DIGITAL TALENT SCHOLARSHIP

THEMATIC ACADEMY

Data Scientist: Artificial Intelligence untuk Dosen dan Instruktur

Pertemuan #3 : Metodologi Data Science

KOMINFO

#JADIJAGOANDIGITAL

Badan Penelitian dan Pengembangan Sumber Daya Manusia

Deskripsi Pelatihan

Tujuan utama dari modul pelatihan ini adalah untuk membahas metodologi data science secara umum untuk mengembangkan suatu aplikasi AI dengan menjelaskan langkah-langkah utama yang diperlukan untuk menyelesaikan suatu masalah organisasi/ bisnis dengan melakukan tugas-tugas yang umumnya terkait dengan data science.

Capaian Pembelajaran

Pada topik ini, kita akan mempelajari:

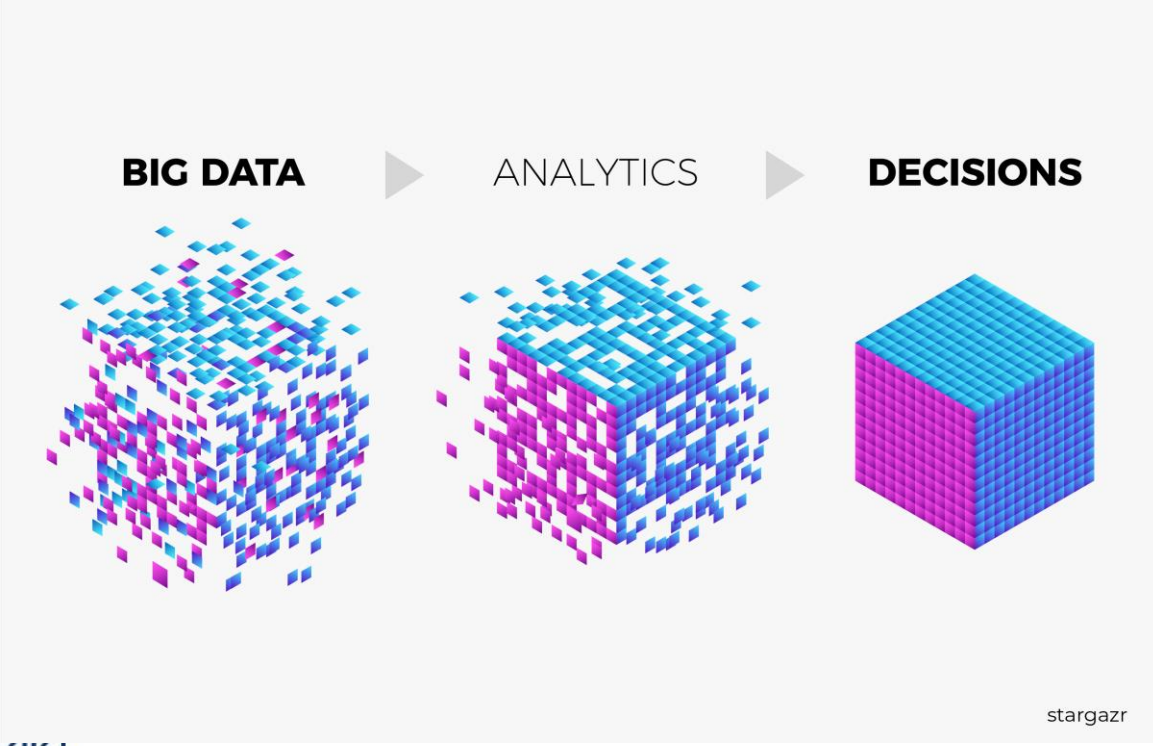
- Metodologi *Data Science*
- Langkah-langkah utama dalam metodologi data science

Agenda

- **Mengapa Metodologi diperlukan**
  - Mengapa Mayoritas Projek AI Gagal
- **Berbagai Metodologi Data Science**
  - Tak semua metodologi sama lengkap
- **Langkah Pengembangan**
  - Dari Masalah Bisnis menjadi Aplikasi AI

