dibuat.

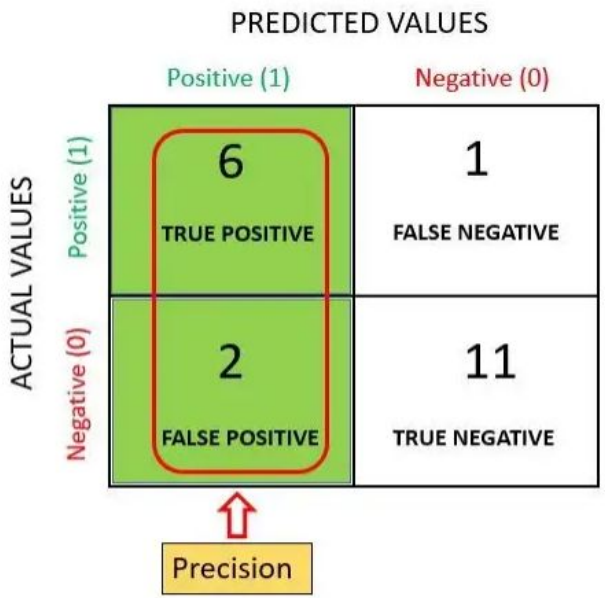


https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

EVALUASI MODEL – KLASIFIKASI

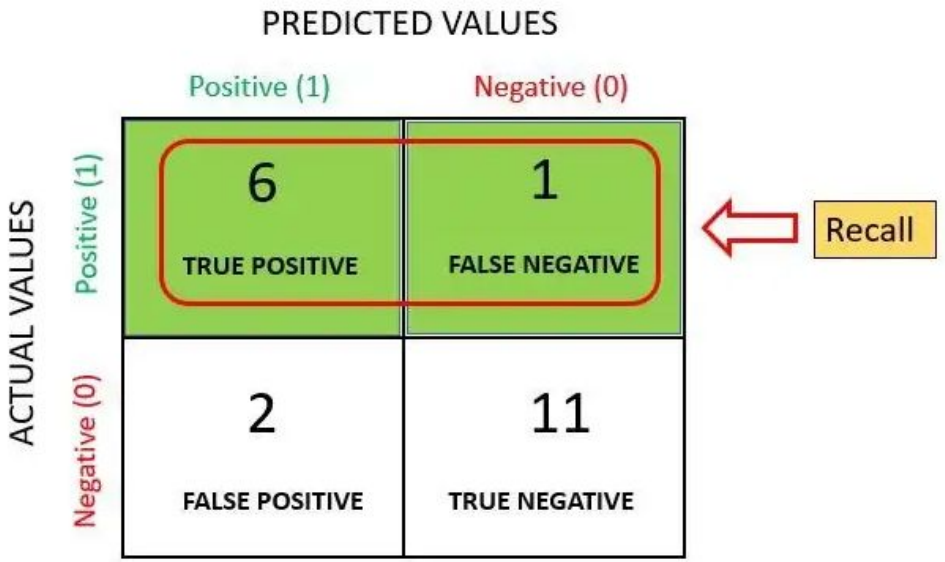
Presisi

Mengukur seberapa akurat prediksi positif yang didapatkan **dari seluruh prediksi yang positif**.



- Recall

Mengukur seberapa akurat prediksi positif yang didapatkan dari **seluruh nilai asli yang positif**.



https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

POSSIBLE STEPS TO IMPROVE MODEL

Cross-Validation

- Melakukan splitting datasets menjadi K bagian, kemudian dilakukan training selama K kali menggunakan K-1 bagian sebagai train dataset dan 1 bagian sisa sebagai test dataset. Setiap bagian harus mendapat giliran menjadi test dataset.

