
INSTITUT TEKNOLOGI BANDUNG
1920

Data Science

Informatics Research Group
School of Electrical Engineering and Informatics
Institut Teknologi Bandung

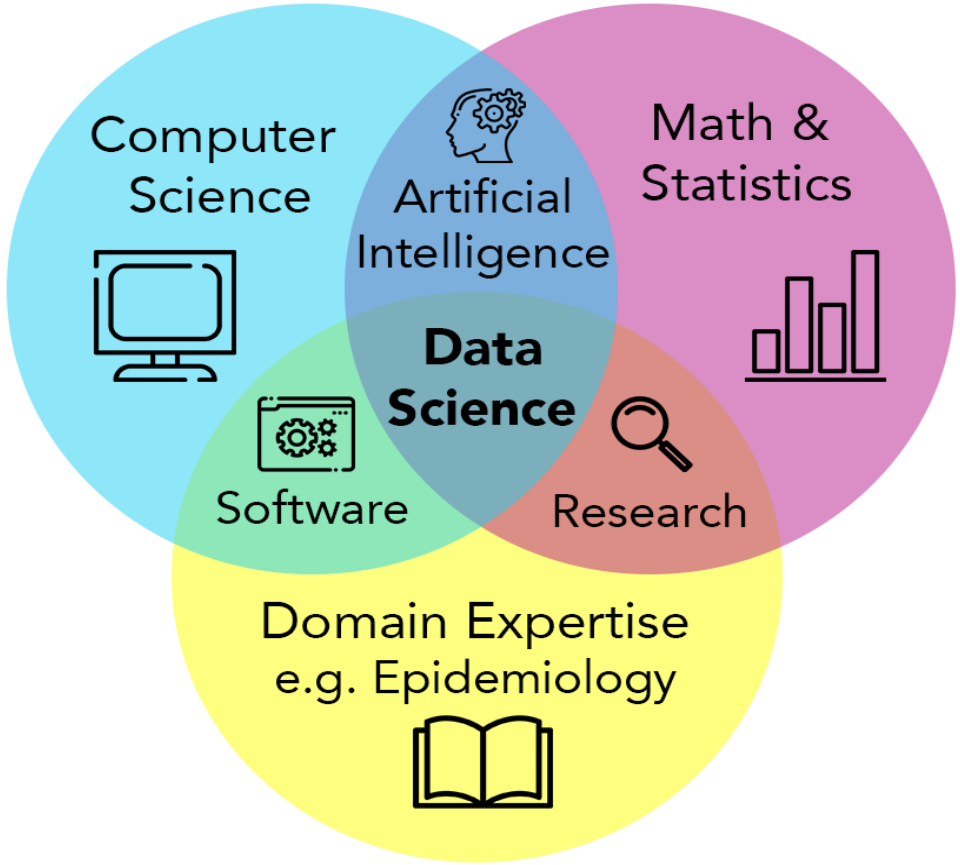
Sources:

Materi Pelatihan Associate Data Science – Pusat Artificial Intelligence ITB

EDUNEX ITB

Definisi Data Science

- Data science adalah bidang studi yang berfokus pada pemahaman, analisis, dan pengambilan keputusan berdasarkan data.
- Data science menggabungkan berbagai disiplin ilmu, termasuk statistika, matematika, komputer, dan pengetahuan domain, untuk menggali wawasan berharga dari data dan memecahkan masalah yang kompleks



Tujuan Task

01

Descriptive:
Menjelaskan keadaan bisnis saat ini melalui data historis.

02

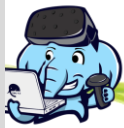
Diagnostic:
Menjelaskan mengapa suatu masalah terjadi dengan melihat data historis.

03

Predictive:
Memproyeksikan atau memprediksi hasil masa depan berdasarkan data historis.

04

Prescriptive:
Menggunakan hasil analitik prediktif dan pengetahuan lain dengan menyarankan upaya terbaik di masa depan.



Perlu Metodologi Pengembangan

Pengembangan Sistem AI berdasar data

≠

Data + Machine Learning (ML) Algorithms

Metodologi Pengembangan:

Metoda *iterative* yang dipakai untuk menyelesaikan masalah dengan menggunakan *data science* melalui urutan langkah yang ditentukan