Mengapa Metodologi diperlukan

Sistem AI berbasis (Big) Data



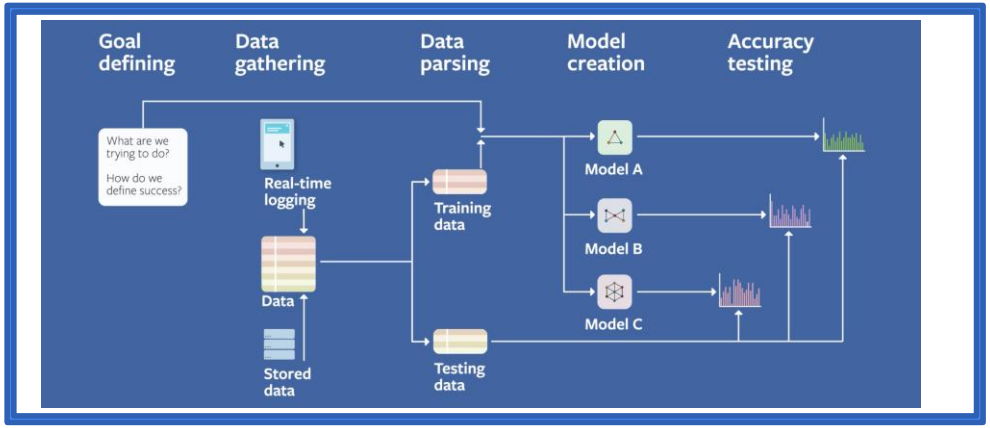
Data

Menjadi

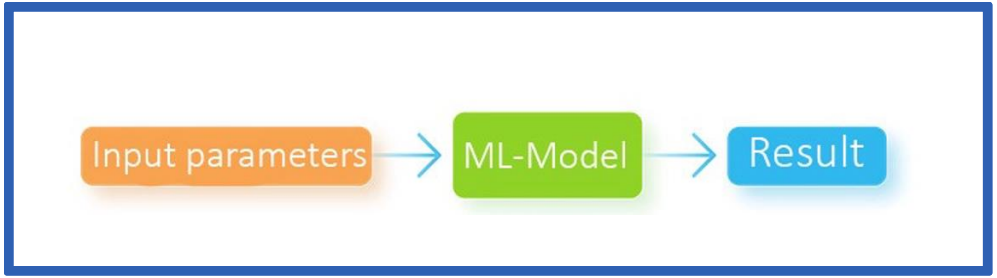
Sistem Intelijen
(berbasis Pengetahuan)

Sistem AI berbasis (Big) Data dikembangkan dalam 2 tahap

1. Pengembangan (Pelatihan)



2. Penggunaan



Tujuan Tugas/ Task yang Biasa Dikembangkan

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.

Jenis Task yang Dikembangkan

- Regression / Estimation
- Classification
- Clustering
- Association
- Anomaly Detection
- Sequence Mining
- Recommendation Systems

Mayoritas Proyek Pengembangan AI/DS Gagal

GARTNER ESTIMATED

85%

of big data projects fail (2017). The initial estimation was 60% (GARTNER 2016)

THROUGH 2020

80%

of AI projects will remain alchemy, run by wizards whose talents will not scale in the organization. (GARTNER 2018)

THROUGH 2022

20%

of analytic insights will deliver business outcomes. (GARTNER 2018)

EXECUTIVE SURVEY

77%

respondents say that "business adoption" of big data and AI initiatives continues to represent a challenge for their organizations (NEWVANTAGE PARTNERS 2019)

https://www.slideshare.net/PMI-Montreal/symposium-2019-gestion-de-projet-en-intelligence-artificielle

Mayoritas Proyek Pengembangan AI/DS Gagal

- PROBLEM yang akan diselesaikan
  - Tidak Jelas; Problem salah; Over promising
- DATA
  - Tidak cukup (jumlah) atau tidak tepat (variabel)
  - Kualitas, tidak mencukupi
  - Tidak mengerti arti (semantic) data
  - Berbagai bias, hubungan antar variabel tidak dipikirkan (sampling, Fairness)
- MODEL yang dikembangkan
  - Terlalu kompleks; Tidak dimengerti
  - Metriks pengukuran tidak tepat
- ALGORITHMS
  - Terlalu sophisticated; Tidak dimengerti secara teknis
  - Tidak tepat
- SUMBER DAYA MANUSIA
  - One man show
  - Dukungan pemangku kepentingan kunci kurang

Perlu Metodologi Pengembangan

Pengembangan Sistem AI berdasar data

≠

Data + Machine Learning (ML) Algorithms

Metodologi Pengembangan

Metoda iterative yang dipakai untuk menyelesaikan masalah dengan mengguna-kan data dan data science melalui urutan langkah yang ditentukan

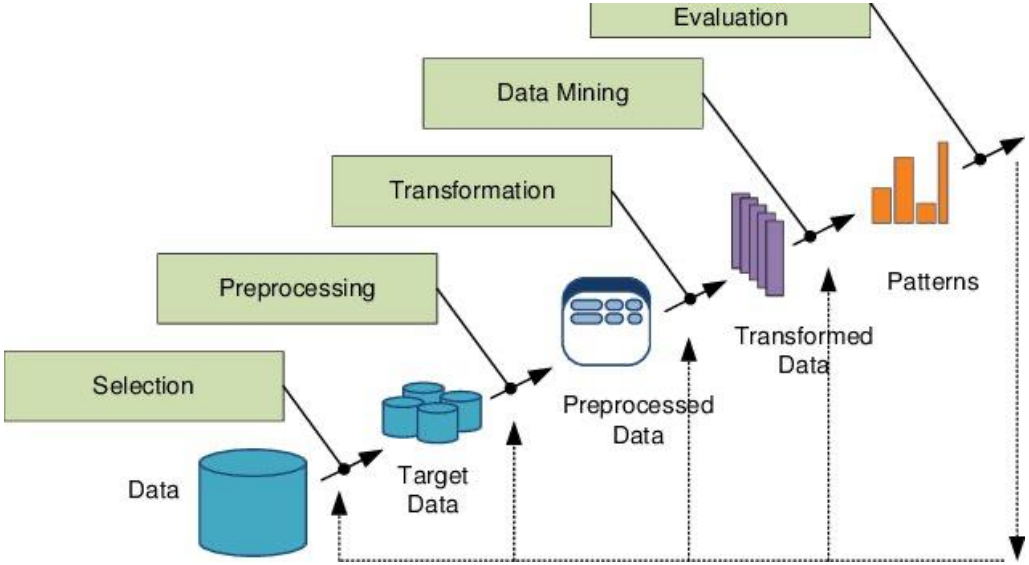
Berbagai Metodologi Data Science

Jenis Metodologi

- Metodologi kegiatanTeknis
- Metodologi kegiatan bisnis (dan teknis)

