



Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi
Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi
Republik Indonesia

DIKTI
SIGAP
MELAYANI

Kampus
Merdeka
INDONESIA JAYA



MICROCREDENTIAL: ASSOCIATE DATA SCIENTIST

01 November – 10 Desember 2021

- Pertemuan ke-3

Metodologi Pengembangan AI Menggunakan Data



[ditjen.dikti](#)



@ditjendikti



[ditjen.dikti](#)



Ditjen Diktiristek



<https://dikti.kemdikbud.go.id/>



Profil Pengajar: I Putu Agus Eka Pratama, S.T., M.T.



- **Jabatan Akademik:** Tenaga Pengajar

- **Latar Belakang Pendidikan:**

- S1: Informatika, Institut Teknologi Telkom (Telkom University)
- S2: Informatika, Institut Teknologi Bandung
- S3: -

- **Riwayat/Pengalaman Pekerjaan:**

- Dosen
- Peneliti
- Konsultan IT
- Penulis buku IT

Contak Pengajar:

Ponsel: 085738336989

WA/Telegram: 085738336989

Email:eka.pratama@unud.ac.id

<https://udayananetworking.unud.ac.id/lecturer/885-i-putu-agus-eka-pratama>



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
- Development Life Cycle (materi tambahan tim instruktur UG)
- Arsitektur Sistem (materi tambahan tim instruktur UG)
- Project Management (materi tambahan tim instruktur UG)
- Beragam model life cycle management (materi tambahan tim instruktur UG)
- Proses Bisnis (materi tambahan tim instruktur UG)



Agenda

- **Mengapa Metodologi diperlukan**
 - Mengapa Majoritas Projek AI Gagal
- **Development Life Cycle, Arsitektur Sistem, Manajemen Proyek** (materi pelengkap/ tambahan tim instruktur UG)
 - Problem Pengembangan
 - Analisis Stakeholder
 - Arsitektur Sistem dan Proses Bisnis
 - Tahapan Mengelola Proyek
- **Berbagai Metodologi Data Science**
 - Tak semua metodologi sama lengkap
- **Langkah Pengembangan**
 - Dari Masalah Bisnis menjadi Aplikasi AI



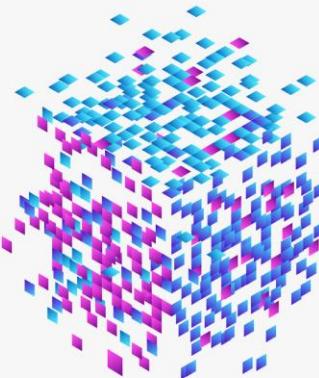
Mengapa Metodologi Diperlukan



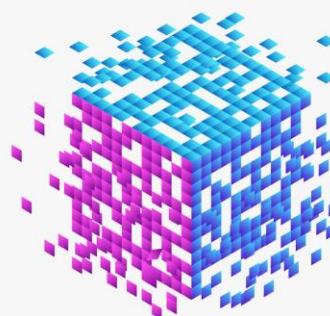


Sistem AI berbasis (Big) Data

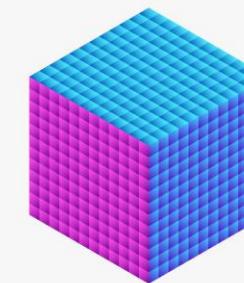
BIG DATA



ANALYTICS



DECISIONS



stargazr

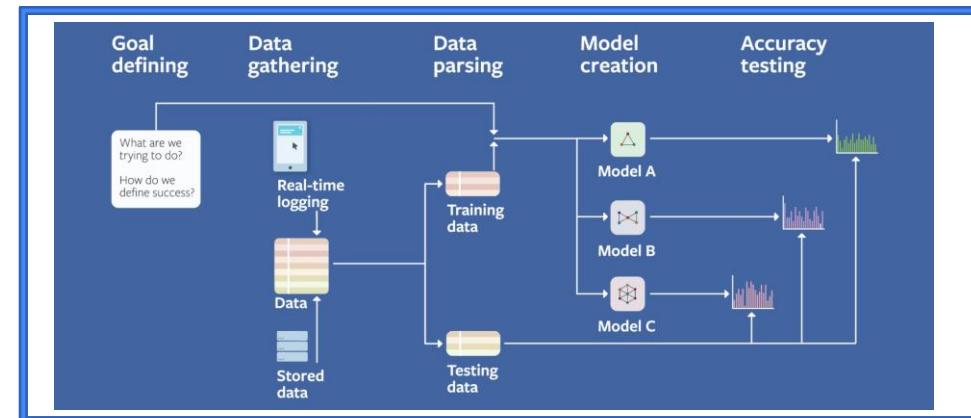
Data

Menjadi

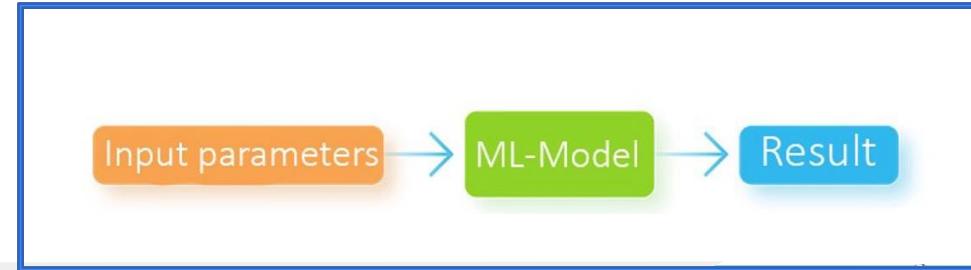
Sistem Intelijen
(berbasis Pengetahuan)

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 menggunakan data dan data science melalui urutan langkah yang ditentukan

Dari “Craft ke Engineering”



- Kutak katik
- No Method
- No Design
- No Documentation



- Terarah
- Method tertentu
- Design before implementing
- Well Documentation

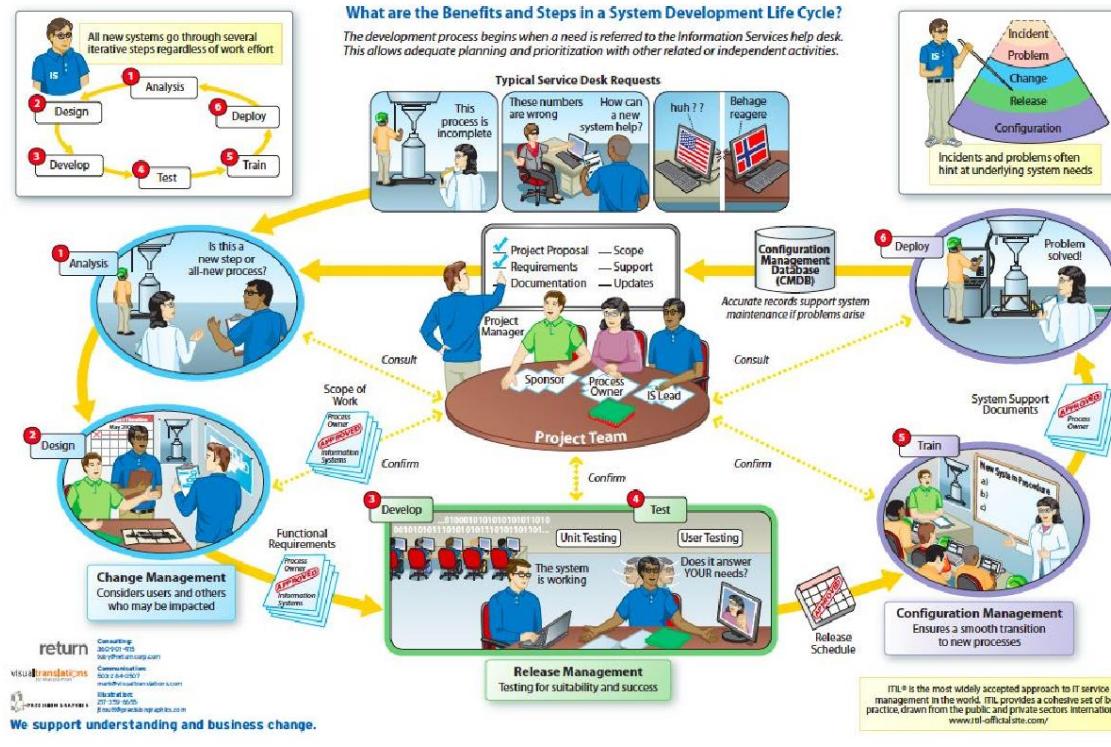


Development Life Cycle

(materi pelengkap/tambahan tim instruktur UG)