Jenis Metodologi *Data Science*

1. Metodologi kegiatan Teknis

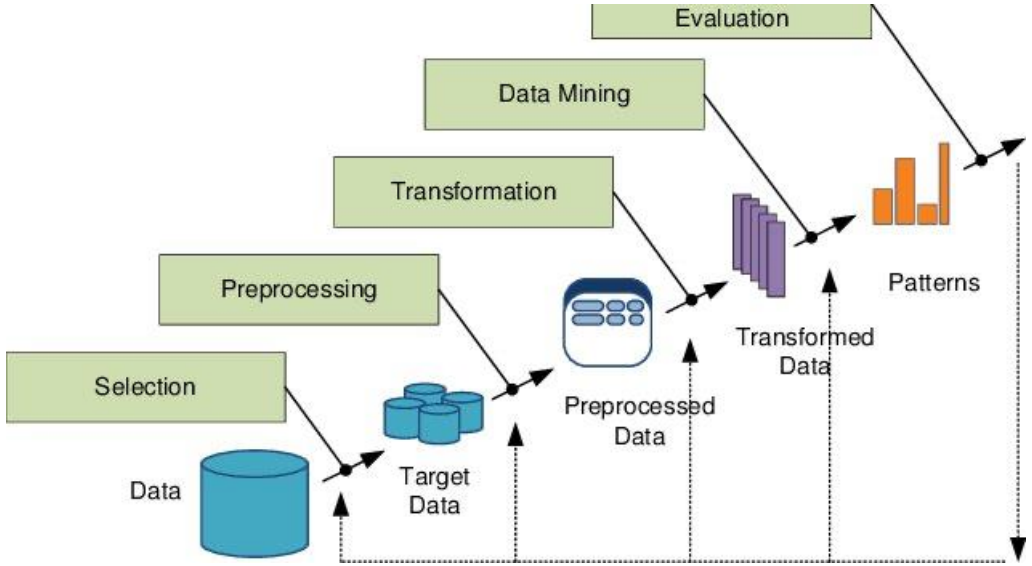
Kegiatan *Data Science* dianggap sebagai kegiatan teknis di bidang Teknologi Informasi ataupun Pengolahan data

2. Metodologi kegiatan bisnis

Kegiatan *Data Science* dianggap sebagai kegiatan bisnis yang terkait dengan penyelesaian suatu masalah organisasi menggunakan pendekatan data

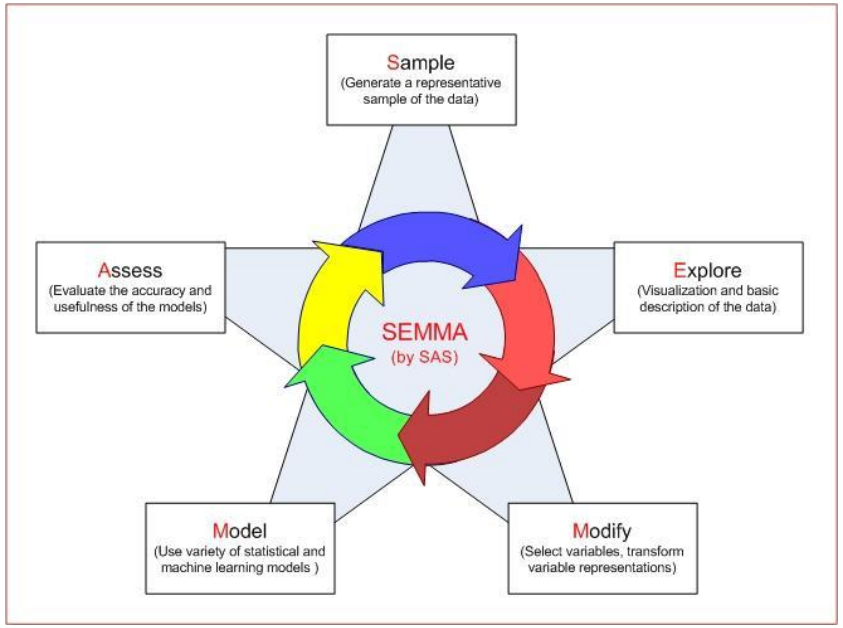
Metodologi Teknis
Kegiatan DS/AI dianggap Kegiatan Teknikal

Knowledge Discovery and Data Mining (KDD)