Metodologi Teknis: Kegiatan DS/AI dianggap Kegiatan Teknikal

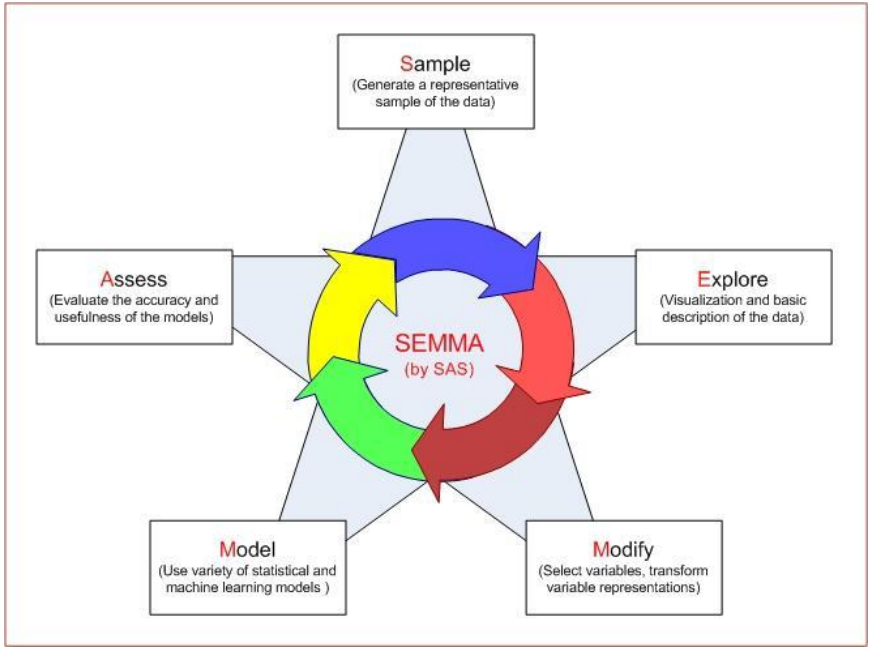
- Knowledge Discovery and Data Mining



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

Metodologi Teknis: Kegiatan DS/AI dianggap Kegiatan Teknikal

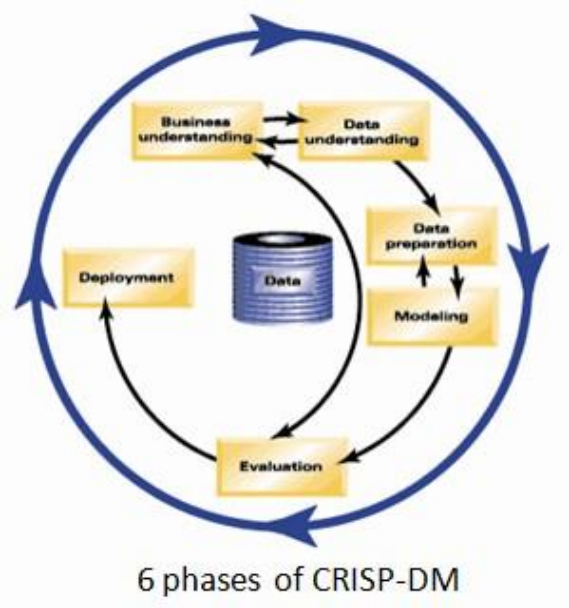
- SEMMA dari SAS Institute



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

Metodologi Lengkap: Kegiatan DS/AI dianggap Kegiatan Bisnis: Masalah Bisnis menjadi Masalah DS/AI