Hyperparameter Tuning

- Contoh: Mengatur jumlah kedalaman pohon dalam Decision Tree

Tambah data

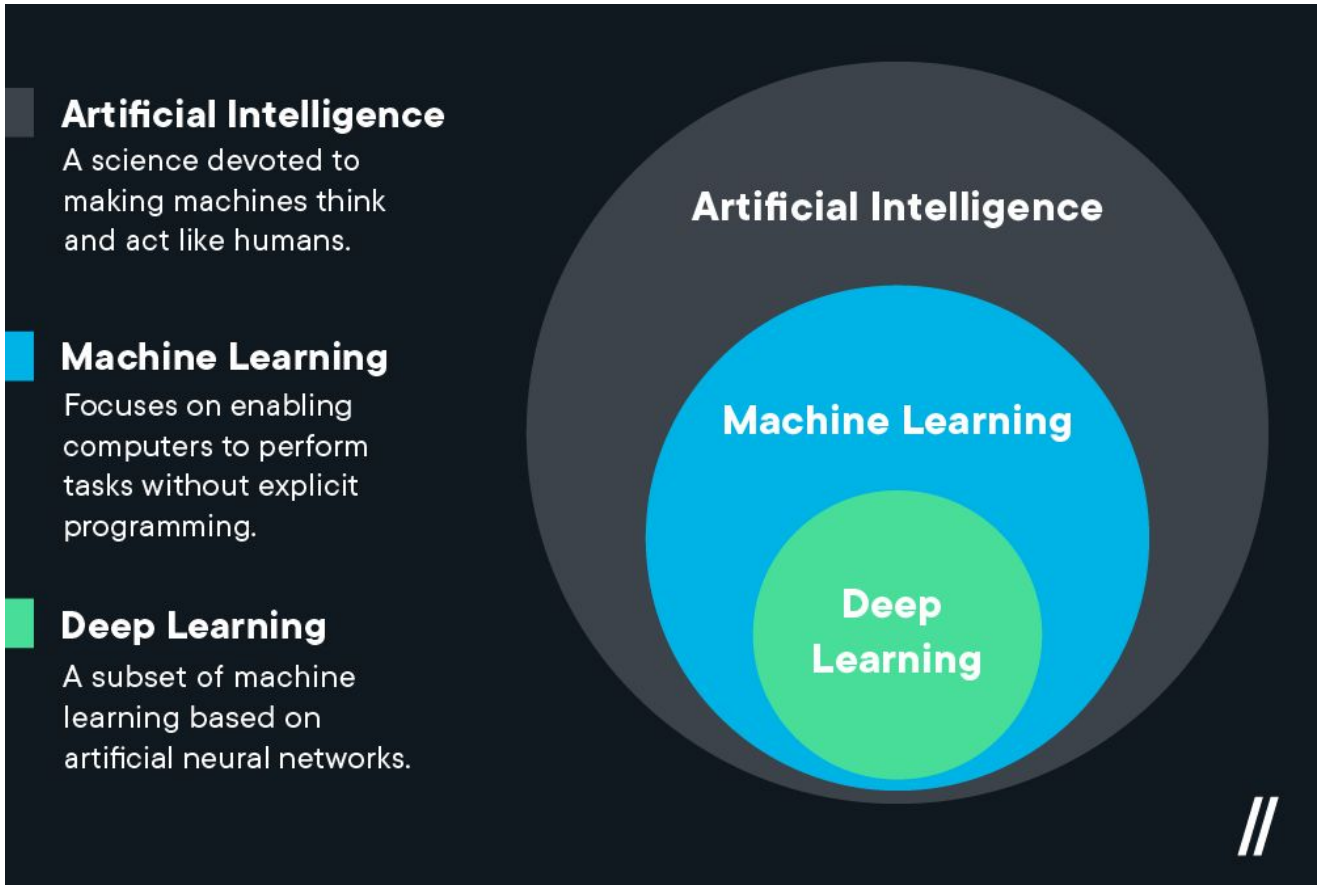
- Jumlah parameter > Jumlah data

Redo Pengolahan fitur/Seleksi fitur

DEMONSTRASI FITTING MODEL....