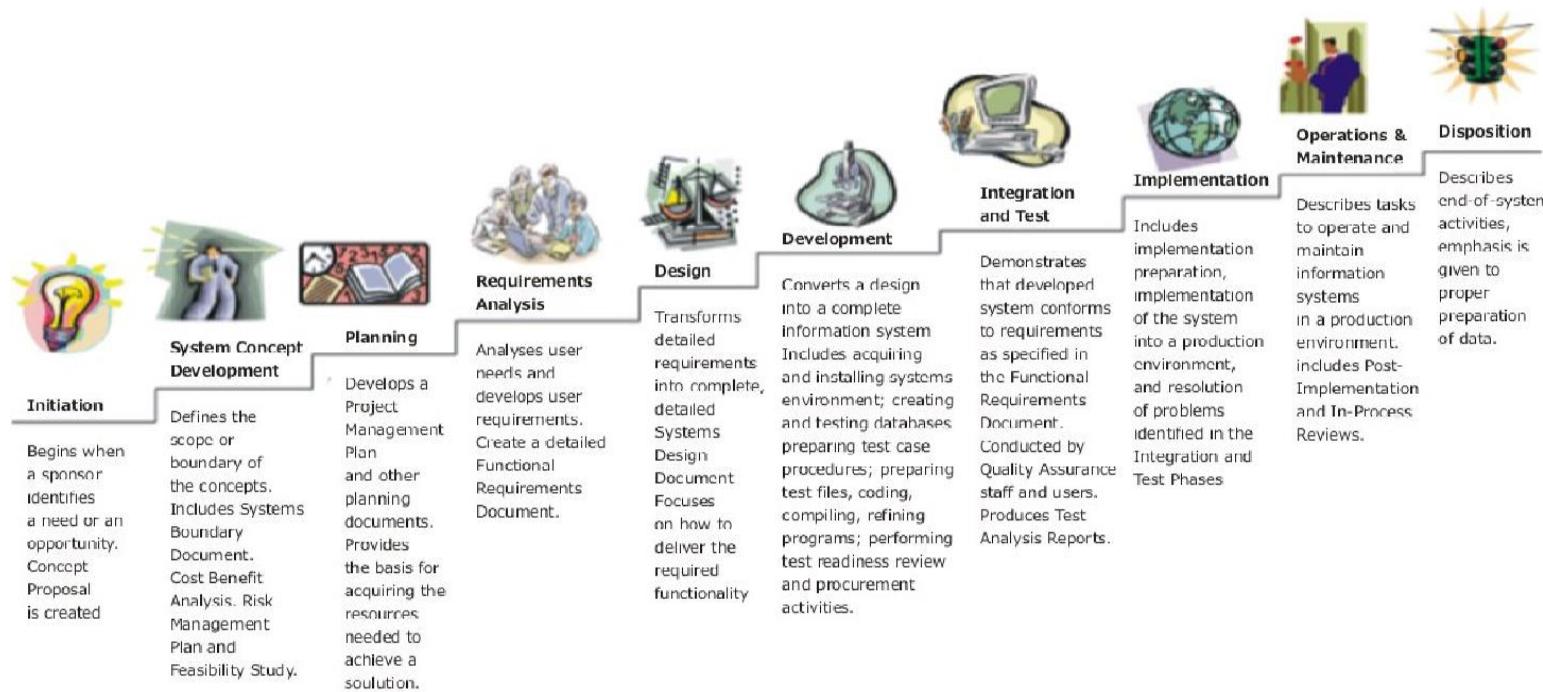


Gambaran Besar/ Big Picture Manajemen Proyek IT



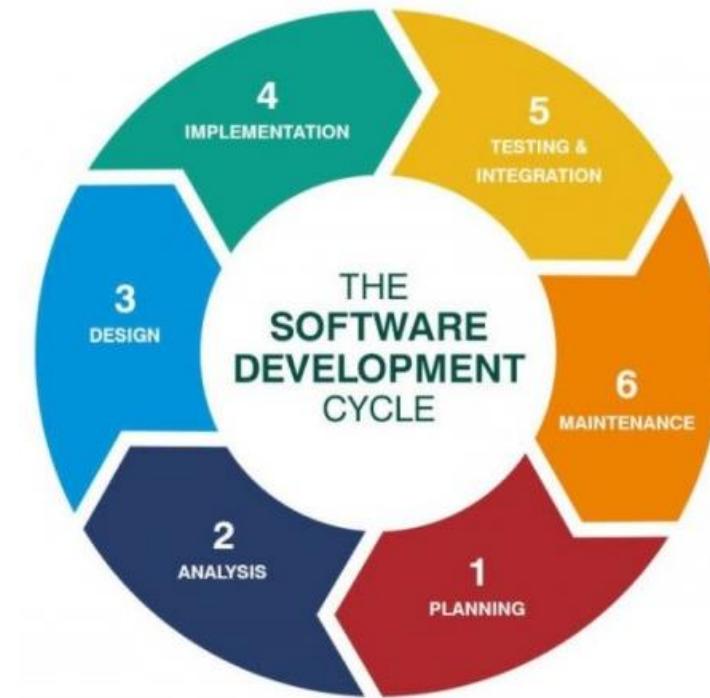


System Development Life Cycle (SDLC)



SDLC

- **Metode SDLC** (Software Development Life Cycle) adalah proses pembuatan dan pengubahan sistem serta model dan metodologi yang digunakan untuk mengembangkan sistem rekayasa perangkat lunak
- Proses logika yang digunakan oleh seorang analis sistem untuk mengembangkan sebuah sistem informasi yang melibatkan requirements, validation, training dan pemilik sistem (Prof. Dr. Sri Mulyani, AK., CA. 2017)
- proses yang memproduksi sebuah software dengan kualitas setinggi-tingginya tetapi dengan biaya yang serendah-rendahnya (Stackify)





Jenis Metode SDLC

- **Waterfall (Air Terjun)**

Metode kerja yang menekankan fase-fase yang berurutan dan sistematis. Disebut waterfall karena proses mengalir satu arah “ke bawah” seperti air terjun. Metode waterfall ini harus dilakukan secara berurutan sesuai dengan tahap yang ada.