- CRISP-DM

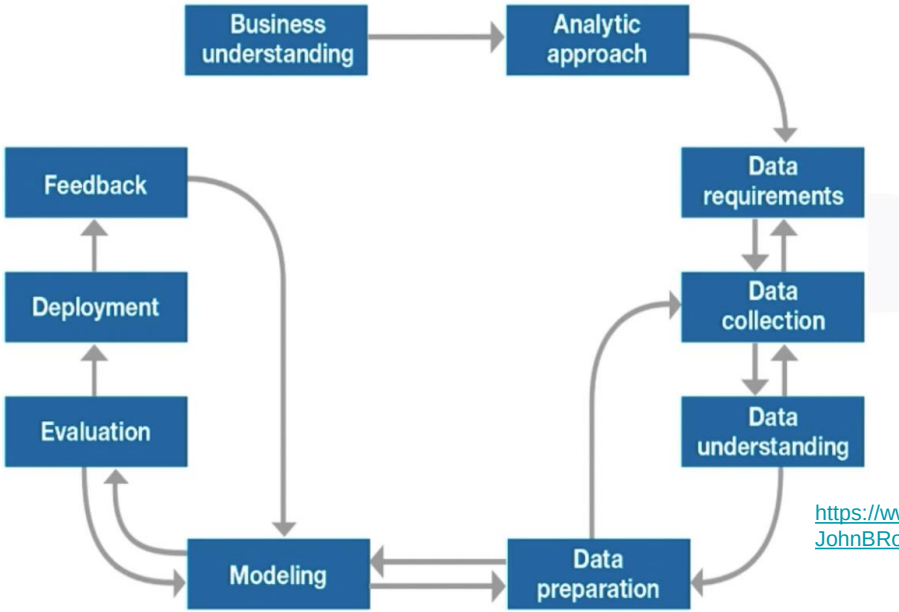


6 phases of CRISP-DM

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

Metodologi Lengkap: Kegiatan DS/AI dianggap Kegiatan Bisnis: Masalah Bisnis menjadi Masalah DS/AI

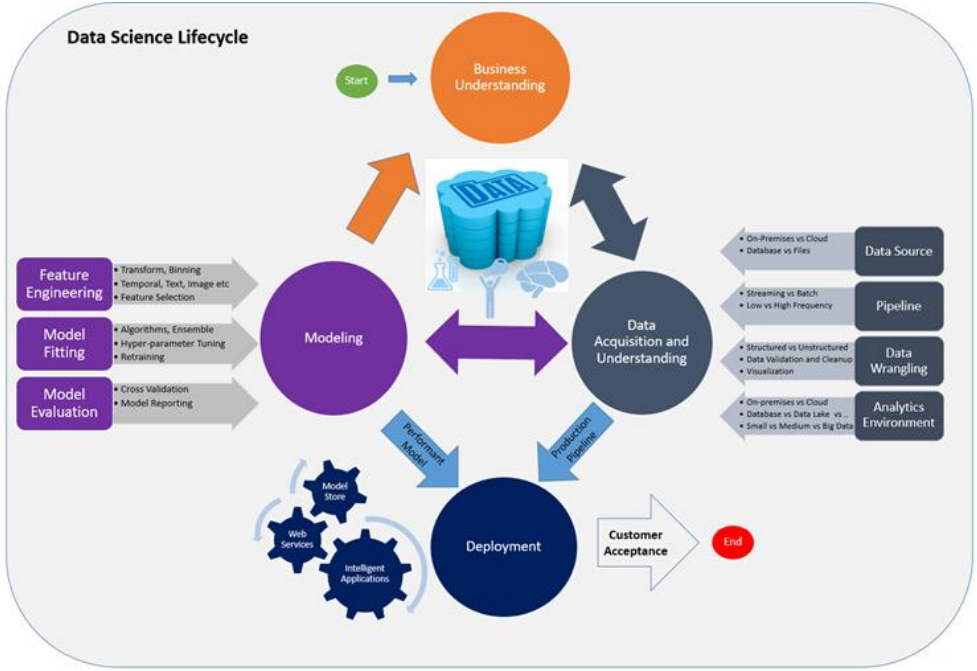
- IBM Data Science Methodology



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

Metodologi Lengkap: Kegiatan DS/AI dianggap Kegiatan Bisnis: Masalah Bisnis menjadi Masalah DS/AI