https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

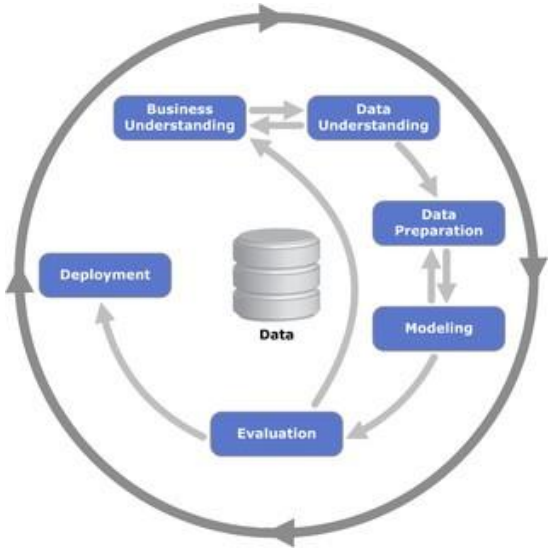
SEMMA dari SAS Institute



https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

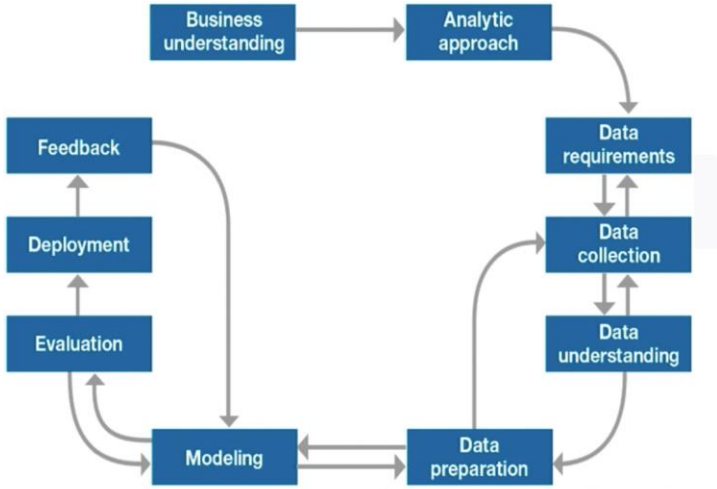
Metodologi Bisnis

CRISP-DM



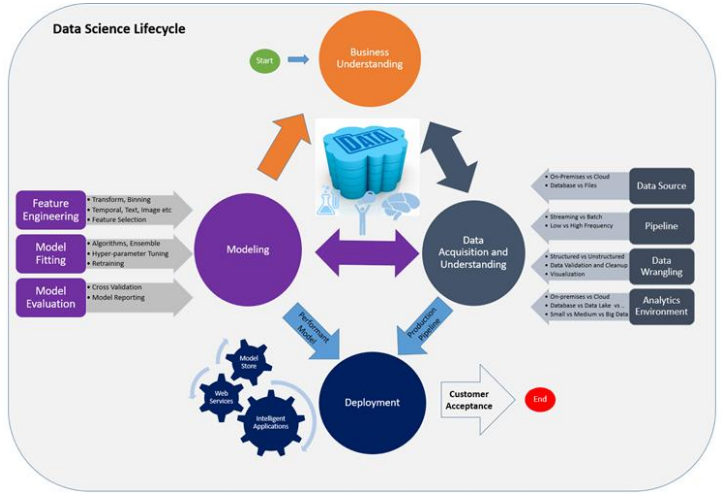
https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en

IBM Data Science Methodology



https://www.slideshare.net/JohnBRollinsPhD/foundational-methodology-for-data-science

Microsoft's Team Data Science Process



https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

Metodologi Data Science

Bagaimana di Indonesia?

MENTERI KETENAGAKERJAAN
REPUBLIK INDONESIA

KEPUTUSAN MENTERI KETENAGAKERJAAN
REPUBLIK INDONESIA
NOMOR 299 TAHUN 2020
TENTANG
PENETAPAN STANDAR KOMPETENSI KERJA NASIONAL INDONESIA
KATEGORI INFORMASI DAN KOMUNIKASI GOLONGAN POKOK AKTIVITAS
PEMROGRAMAN, KONSULTASI KOMPUTER DAN KEGIATAN YANG
BERHUBUNGAN DENGAN ITU (YBDI) BIDANG KEAHLIAN *ARTIFICIAL INTELLIGENCE* SUBBIDANG *DATA SCIENCE*

TUJUAN UTAMA	FUNGSI KUNCI	FUNGSI UTAMA	FUNGSI DASAR
Menemukan pengetahuan, *insight* atau pola yang bermanfaat dari data untuk berbagai keperluan (orang mengambil keputusan atau sistem memproses lebih lanjut)	Menganalisis Kebutuhan (Requirements) Organisasi	*Business Understanding*	1. Menentukan objektif bisnis 2. Menentukan tujuan teknis 3. Membuat rencana proyek
		Data Understanding	4. Mengumpulkan data 5. Menelaah data 6. Memvalidasi data
	Mengembangkan model	*Data Preparation*	7. Memilah data 8. Membersihkan data 9. Mengkonstruksi data 10. Menentukan Label Data 11. Mengintegrasikan data
		Modeling	12. Membangun skenario pengujian 13. Membangun model
		Model Evaluation	14. Mengevaluasi hasil pemodelan 15. Melakukan review proses pemodelan
	Menggunakan model yang dihasilkan	*Deployment*	16. Membuat rencana deployment model 17. Melakukan deployment model 18. Melakukan rencana pemeliharaan 19. Melakukan pemeliharaan
		Evaluation	20. Melakukan review proyek 21. Membuat laporan akhir proyek

Standard Kompetensi Kerja Nasional (SKKNI):

KepMen Ketenagakerjaan No 299 thn 2020

Siapa yang terlibat dalam Kegiatan Data Science

Data Scientist
Mengembangkan model terbaik dari data untuk menjawab permasalahan bisnis

Project/ Product Manager
Mengelola projek/ produk berbasis data.

Data Engineer
Menyiapkan (big) data untuk diolah/ dimodelkan

Domain Expert
Memberi arahan tentang domain permasalahan

Data Analyst
Menganalisis/ mencari insight dari data (dan menampilkannya dalam dashboard)

IT People
Menyiapkan infrastruktur IT (terutama deployment)

Kompetensi (Minimal) bagi seorang Data Scientist

TUGAS Menganalisis & mengembangkan model pengetahuan terbaik dari data

LEVEL KKNI 7

KODE UNIT	JUDUL UNIT KOMPETENSI
J.62DMI00.001.1	Menentukan Objektif Bisnis
J.62DMI00.002.1	Menentukan Tujuan Teknis DS
J.62DMI00.005.1	Menelaah Data
J.62DMI00.006.1	Memvalidasi Data
J.62DMI00.007.1	Menentukan Objek Data
J.62DMI00.008.1	Membersihkan Data
J.62DMI00.009.1	Mengkonstruksi Data
J.62DMI00.012.1	Membangun Skenario Model
J.62DMI00.013.1	Membangun Model
J.62DMI00.014.1	Mengevaluasi Hasil Pemodelan
J.62DMI00.015.1	Melakukan Proses Review Pemodelan