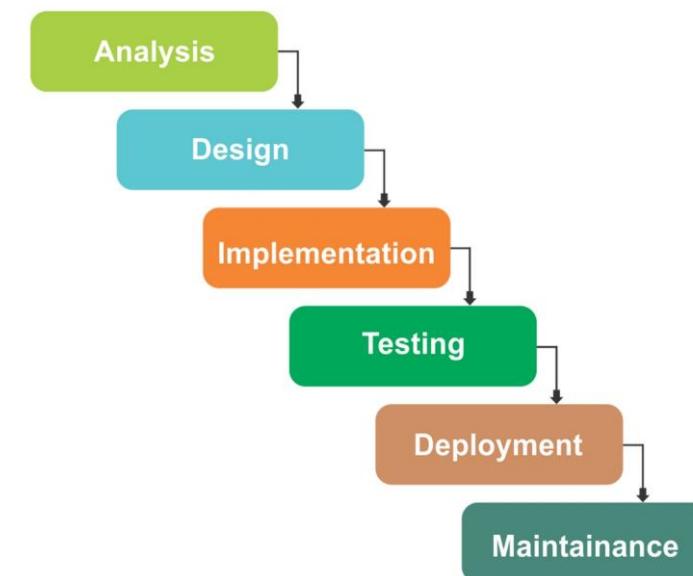
Pro:

- Paling handal dan paling lama digunakan.
- Cocok untuk sistem software dengan kompleksitas rendah
- Penggeraan project sistem terjadwal dengan baik dan mudah dikontrol (keteraturan dan jadwal rapih)

Kontra:

- Waktu pengembangan lama, harus menunggu tahap sebelumnya selesai. shg
- Biaya juga mahal,
- Kaku, tahapan pada waterfall tidak dapat berulang, maka model ini tidak cocok untuk proyek dengan kompleksitas tinggi

SDLC- WATERFALL MODEL



Jenis Metode SDLC

- **Prototype** (Purwarupa)

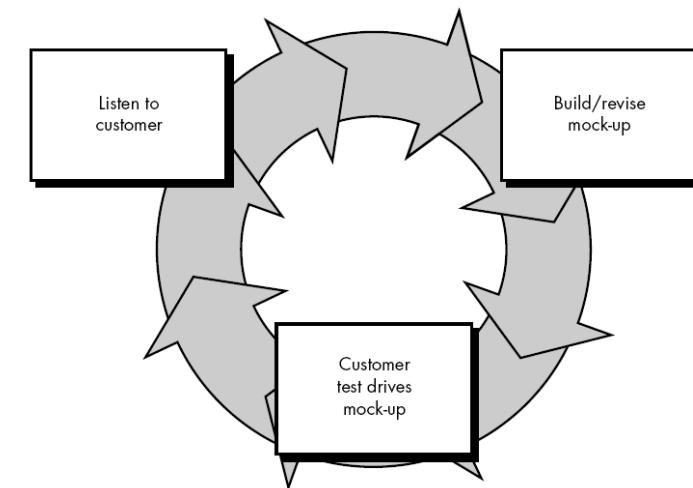
metode yang memungkinkan pengguna atau user memiliki gambaran awal tentang perangkat lunak yang akan dikembangkan, serta pengguna dapat melakukan pengujian di awal sebelum perangkat lunak dirilis. bertujuan: mengembangkan model menjadi perangkat lunak yang final. Artinya sistem akan dikembangkan lebih cepat dan biaya yang dikeluarkan lebih rendah

Pro:

- Mempersingkat waktu pengembangan perangkat lunak
- Penerapan fitur menjadi lebih mudah, karena pengembang mengetahui apa yang diharapkan

Kontra:

- Proses yang dilakukan untuk analisis dan perancangan terlalu singkat
- Kurang fleksibel jika terjadi perubahan



Jenis Metode SDLC

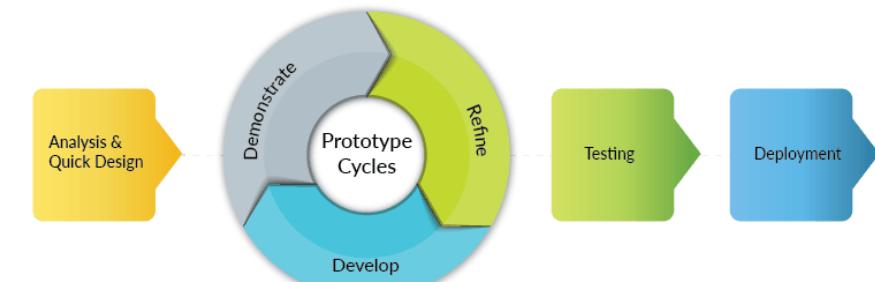
- **RAD** (Rapid Application Development) menggunakan pendekatan iteratif dan inkremental, dan menekankan pada tenggat waktu dan efisiensi biaya yang sesuai dengan kebutuhan

Pro:

- dianggap lebih singkat. semua pihak, baik pelanggan maupun pengembang, terus terlibat secara aktif dalam setiap proses hingga hasil dapat tercapai.
- tahapan kerja pada lebih sedikit.

Kontra:

- segi konsistensi dan kemampuan personel butuh usaha lebih
- kurang cocok utk proyek skala besar



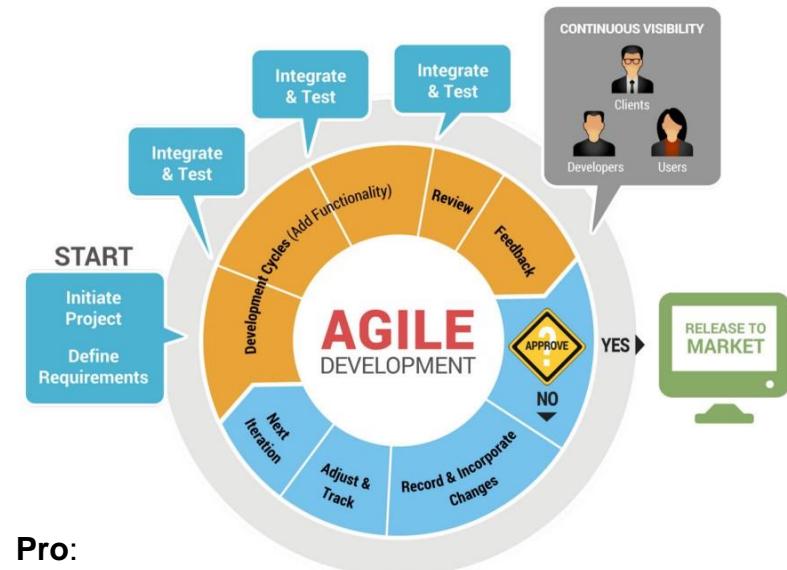
Jenis Metode SDLC

- **Agile**

model pengembangan jangka pendek yang memerlukan adaptasi cepat dan pengembangan terhadap perubahan dalam bentuk apapun induk dari model *Scrum*

poin utama:

- Interaksi antar personal lebih penting daripada proses dan alat.
- Software yang berfungsi lebih penting daripada dokumentasi yang lengkap
- Kolaborasi dengan klien lebih penting daripada negoisasi kontrak.
- Sikap tanggap lebih penting daripada mengikuti rencana/plan.
- Dokumentasi harus tersusun rapi dan terstruktur



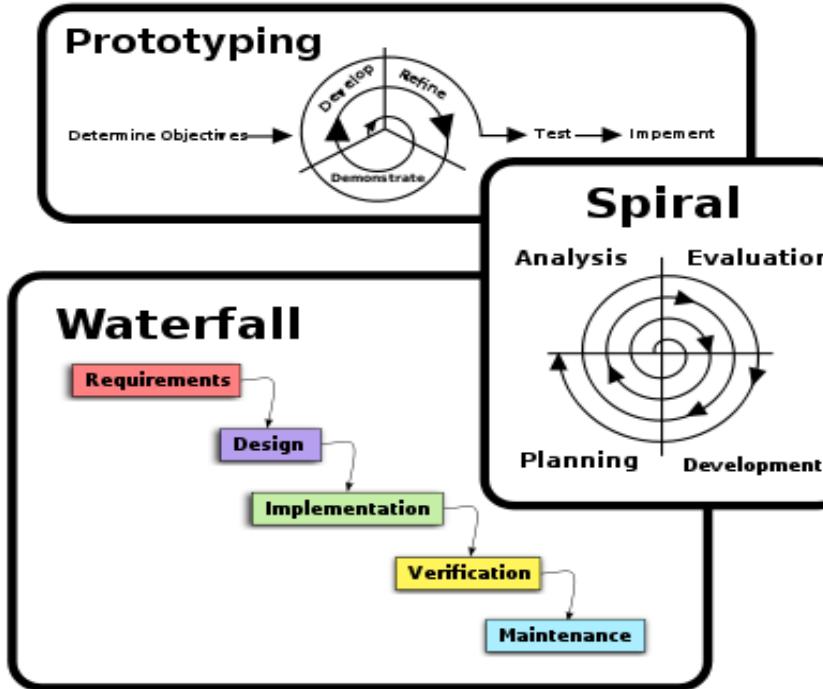
Pro:

- Functional dapat dibuat dengan cepat dan dilakukan testing
- Perubahan dengan cepat ditangani

Kontra:

- Analisis, desain, dan pengembangan sulit diprediksi
- Dapat memunculkan permasalahan dari arsitektur maupun desain.