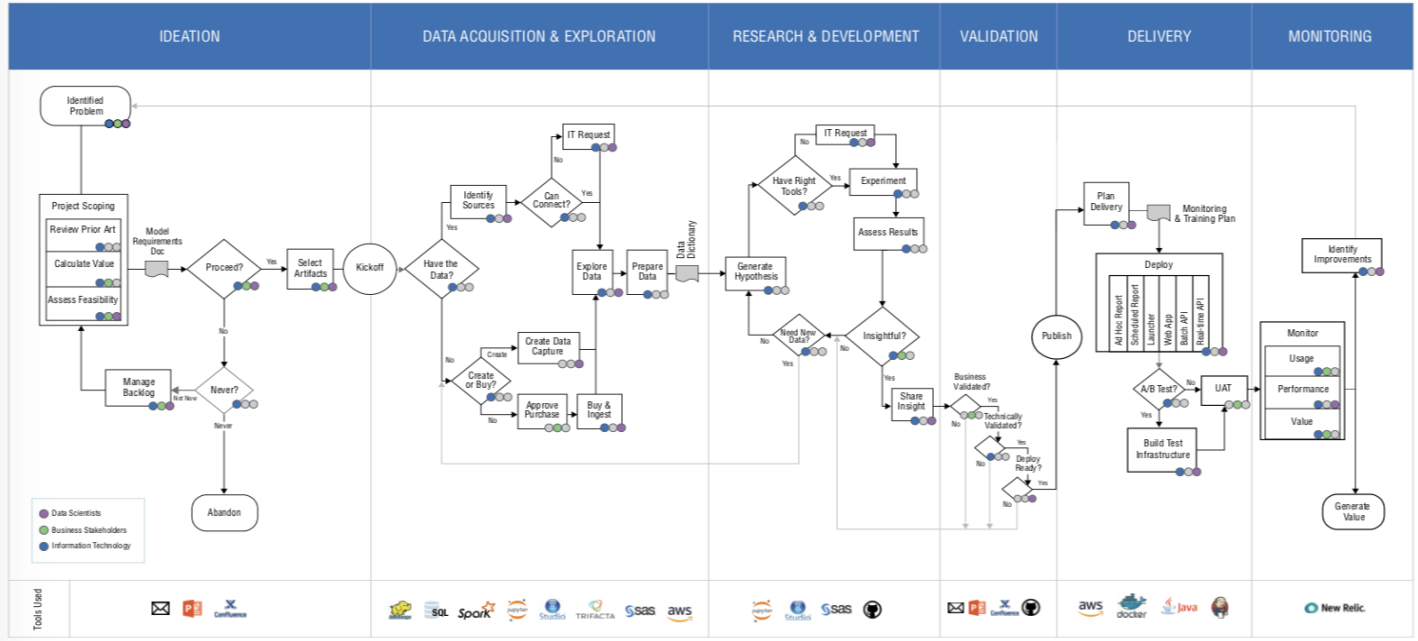
- Microsoft's Team Data Science Process



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

Metodologi Lengkap: Kegiatan DS/AI dianggap Kegiatan Bisnis: Masalah Bisnis menjadi Masalah DS/AI

- Domino DataLab Methodology



https://www.dominodatalab.com/

Bagaimana di Indonesia?

Standard Kompetensi Kerja Nasional:
KepMen Ketenagakerjaan No 299 thn 2020

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

Tim Pengembang: Kegiatan Bersama

01 **Data Scientist**
Mengembangkan model terbaik dari data untuk menjawab permasalahan bisnis

02 **Data Engineer**
Menyiapkan (big) data untuk diolah/ dimodelkan

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

04 **Project/ Product Manager**
Mengelola projek/ produk berbasis data.

05 **Domain Expert**
Memberi arahan tentang domain permasalahan

06 **IT People**
Menyiapkan infrastruktur IT (terutama deployment)

Langkah Pengembangan

1. Business Understanding: Menentukan Masalah Bisnis

Kasus: Kegagalan Kredit

westonlegal.com

Problem:

Bagaimana menurunkan NPL suatu bank

Pertanyaan:

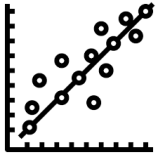
Bagaimana memperbaiki perhitungan Credit score

Measurable outcomes:

% Penurunan kredit gagal bayar

1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis?

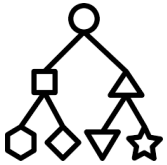


A. Regresi/Estimasi: Memprediksi nilai kontinyu dari kasus

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

1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis?



B. Klasifikasi: Memprediksi kelas/ kategori dari kasus

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

1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan
untuk menjawab permasalahan bisnis?

C. Klastering: Mengelompokkan kasus berdasar kemiripan

- Segmentasi nasabah perbankan
- Pengelompokkan pasien yang mirip kasusnya

1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis?

D. Asosiasi: Memprediksi kumpulan item/ kejadian yang biasa terjadi bersama

- Mencari barang jualan yang biasa dibeli bersama
- Menyusun portofolio saham

1. Business Understanding: Menentukan Tugas Analytics

A. Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis?

E. Anomali Detection: Menemukan kasus abnormal/ tidak biasa terjadi

- Pendeteksian transaksi illegal penggunaan kartu kredit
- Pendeteksian penerobosan jaringan

1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan
untuk menjawab permasalahan bisnis?

F. Sequence Mining: Memprediksi apa yang akan terjadi dari keadaan saat ini

- Prediksi apakah nasabah akan berhenti berlangganan
- Menentukan alur pada transaksi e-commerce

1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis?

G. Rekomendasi: Memberikan rekomendasi pengguna berdasar asosiasi preferensi dengan pengguna lain yang memiliki 'taste' yang sama

- Rekomendasi film untuk ditonton
- Rekomendasi saham untuk dibeli

1. **Business Understanding:** Menentukan Tugas Analytics

Pengukuran Performansi tergantung Jenis Task Analytics

Metriks Performansi: Ukuran keberhasilan dari proses data science yang dilakukan

Contoh: Root Mean Squared Error (RMSE)
R-Square
Jackard Index
Log-loss
Precision
Recall
F1-Score

1. Business Understanding: Menentukan Tugas Analytics

Kasus: Kegagalan Kredit

Apa Tugas Analitiks yang perlu diselesaikan untuk menjawab permasalahan bisnis tersebut?

westonlegal.com

Problem:

Bagaimana menurunkan NPL suatu bank

Pertanyaan:

Bagaimana memperbaiki perhitungan Credit score

Tugas Analitik:

Klasifikasi

Performance Metrics:

F1-Score

1. Business Understanding: Menentukan Kebutuhan Data

Data apa yang diperlukan?
Dari mana bisa diperoleh?