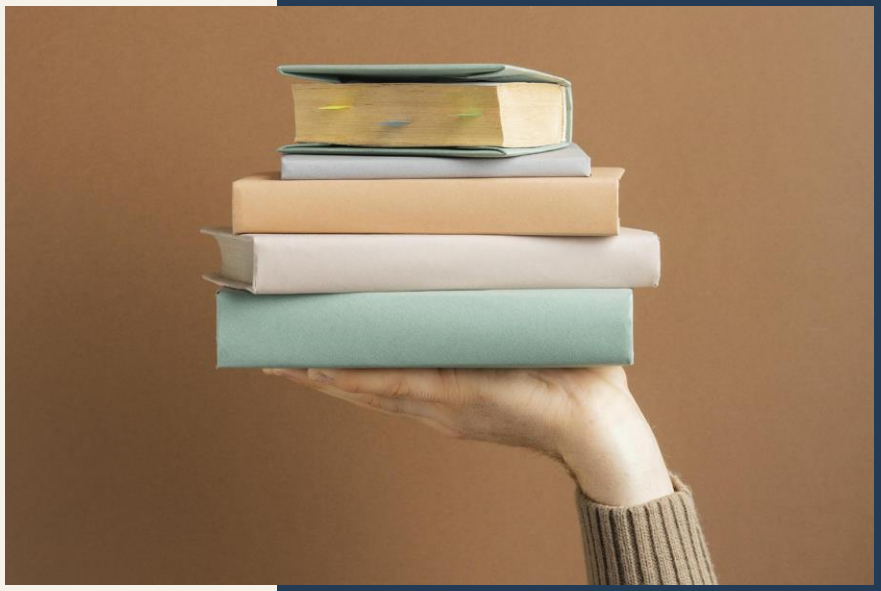
Kompetensi (Minimal) bagi seorang Data Analyst

TUGAS Membantu Data Scientist dalam mengembangkan pengetahuan, pola, dan insight dari data