Berbagai Metodologi SDLC



- Setiap metodologi cocok untuk permasalahan dan constraint tertentu
- Setiap metodologi membutuhkan personal (perencanaan SDM) dan tools yang berbeda
- Setiap metodologi membutuhkan penjadwalan (perencanaan waktu) yang berbeda



Proses Bisnis (probis)

(materi pelengkap/tambahan tim instruktur UG)





Proses Bisnis

- Start Transformasi Teknologi *Transformasi Proses Bisnis*
- *Proses Bisnis*: Nadi dari Solusi Bisnis dan Teknologi
- Proses bisnis dapat dipahami stakeholder internal organisasi dan eksternal (pelanggan, investor, regulator, dll)
- Salah satu solusi penggambaran proses bisnis: menggunakan model/notasi standar
- BPMN: *Business Process Modeling Notation*



Proses Bisnis dengan BPMN

- Standar untuk pemodelan proses bisnis yang menyediakan notasi grafis untuk menentukan proses bisnis dalam Business Process Diagram (BPD)
- Berbasis teknik flowchart dan similar dengan diagram aktivitas di UML
- Sebagai bahasa standar komunikasi antara desain vs implementasi
- Tujuan: mendukung manajemen proses bisnis, baik untuk pengguna teknis dan pengguna bisnis, dengan memberikan notasi yang intuitif untuk pengguna bisnis, namun mampu mewakili semantik proses yang kompleks.
- Manfaat:
 - memberikan bahasa yang sama /notasi standar yang mudah dipahami oleh semua pemangku kepentingan bisnis:
 - internal: analis bisnis, pengembang teknis, dan manajer bisnis.
 - eksternal: pelanggan/pengguna lain, investor, mitra (ABG)
 - Untuk memvisualisasikan proses bisnis
 - Untuk mendokumentasikan sebuah proses
 - Untuk melakukan analisis pada proses bisnis

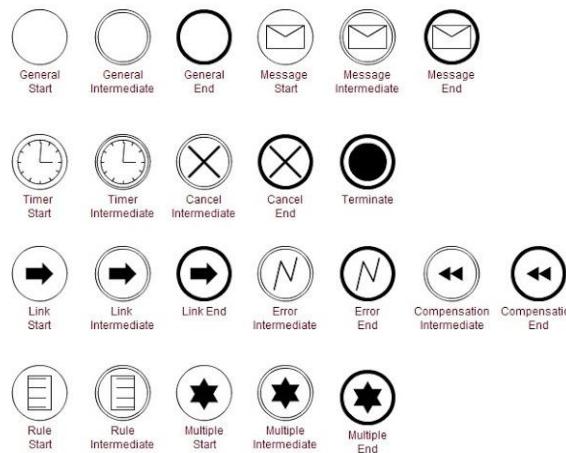
Proses Bisnis dengan BPMN

- **Pro:**
 - Less dependent by vendor. Tidak dimiliki satu atau sebagai perusahaan, tapi oleh Institusi [OMG](#) yang mapan dengan standar dunia, seperti UML
 - Mudah digunakan dan memahami notasi sangat cepat
 - Power of expression: Jika perlu, Anda dapat menjelaskan dengan tepat bagaimana suatu proses berfungsi dengan BPMN.
 - BPMN terutama dikembangkan untuk mendukung implementasi teknis proses (“Otomasi Proses”). Semakin penting TI dalam suatu perusahaan, semakin membantu penggunaan BPMN.
- **Tools:**
 - [camunda](#)
 - [draw io](#)
 - [lucidchart](#)
 - [bpmn io](#)
 - dll



Proses Bisnis dengan BPMN

Notasi



Kategori Notasi

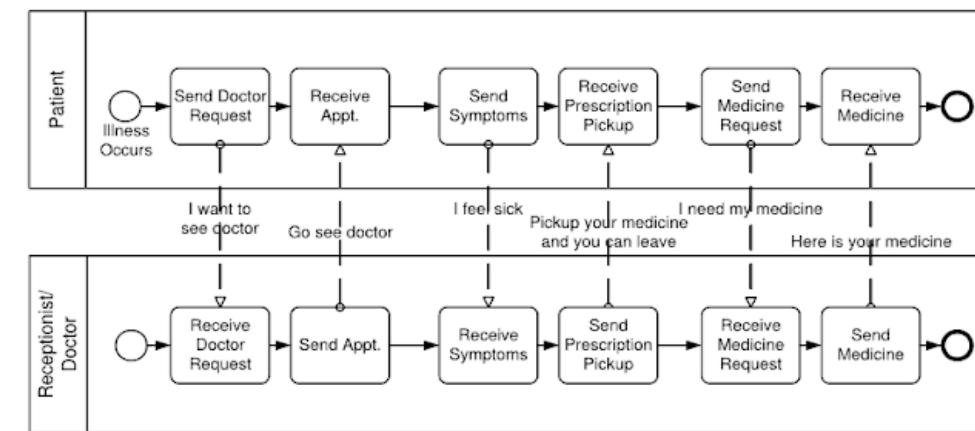


Proses Bisnis dengan BPMN

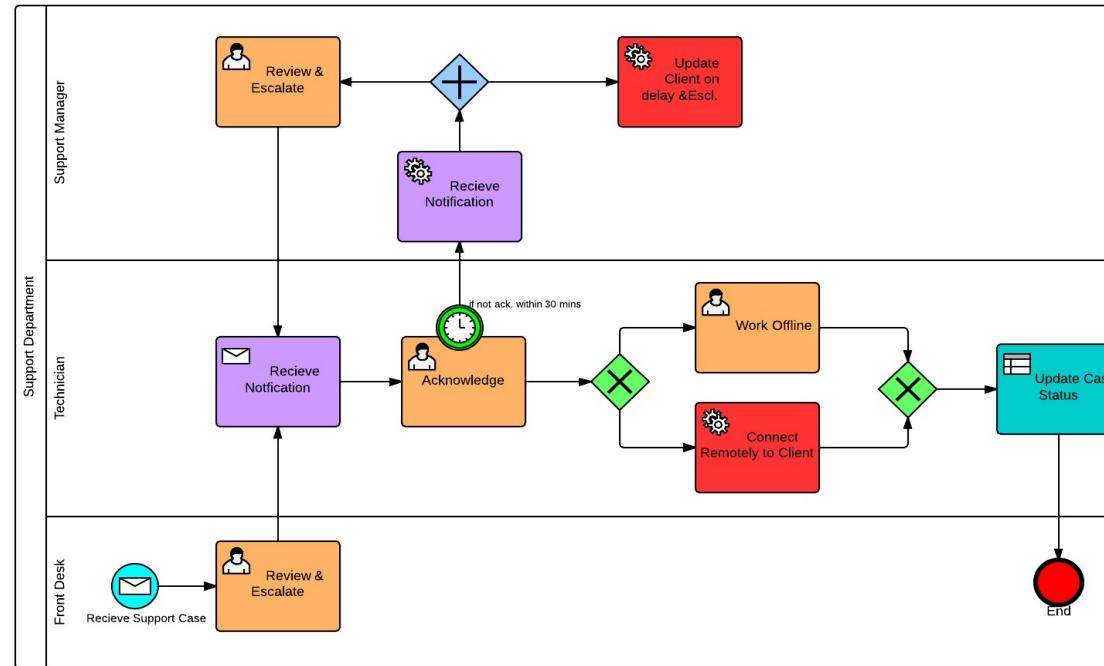
Contoh



Contoh



Proses Bisnis dengan BPMN



Contoh





Bussines Goal vs User needs

Assumptions Worksheet

Business Assumptions

1. I believe my customers have a need to _____.
2. These needs can be solved with _____.
3. My initial customers are (or will be) _____.
4. The #1 value a customer wants to get out of my service is _____.
5. The customer can also get these additional benefits _____.
6. I will acquire the majority of my customers through _____.
7. I will make money by _____.
8. My primary competition in the market will be _____.
9. We will beat them due to _____.
10. My biggest product risk is _____.
11. We will solve this through _____.
12. What other assumptions do we have that, if proven false, will cause our business/project to fail? _____.