Struktur Data: Bagaimana deskripsi data (atribut) yang diperlukan

Jumlah Data: Berapa banyak (record) data yang diperlukan

Sumber Data: Darimana data bisa diperoleh? Apakah sudah tersedia?

- Internal: Sistem Informasi/ ERP, Excel, dokumen
- Eksternal: Web API, Web Scraping
- Dataset via public data
- Dataset via open data

1. Business Understanding: Merencanakan Manajemen Projek

Bagaimana rencana pelaksanaan projeknya?

Cost Benefit Analysis: Apakah menguntungkan untuk melakukannya?

Situation Assessment: Analisa keadaan organisasi

Project Plan: Scope (WBS), Time, Schedule, Tim Pengembang

2. Data Understanding :

Mengenali/ mendalami data yang dimiliki

01 **Mengumpulkan Data**
Mengumpulkan Data yang Diperlukan

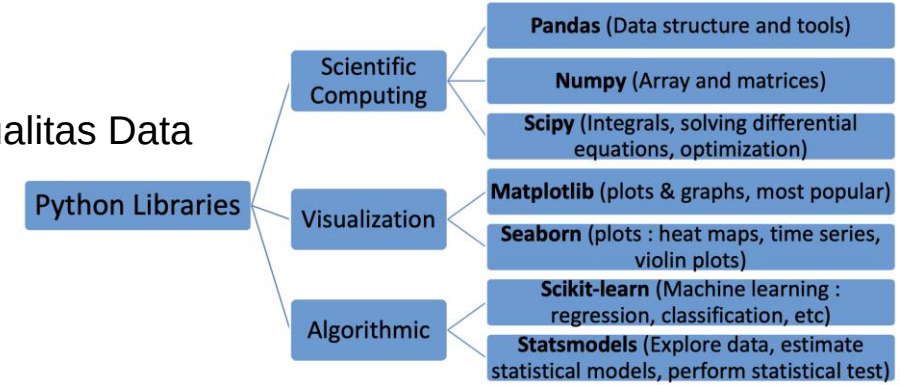
Jumlah Data (Baris dan Kolom)
Deskripsi data

02 **Menelaah data**
Menganalisa data secara eksploratif

Karakteristik atribut/ fitur
Keterkaitan antar data

03 **Memvalidasi Data**
Menilai kesesuaian kualitas data dengan masalah yang akan dipecahkan

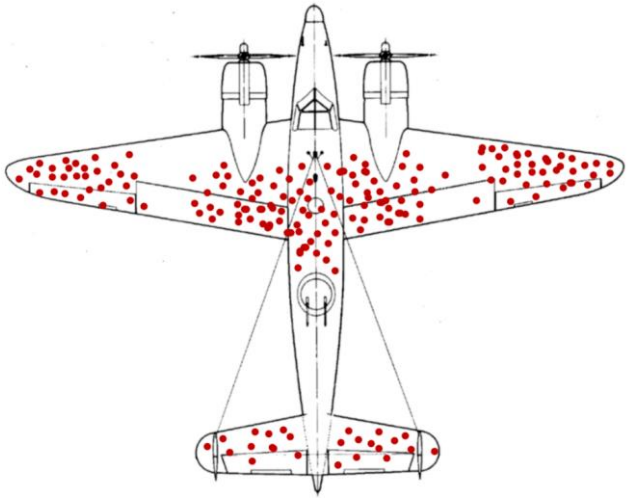
Kualitas Data



2. Data Understanding :

Mengapa Perlu Mengenali/ mendalami data yang dimiliki

- The United States armed forces faced a dilemma during the war, because returning bomber planes were riddled with bullet holes and they needed better ways to protect them
- “Where should they put it?”
- When they plotted out the damage these planes were incurring, it was spread out, but largely concentrated around the tail, body and wings.
- Should they upgrade these sections?



2. Data Understanding : Mengumpulkan Data

Mengumpulkan Data yang Diperlukan

Jumlah Data: Berapa banyak yang dapat diperoleh

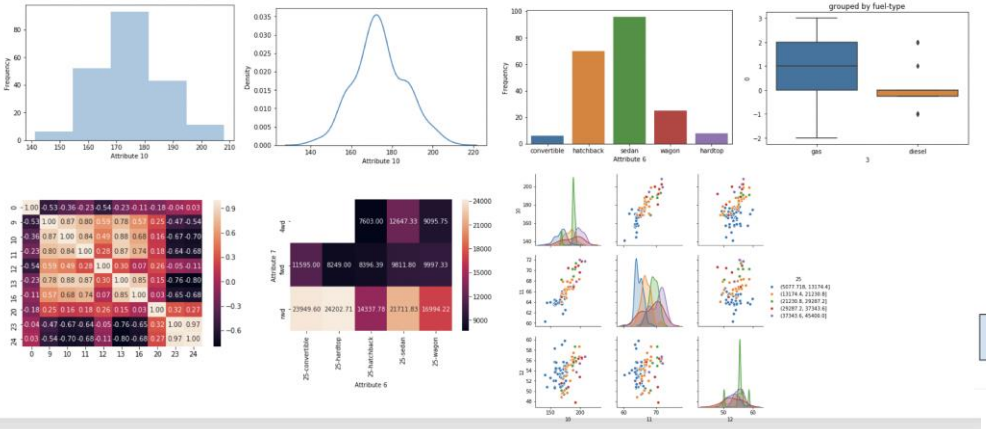
Deskripsi Data: Penjelasan arti atribut/ fitur