DEEP LEARNING

DEEP LEARNING



- Subset dari ML
- Implementasi ML menggunakan Neural Network
- Model jauh lebih kompleks
- Data yang dimiliki tidak terstruktur

https://flatironschool.com/blog/deep-learning-vs-machine-learning/

DEEP LEARNING

Machine learning	Deep learning
A subset of AI	A subset of machine learning
Can train on smaller data sets	Requires large amounts of data
Requires more human intervention to correct and learn	Learns on its own from environment and past mistakes
Shorter training and lower accuracy	Longer training and higher accuracy
Makes simple, linear correlations	Makes non-linear, complex correlations
Can train on a CPU (central processing unit)	Needs a specialized GPU (graphics processing unit) to train

https://www.coursera.org/articles/ai-vs-deep-learning-vs-machine-learning-beginners-guide

DEEP LEARNING

Artificial Neural Network (ANN)

Convolutional Neural Network (CNN)

- Buat data gambar
- Bisa buat data barisan/*sequence*

Recurrent Neural Network (RNN)

- Buat data barisan/*sequence*
- Contoh: bahasa, data time-series

STUDI KASUS

DATASET:

https://s.id/DatasetStudiKasusMC2_2022

STUDI KASUS – REGRESI

1. Prediksi harga rumah

Prediksikan harga suatu rumah berdasarkan data yang telah diberikan.

https://www.kaggle.com/competitions/neolen-house-price-prediction

STUDI KASUS – KLASIFIKASI

2. Prediksi klasifikasi harga rumah

Klasifikasikan harga suatu rumah berdasarkan data yang telah diberikan pada studi kasus regresi. Namun, data harga diubah dari yang tadinya numerik menjadi kategorik menjadi: Murah, Sedang, dan Mahal.

Syarat kategorisasi data numerik menjadi Murah, Sedang, atau Mahal tergantung peserta.

https://www.kaggle.com/competitions/neolen-house-price-prediction

PENGERJAAN STUDI KASUS....