User Assumptions

1. Who is the user?
2. Where does our product fit in his work or life?
3. What problems does our product solve?
4. When and how is our product used?
5. What features are important?
6. How should our product look and behave?



Berbagai Metodologi Data Science





Jenis Metodologi

- Metodologi kegiatan Teknis
- Metodologi kegiatan bisnis (dan teknis)



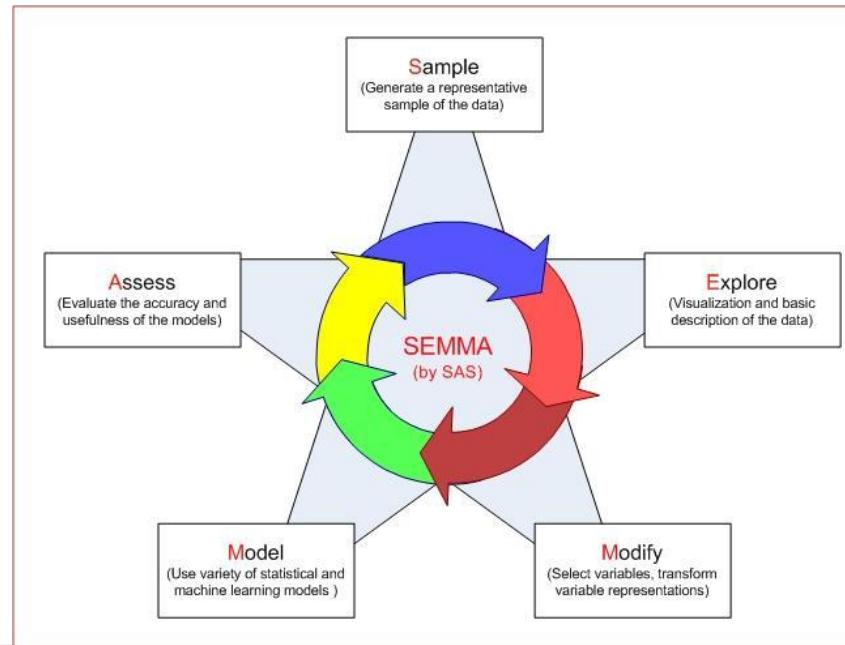
Mengapa harus ada standar proses

- Proses data mining harus handal dan dapat diulang oleh orang dengan latar belakang data mining yang sedikit.
- Framework untuk merekam pengalaman → memungkinkan proyek diulangi
- Alat bantu untuk perencanaan proyek dan manajemen
- Bagi pengembang baru akan memudahkan
- Menunjukkan maturitas pekerjaan data mining
- Meminimalkan kebergantungan pada personal utama



Metodologi Teknis: Kegiatan DS/AI dianggap Kegiatan Teknikal

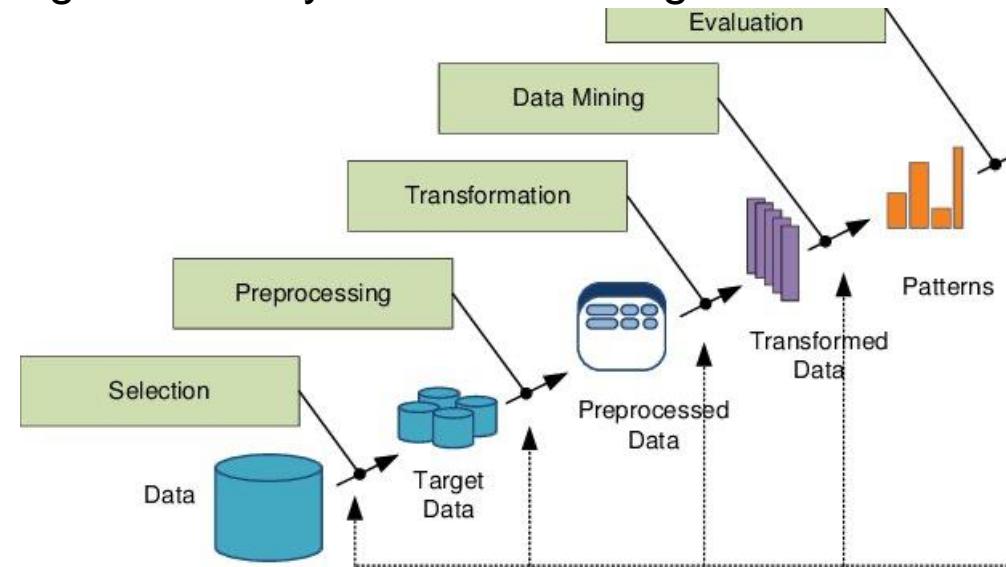
- SEMMA dari SAS Institute



<https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnij8bbjjm1a2.htm&docsetVersion=14.3&locale=en>