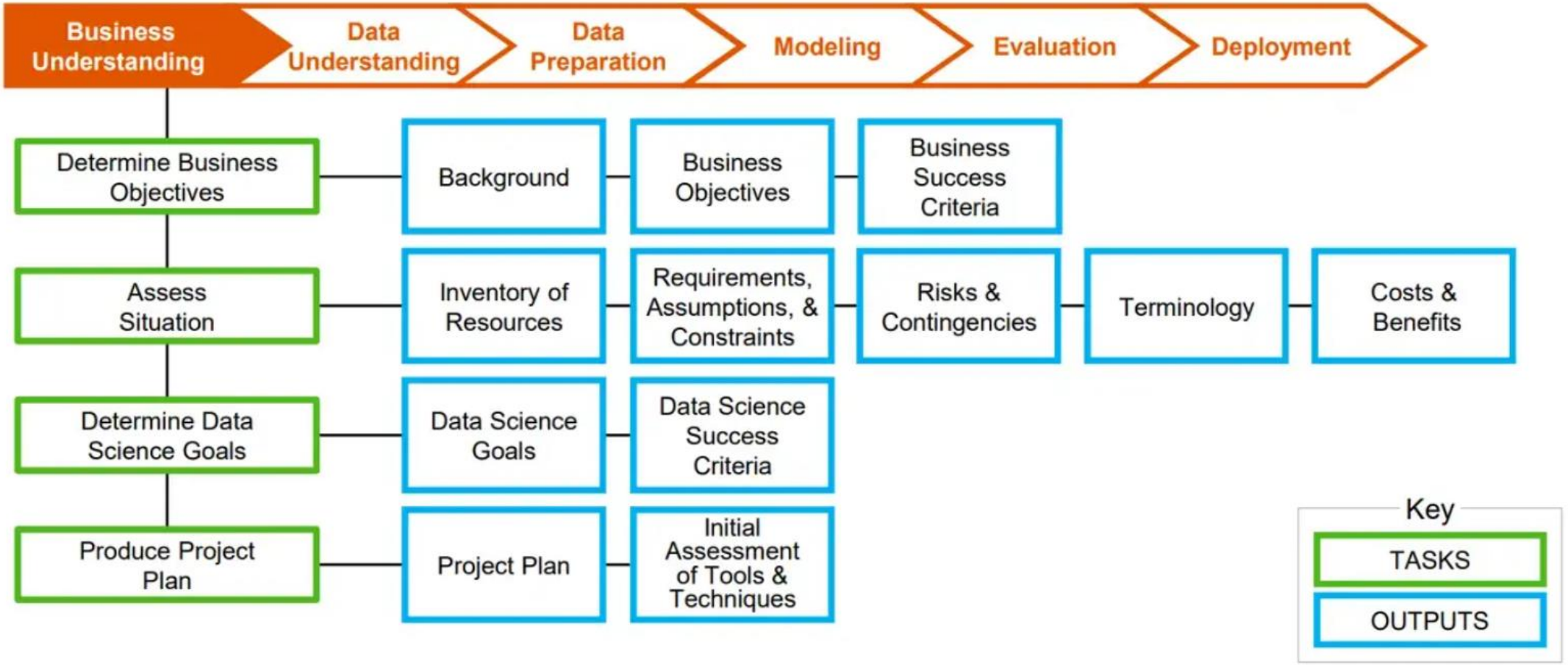
LEVEL KKNI 6

KODE UNIT	JUDUL UNIT KOMPETENSI
J.62DMI00.004.1	Mengumpulkan Data
J.62DMI00.005.1	Menelaah Data
J.62DMI00.006.1	Memvalidasi Data
J.62DMI00.007.1	Menentukan Objek Data
J.62DMI00.008.1	Membersihkan Data
J.62DMI00.009.1	Mengkonstruksi Data
J.62DMI00.010.1	Menentukan Label data
J.62DMI00.013.1	Membangun Model
J.62DMI00.014.1	Mengevaluasi Hasil Pemodelan

Langkah Pengembangan Aplikasi dengan Data Science



CRISP-DM – Phase 1: Business Understanding



© 2020 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

1. Business Understanding: Menentukan Masalah Bisnis

Kasus: Kegagalan Kredit

westonlegal.com

Problem

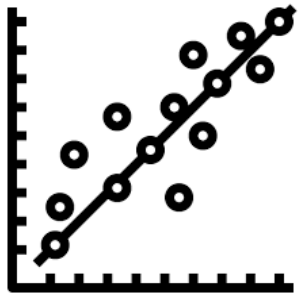
Bagaimana menurunkan NPL suatu bank?

Pertanyaan

Bagaimana memperbaiki perhitungan Credit score

Measurable outcomes

% Penurunan kredit gagal bayar

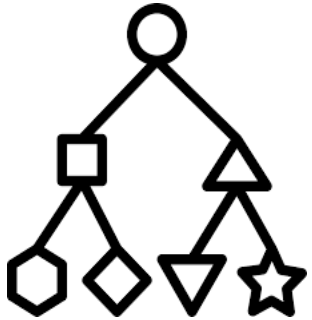


Regresi/Estimasi:

Memprediksi nilai kontinyu dari kasus

- Prediksi harga rumah berdasar karakteristik tertentu
- Prediksi harga saham besok

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE), dll



Klasifikasi:

Memprediksi kelas/ kategori dari kasus

- Prediksi kolektibilitas suatu pinjaman
- Prediksi kebangkrutan suatu perusahaan di tahun depan

- Recall, Precision
- F1 Score, dll

Klastering:

Mengelompokkan kasus berdasar kemiripan

- Segmentasi nasabah perbankan
- Pengelompokkan pasien yang mirip kasusnya

- Silhouette Score
- Davies-Bouldin Index (DBI), dll

2. Data Understanding :
Mengenali/ mendalami data yang dimiliki

Mengumpulkan Data
Mengumpulkan Data yang Diperlukan

Menelaah data
Menganalisa data secara eksploratif

Memvalidasi Data
Menilai kesesuaian kualitas data dengan masalah yang akan dipecahkan

2. Data Understanding : Mengumpulkan data

Mengumpulkan Data yang Diperlukan

Jumlah Data: Berapa banyak yang dapat diperoleh

Deskripsi Data: Penjelasan arti atribut/ fitur

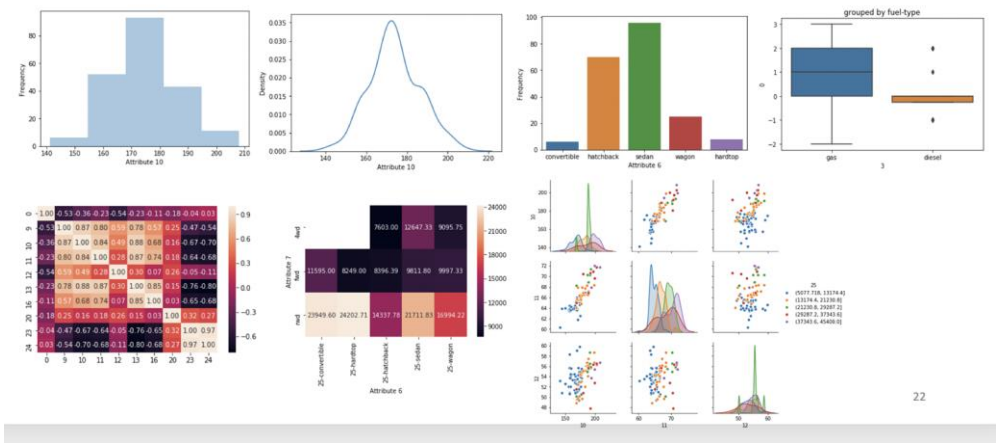
ID	CONTROL ACCOUNTS	PREDECESSOR	DURATION	START	FINISH	RESOURCES
1	Notify important stakeholders		2 Weeks	08/01/2022	08/12/2022	Marketing team, external venue, catering service, sound equipment, facilitator
2	Identify new business name	1	2 Weeks	08/15/2022	08/26/2022	Marketing team, external legal service
3	Design new business name	2	4 Weeks	08/29/2022	09/23/2022	Marketing team, expert designer
4	Trademark name	3	3 Weeks	09/26/2022	10/14/2022	External legal service
5	Implement new name	4	4 Weeks	10/17/2022	11/11/2022	Marketing team, domain host, printing company, signage company
6	Rebranding VIP happening	4	3 Weeks	10/17/2022	11/04/2022	Marketing team, management & staff, external venue, catering service, music entertainment, printing company
7	Evaluation & closure	5, 6	1 Week	11/14/2022	11/18/2022	Marketing team, management & staff, external venue, catering service, facilitator