PENUTUP

ABOUT MACHINE LEARNING

Machine Learning hanya masalah *practice*

Beberapa hal yang penting untuk diingat

- Pertimbangan cost training dan deploying model
- Interpretasi hasil model

Data yang jelek lebih buruk daripada model yang jelek

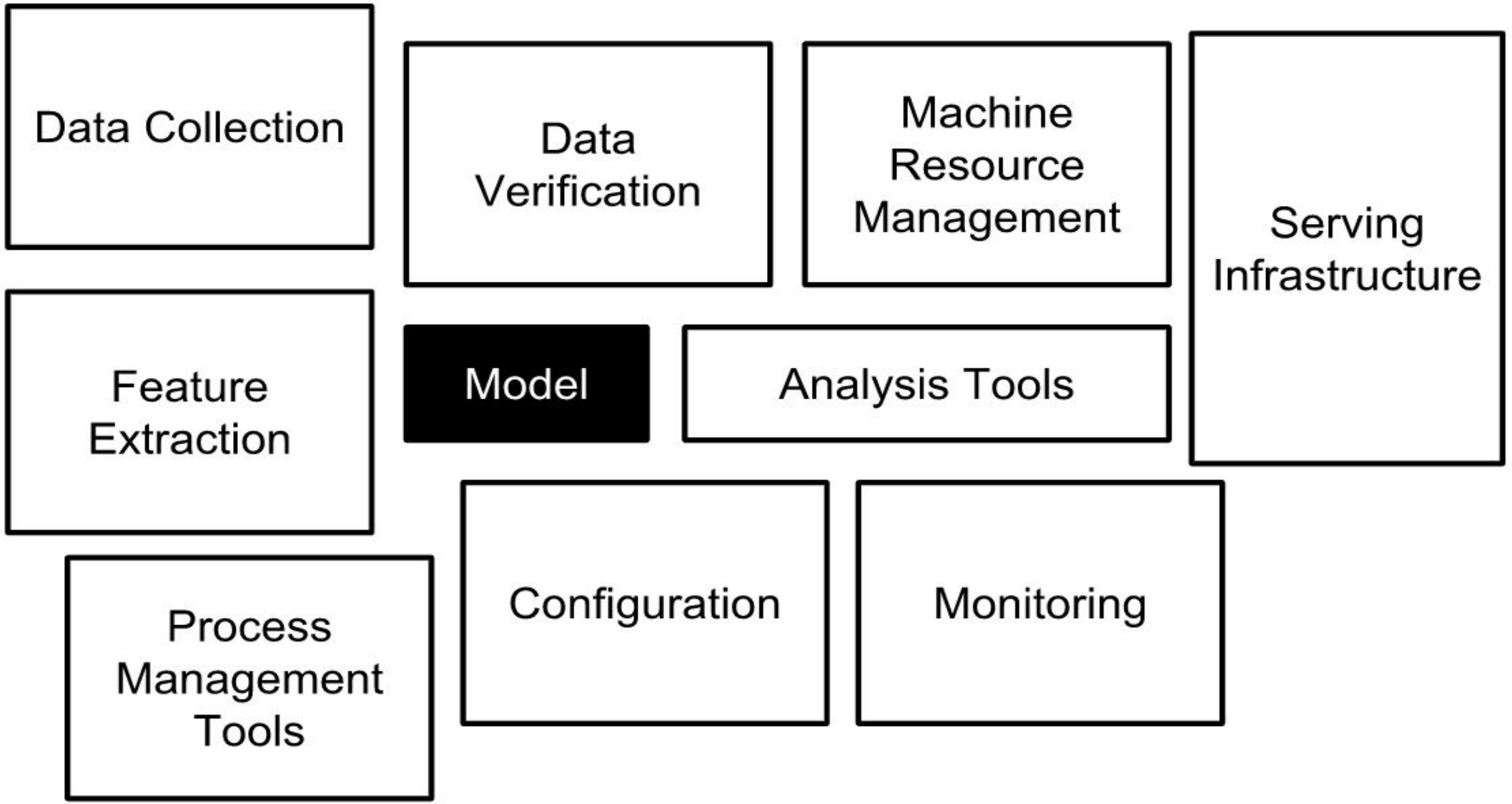
Explorasi Data (Exploratory Data Analysis) dilakukan di setiap fase

Dalam project ML, waktu paling banyak habis di bagian data

DATA SCIENCE PROJECT IS BIG

Role Data Scientist berhubungan dengan:

- Data Engineer
- Machine Learning Engineer
- DevOps/MLOps Engineer



https://developers.google.com/machine-learning/testing-debugging/pipeline/overview

RANGKUMAN

Machine Learning adalah cabang dari Artificial Intelligence yang mana berfokus untuk menciptakan sebuah model yang dapat memberikan prediksi/*insight* berdasarkan data yang telah ada

Model-model Machine Learning dapat dibagi berdasarkan permasalahan, yaitu regresi, klasifikasi, dan *clustering*

Project Machine Learning terdiri dari tahapan sebagai berikut:

- Data acquisition
- Data cleaning
- Feature engineering
- Feature selection
- Model fitting & evaluation

TEMPAT BELAJAR MACHINE LEARNING

Website

- www.medium.com
- www.towardsdatascience.com
- www.analyticsvidhya.com
- www.machinelearningmastery.com

E-book

- www.d2l.ai

Courses

- Courses from Deep Learning.ai
- Coursera
- Kaggle Courses
- EdX

YouTube Channel

- Krish Naik

WHAT'S NEXT

Data Synthetic

- SMOTE

Data Preprocessing

- Tergantung permasalahan (bahasa, gambar, signal, etc.)
- Tergantung domain knowledge

WHAT'S NEXT

Deep Learning

- Computer Vision
- Natural Language Processing

Advanced Model

- Hybrid Model
- Ensemble Learning

WHAT'S NEXT

Model Validation Techniques

- K-fold Cross Validation

Model Deployment

- API

Model Monitoring

- Data Drift
- Concept Drift

Model Re-Training dan Model Re-Deployment

TERIMA KASIH