Metodologi Teknis: Kegiatan DS/AI dianggap Kegiatan Teknikal

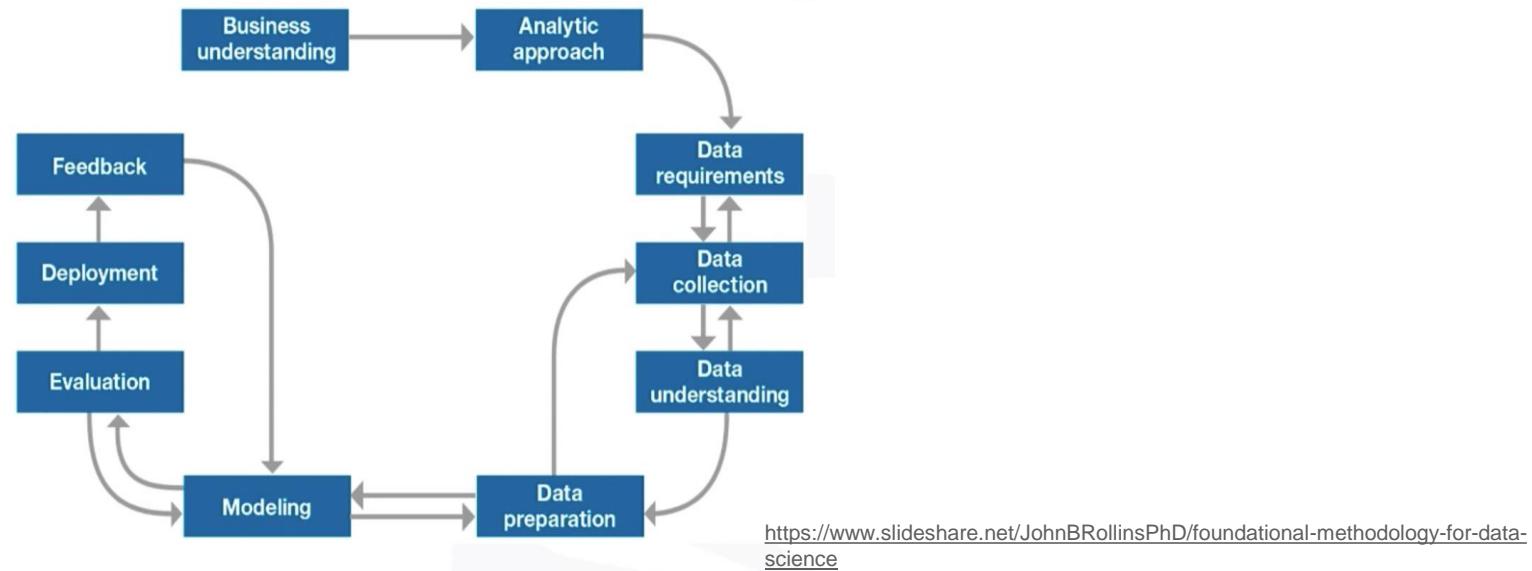
- Knowledge Discovery and Data Mining



<https://www.kdnuggets.com/gpsspubs/aimag-kdd-overview-1996-Fayyad.pdf>

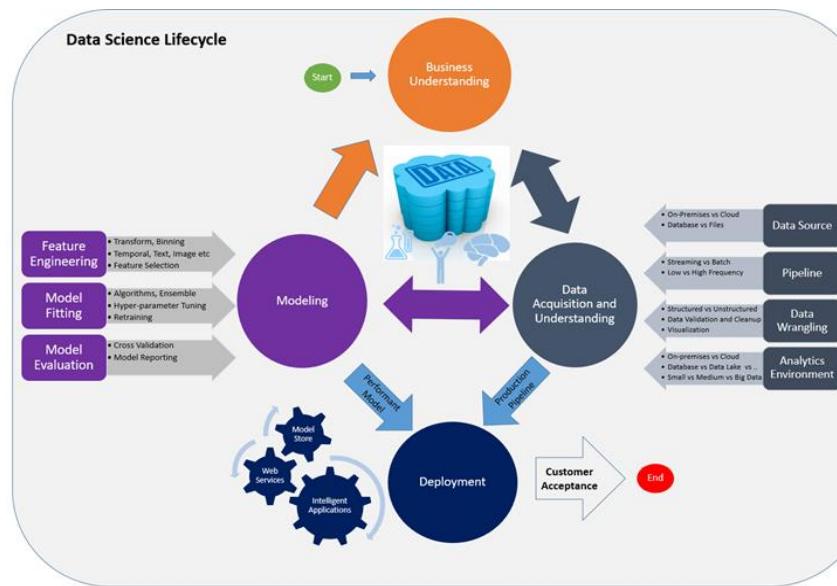
Metodologi Lengkap: Kegiatan DS/AI dianggap Kegiatan Bisnis: Masalah Bisnis menjadi Masalah DS/AI

- IBM Data Science Methodology



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

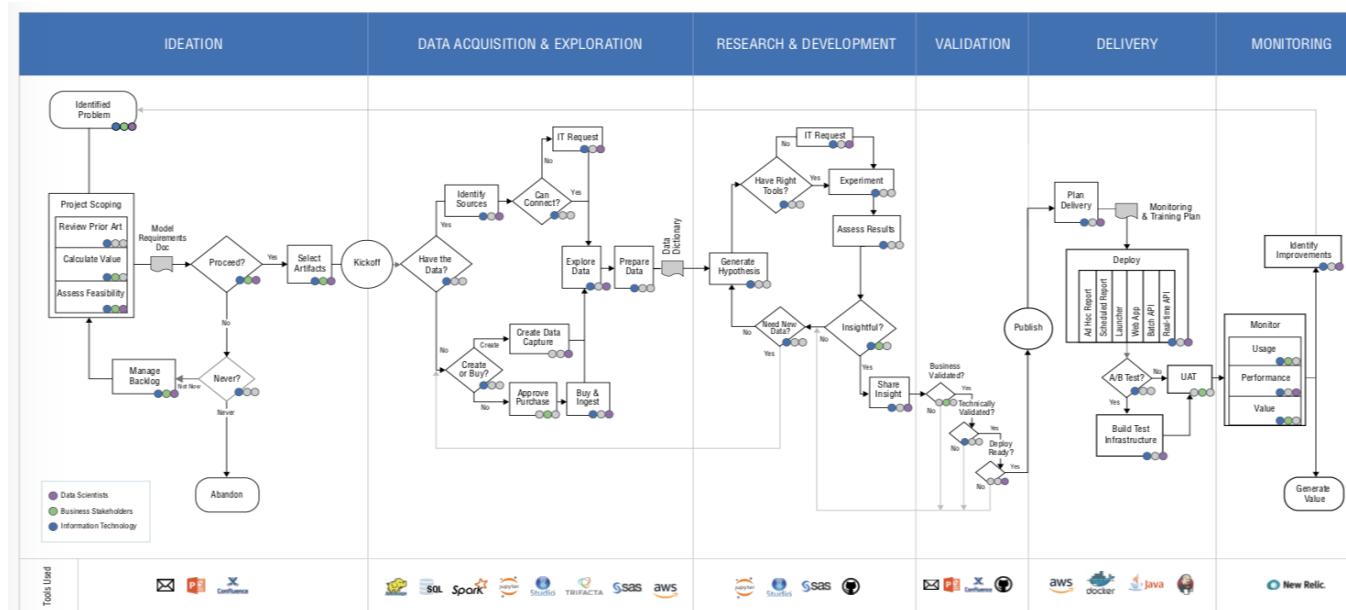
- Microsoft's Team Data Science Process



<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

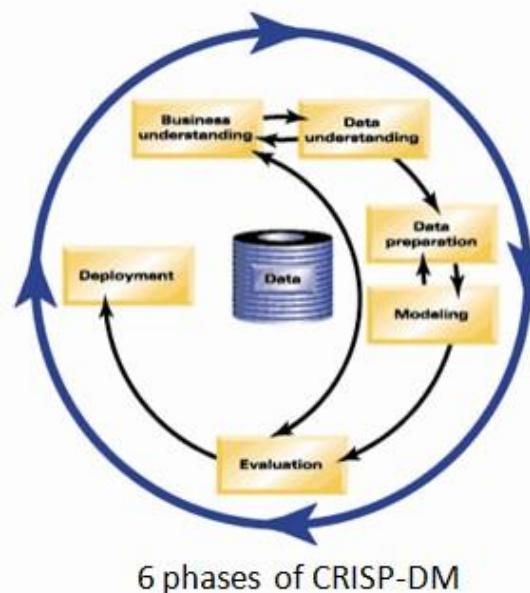
- Domino DataLab Methodology



<https://www.dominodatalab.com>

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

- **CRISP-DM: Cross Industry Standard Process for Data Mining**



<https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jn8bbijm1a2.htm&docsetVersion=14.3&locale=en>



Bagaimana di Indonesia?

Standard Kompetensi Kerja Nasional:
KepMen Ketenagakerjaan No 299 thn 2020



MENTERI KETENAGAKERJAAN
REPUBLIK INDONESIA

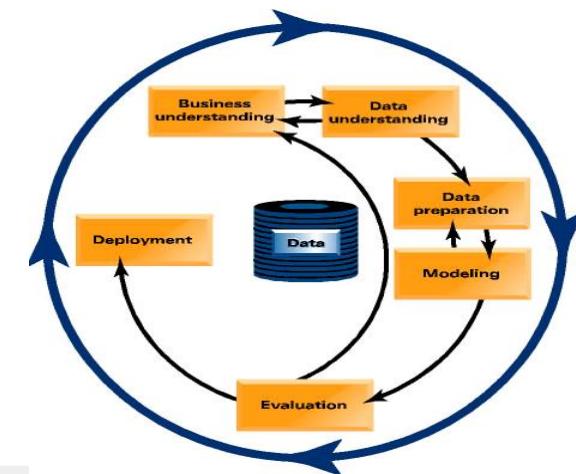
KEPUTUSAN MENTERI KETENAGAKERJAAN
REPUBLIK INDONESIA
NOMOR 299 TAHUN 2020
TENTANG

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

CRISP - DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
- As well as technical analysis
- Framework for guidance
- Experience base
- Templates for Analysis
- Data Mining methodology
- Process Model
- For anyone
- Provides a complete blueprint
- Life cycle: 6 phases