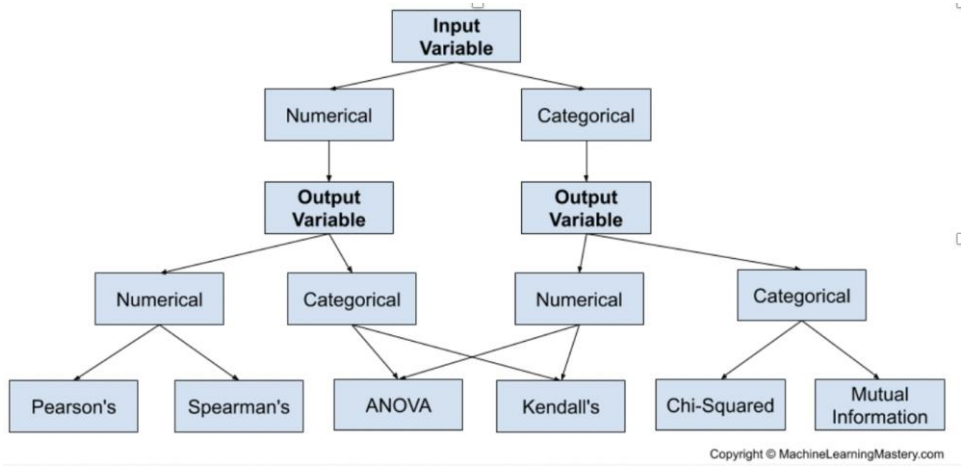
2. Data Understanding : Menelaah Data

Menganalisa data secara eksploratif (EDA)

Karakteristik Atribut: Deskripsi data (atribut) yang diperoleh

Keterkaitan antar Data: Analisis statistik korelasi, Anova, Chi-Squared,...





2. Data Understanding : Memvalidasi Data

Menilai kesesuaian kualitas data dengan masalah yang akan dipecahkan

Laporan Kualitas Data:

- Ukuran Data (Atribut/ fitur dan Jumlah record)
- Deskripsi statistical atribut
- Relasi antar atribut (dan label)
- Visualisasi data

3. Data Preparation

Memperbaiki kualitas data untuk Pemodelan

Memilih dan memilah data

Memilih data yang akan dipergunakan

→

- Record terpakai
- Atribut terpakai

Membersihkan Data

Meminimalkan noise (tidak lengkap, salah)

→

- Data lengkap
- Data yang diperbaiki
- Data Pecilan

Mengkonstruksi data

Menambahkan fitur dan transformasi data

→

- Fitur tambahan (Feature Engineering)
- Transformasi data (standardisasi, transformasi)

Integrasi Data

Menggabungkan data

4. Modeling : Mengembangkan Model (Pengetahuan)

Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

A. Memilih Algoritma : **Disesuaikan dengan Tugas Analytics yang dipilih**

1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

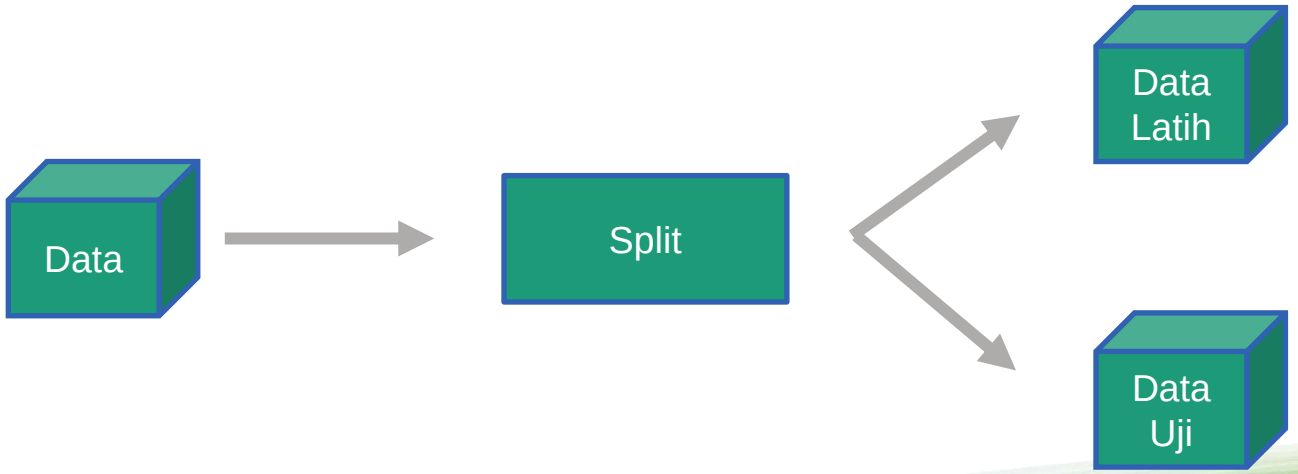
4. Modeling : Mengembangkan Model (Pengetahuan)

Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

B. Membagi data: Sesuai dengan ketersediaan data

1. Data Latih: Untuk mengembangkan model
2. Data Uji: Untuk Mengukur performansi model



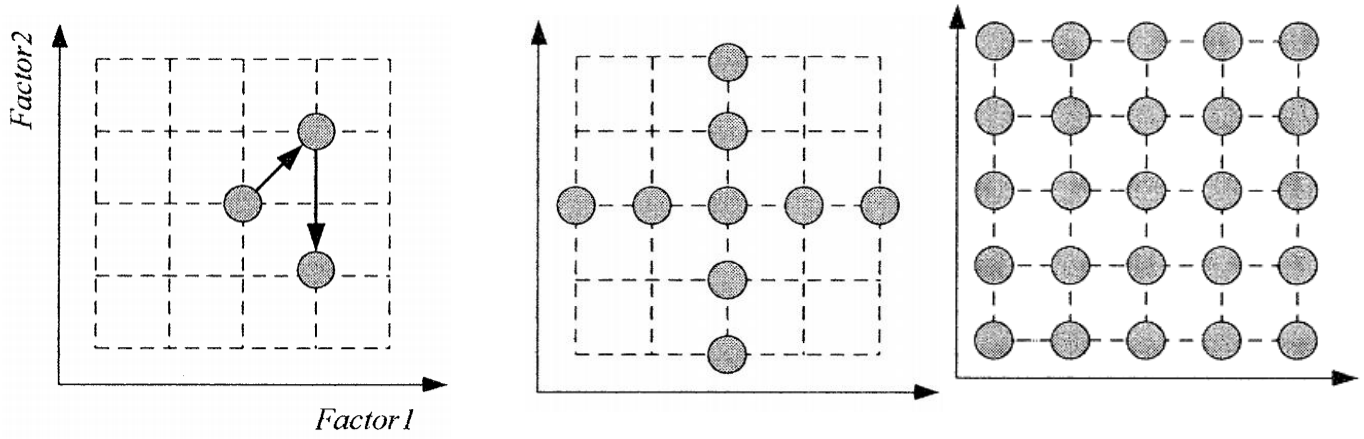
4. Modeling : Mengembangkan Model (Pengetahuan)

Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

C. Menentukan Langkah Eksperimen:

Untuk mendapatkan model terbaik secara efisien dan efektif



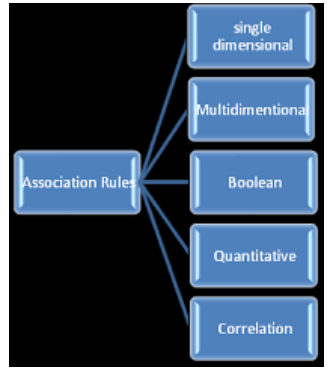
Best Guess **One Factor at A Time** **Grid Search**

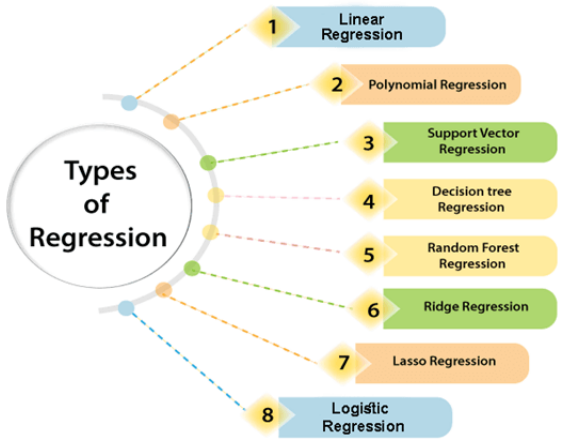
4. Modeling : Mengembangkan Model (Pengetahuan)

Membangun model

Mengembangkan model dengan Teknik ML

Pemilihan Algoritma Machine Learning (ML)
Pembagian Data
Penentuan Langkah Eksperimen





Tidak ada Algoritma yang SELALU TERBAIK untuk setiap dataset! Coba beberapa algoritma!!