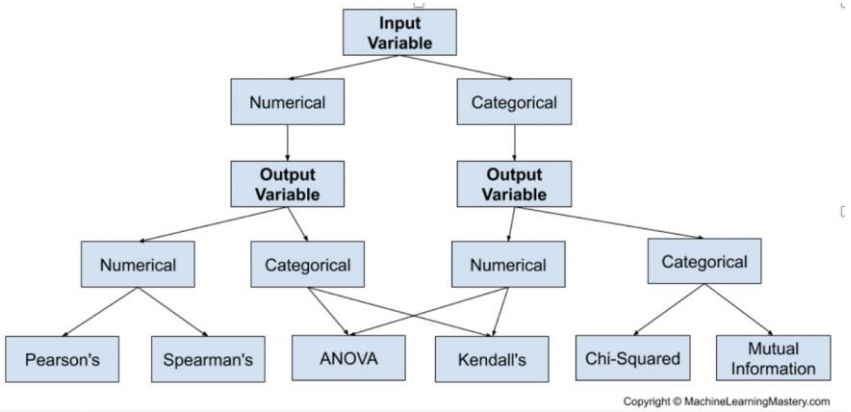
2. Data Understanding : Menelaah Data

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 de-
ngan 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

01	**Memilih dan memilah data** Memilih data yang akan dipergunakan	Rekord terpakai Atribut terpakai
02	**Membersihan Data** Meminimalkan noise (tidak lengkap, salah)	Data lengkap Data yang diperbaiki Data Pecilan
03	**Mengkonstruksi data** Menambahkan fitur dan transformasi data	Fitur tambahan (Feature Engineering) Transformasi data (standardisasi, transformasi)
04	**Integrasi Data** Menggabungkan data	Gabungan data

4. Modeling :
Mengembangkan Model (Pengetahuan)

01

Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

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

02

Membangun model

Mengembangkan model dengan Teknik ML

Eksekusi Algoritma
Pengaturan Parameter
Pengukuran Performance Metrics

4. Modeling : Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

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

4. Modeling : Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

A. Memilih Algoritma: Disesuaikan dengan Tugas Analytics yang dipilih

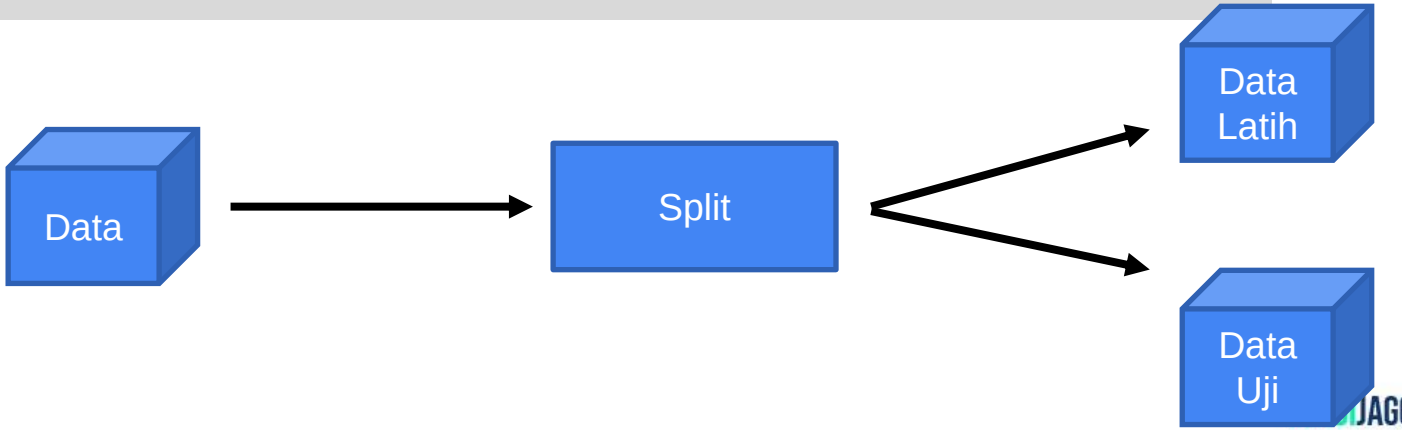
1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