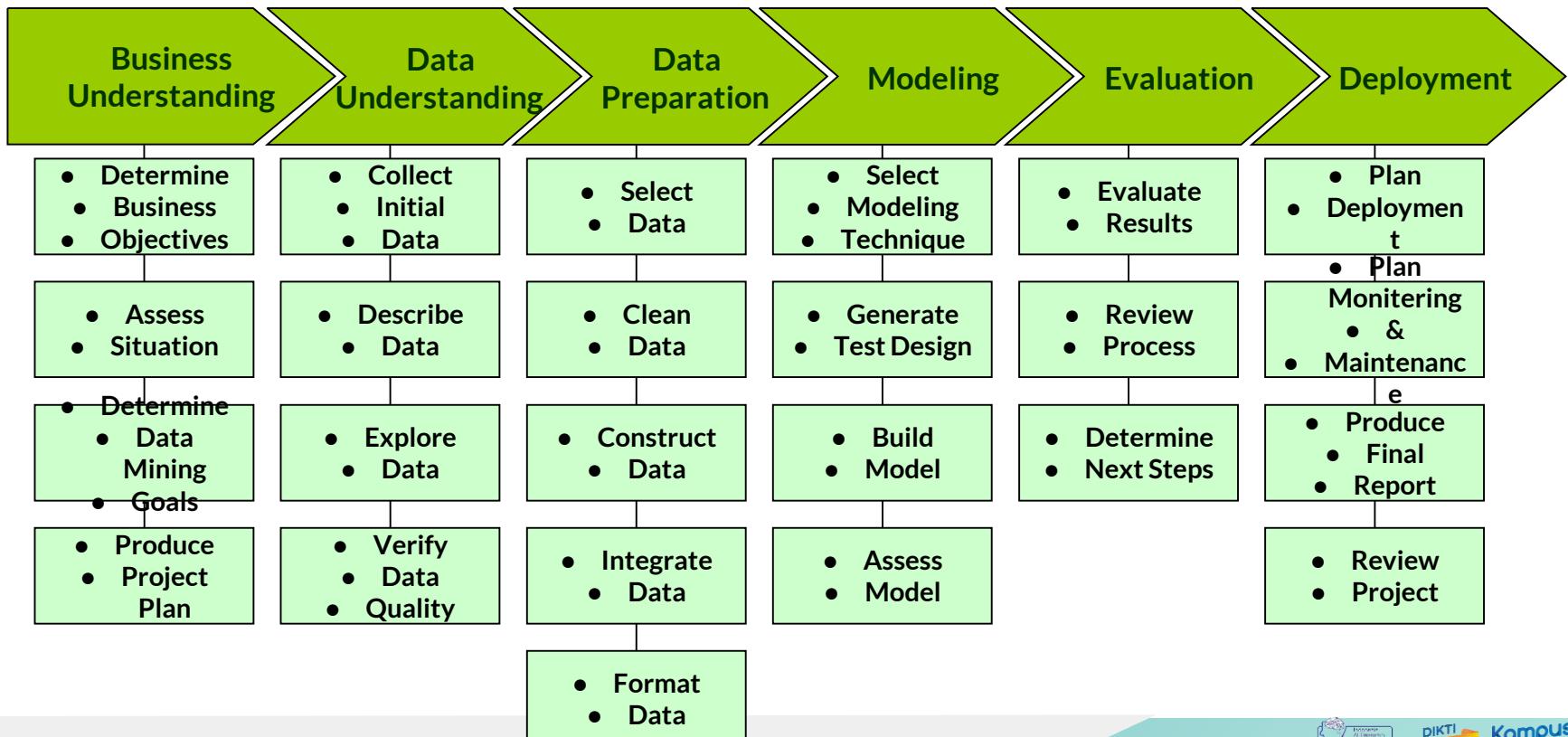


Standardisasi Proses Data Mining

- **Initiative launched in late 1996 by three “veterans” of data mining market.**
- Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR
- **Developed and refined through series of workshops (from 1997-1999)**
- **Over 300 organization contributed to the process model**
- Published CRISP-DM 1.0 (1999)
- **Over 200 members of the CRISP-DM SIG worldwide**
 - **DM Vendors** - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, etc.
 - **System Suppliers / consultants** - Cap Gemini, ICL Retail, Deloitte & Touche, etc.
 - **End Users** - BT, ABB, Lloyds Bank, AirTouch, Experian, etc.



Fase dan Pekerjaan





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



Problem:

Bagaimana menurunkan NPL suatu bank

Pertanyaan:

Bagaimana memperbaiki perhitungan Credit score

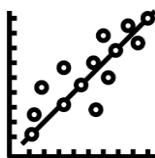
Measurable outcomes:

% Penurunan kredit gagal bayar



1. Business Understanding: Menentukan Tugas Analytics

Apa Tugas Analitiks yang perlu diselesaikan
untuk menjawab permasalahan bisnis?



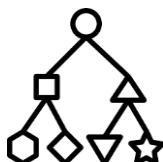
A. Regresi/Estimasi: Memprediksi nilai kontinyu dari kasus

- 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?



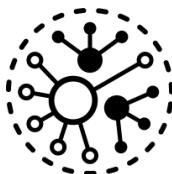
B. Klasifikasi: Memprediksi kelas/ kategori dari kasus

- Prediksi kolektibilitas suatu pinjaman
- Prediksi kebangkrutan suatu perusahaan 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
- Pengelompokan 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

- Pendekripsi transaksi illegal penggunaan kartu kredit
- Pendekripsi 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?