4. Modeling : Mengembangkan Model (Pengetahuan)

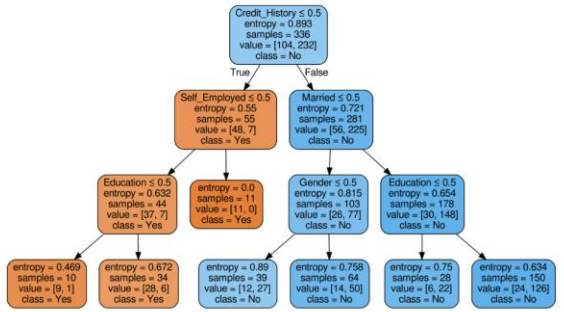
Membangun model

Mengembangkan model dengan Teknik ML

A. Proses Pelatihan : Untuk mendapatkan model

Variable	Type	Definition
BAD	Num	BAD: 1 = applicant defaulted on loan or seriously delinquent; 0 = applicant paid loan
LOAN	Num	LOAN: Amount of the loan request
MORTDUE	Num	MORTDUE: Amount due on existing mortgage
VALUE	Num	VALUE: Value of current property
REASON	Char	REASON: DebtCon = debt consolidation; HomeImp = home improvement
JOB	Char	JOB: Occupational categories
YOJ	Num	YOJ: Years at present job
DEROG	Num	DEROG: Number of major derogatory reports
DELINQ	Num	DELINQ: Number of delinquent credit lines
CLAGE	Num	CLAGE: Age of oldest credit line in months
NINQ	Num	NINQ: Number of recent credit inquiries
CLNO	Num	CLNO: Number of credit lines
DEBTINC	Num	DEBTINC: Debt-to-income ratio

1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

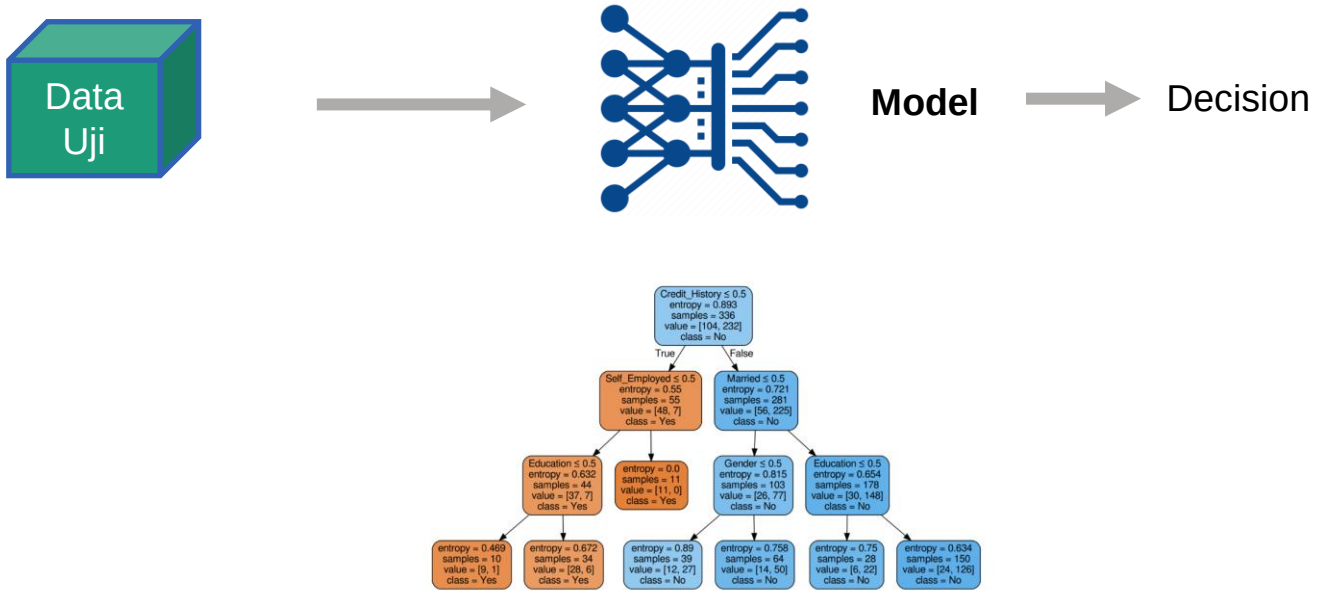


4. Modeling : Mengembangkan Model (Pengetahuan)

Membangun model

Mengembangkan model dengan Teknik ML

B. Proses Pengujian : Untuk mengukur Performansi



Metrik Performansi

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

5. Model Evaluation

Mengevaluasi Performansi Model Yang Dihasilkan

Mengevaluasi Model

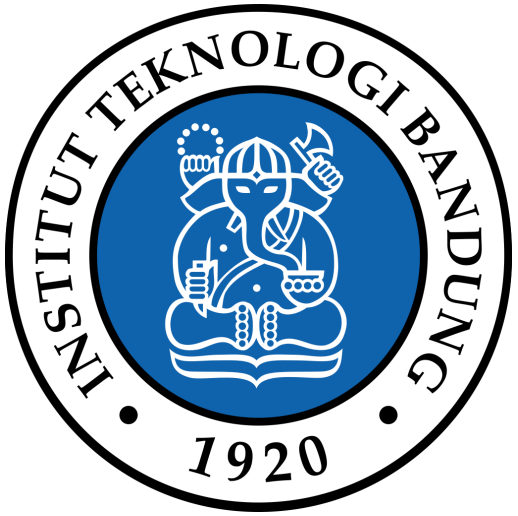
Mengukur performansi model

Performansi Capaian vs Target
Memilih Model terbaik

Mengevaluasi Proses

Menilai apakah proses sudah maksimal

Review Proses untuk mencari batasan atau kekurangan model



Salam

Semoga Bermanfaat