4. Modeling : Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

B. Membagi data: Sesuai dengan ketersediaan data

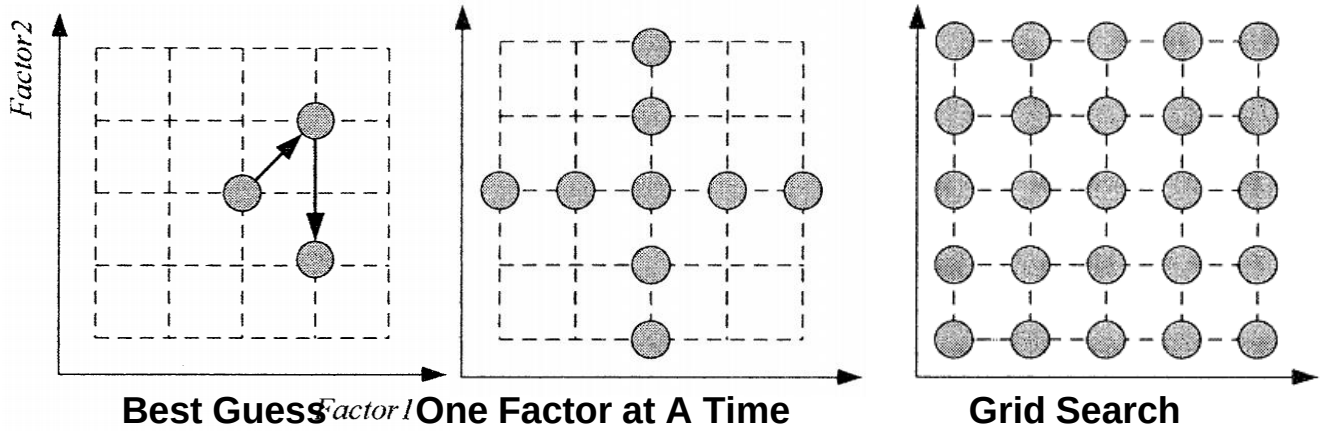
1. Data Latih: Untuk mengembangkan model
2. Data Uji: Untuk Mengukur performansi model



4. Modeling : Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

C. Menentukan Langkah Eksperimen: Untuk mendapatkan model terbaik secara efisien dan efektif



4. Modeling : Membangun model

Mengembangkan model dengan Teknik ML

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

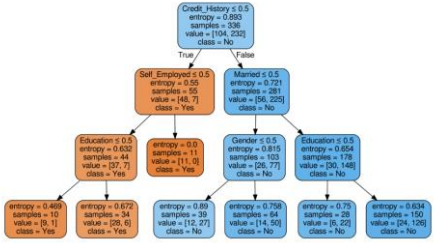
4. Modeling : Membangun model

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

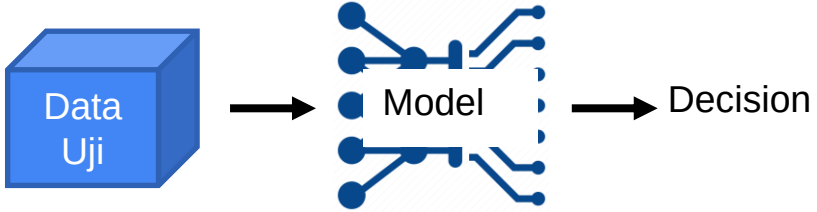
1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

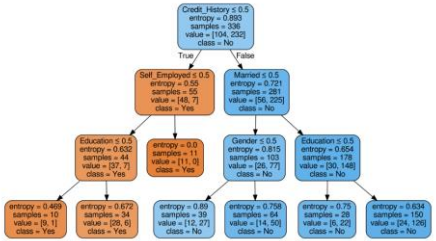


4. Modeling : Membangun model

Mengembangkan model dengan Teknik ML

B. Proses Pengujian : Untuk mengukur 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

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

5. Model Evaluation

Mengevaluasi Performansi Model Yang Dihasilkan

01 **Mengevaluasi Model**
Mengukur performansi model

Performansi Capaian vs Target
Memilih Model terbaik

02 **Mengevaluasi Proses**
Menilai apakah proses sudah maksimal

Review Proses untuk mencari batasan atau kekurangan model

Summary

Pada topik ini, kita sudah mempelajari:

- Langkah-langkah utama dalam menggunakan data untuk membuat suatu aplikasi AI berdasar metodologi data science
- Pengembangan sistem Ai berdasar data bukan hanya masalah teknis (terkait data) namun merupakan masalah bisnis/ organisasi
- Pengembangan sistem melibatkan Pakar Domain, Pakar Data Science/ AI, Pakar Manajemen Projek, dan Pakar TI dalam satu Tim

Tools / Lab Online

Referensi

- Standard Kompetensi Kerja Nasional Indonesia Bidang AI sub bidang Data Science
  - https://skkni.kemnaker.go.id/tentang-skkni/dokumen
- CRISP-DM
  - http://crisp-dm.eu/
- IBM Data Science Methodology
  - https://www.slideshare.net/JohnBRollinsPhD/foundational-methodology-for-data-science
- Microsoft Methodology
  - https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview
- Domino Methodology
  - https://www.dominodatalab.com/

Team Teaching

- Windy Gambetta, Ir., MBA (Institut Teknologi Bandung)
  - Email: windy@staff.stei.itb.ac.id

Quiz / Games

- Quiz dapat diakses melalui LMS (https://lms.kominfo.go.id/)

#JADIJAGOANDIGITAL

TERIMA KASIH

digitalent.kominfo

DTS_kominfo

digitalent.kominfo

digital talent scholarship