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

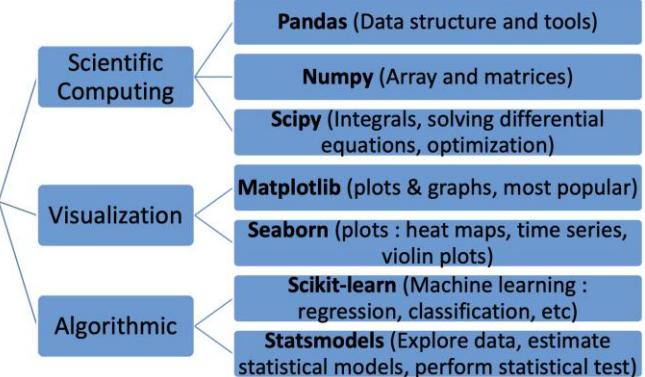
03

Memvalidasi Data

Menilai kesesuaian kualitas data dengan masalah yang akan dipecahkan

Kualitas Data

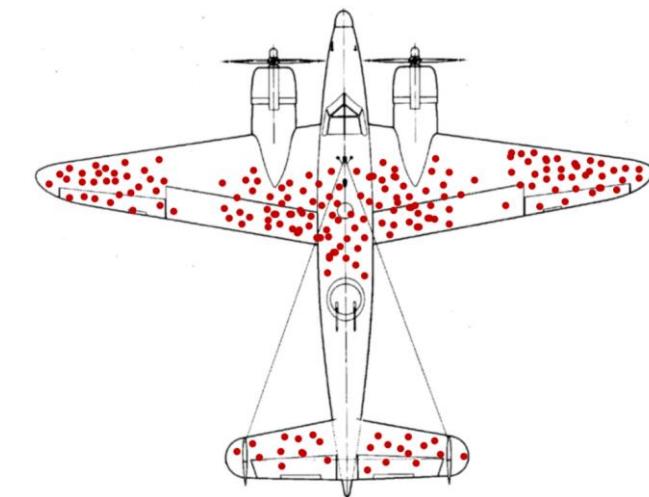
Python Libraries



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

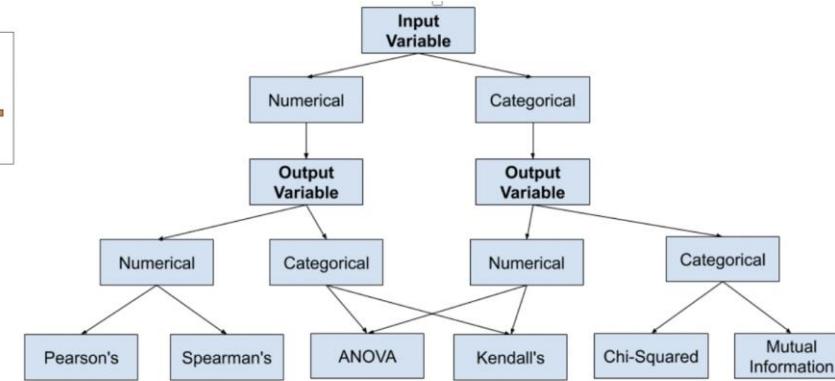
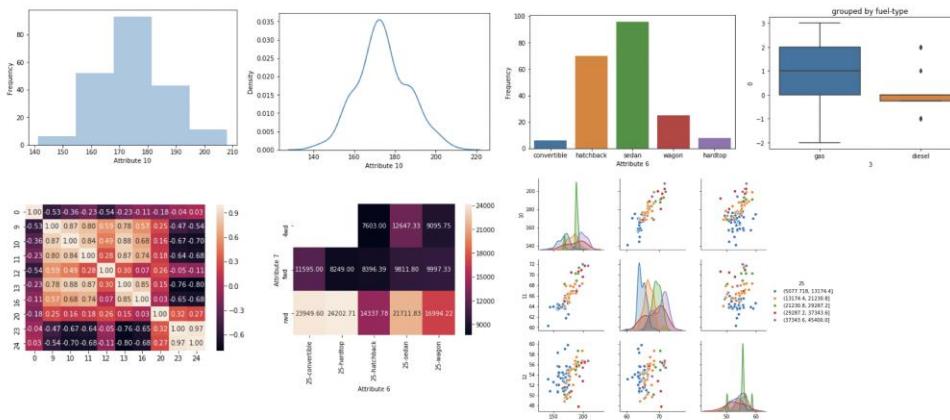
Deskripsi Data: Penjelasan arti atribut/ fitur

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,...



Copyright © MachineLearningMastery.com



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

01

Memilih dan memilah data

Memilih data yang akan dipergunakan

02

Membersihan Data

Meminimalkan noise (tidak lengkap, salah)

03

Mengkonstruksi data

Menambahkan fitur dan transformasi data

04

Integrasi Data

Menggabungkan data

Rekord terpakai
Atribut terpakai

Data lengkap
Data yang diperbaiki
Data Pecilan

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

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

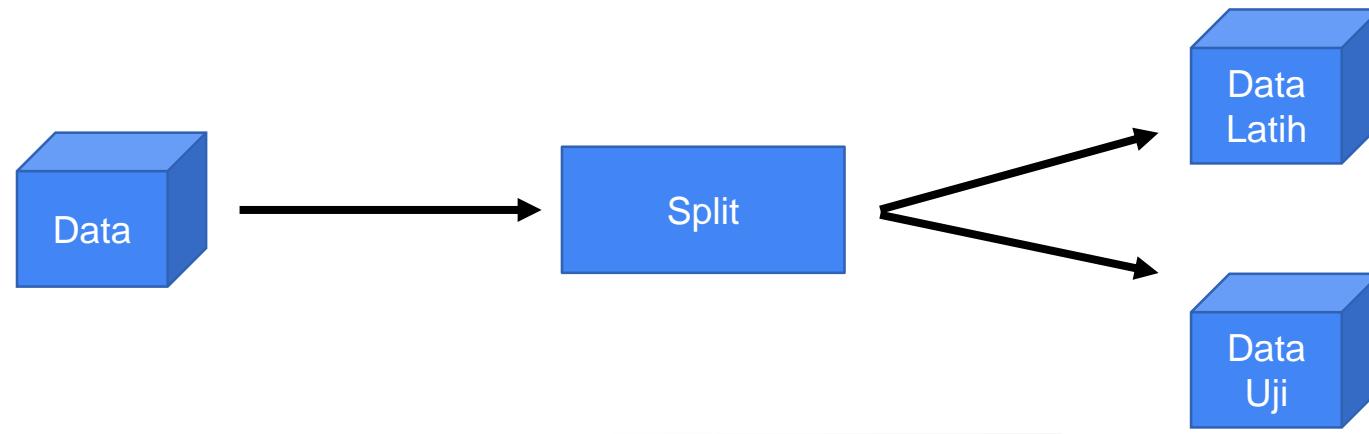
1. k-Nearest Neighbor (k-NN)
2. Naïve Bayes
3. Regression Techniques
4. Support Vector Machines (SVMs)
5. Decision Trees
6. Random Forests
7. Deep Learning Algorithms
8. . . .

4. Modeling : Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

B. **Membagi data:** Sesuai dengan ketersediaan data

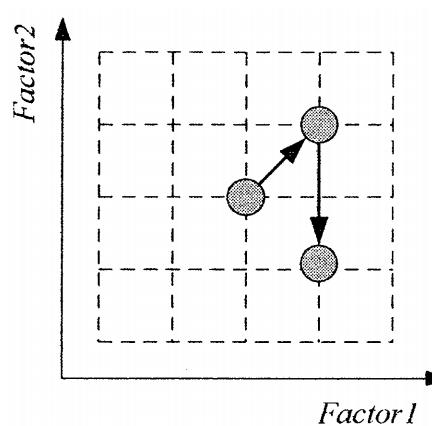
1. Data Latih: Untuk mengembangkan model
2. Data Uji: Untuk Mengukur performansi model



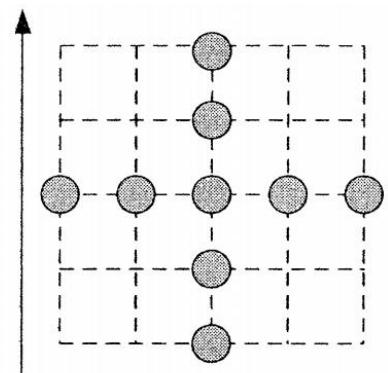
4. Modeling : Membangun Skenario Pemodelan

Membuat strategi pencarian model terbaik

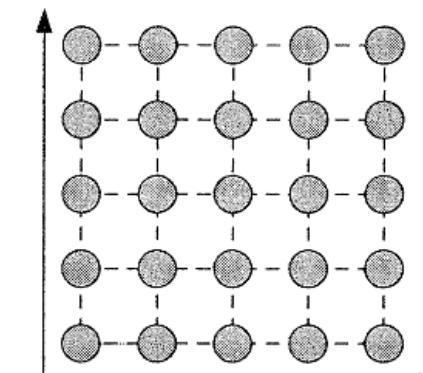
C. Menentukan Langkah Eksperimen: Untuk mendapatkan model terbaik secara efisien dan efektif



Best Guess



One Factor at A Time



Grid Search



4. Modeling : Membangun model

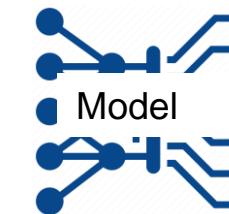
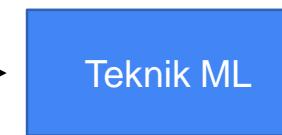
Mengembangkan model dengan Teknik ML

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

4. Modeling : Membangun model

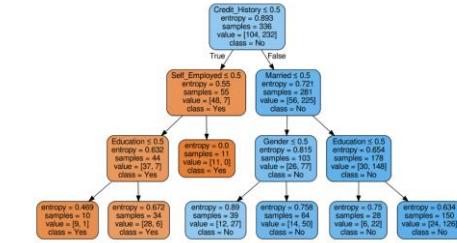
Mengembangkan model dengan Teknik ML

A. Proses Pelatihan : Untuk mendapatkan model



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. ...

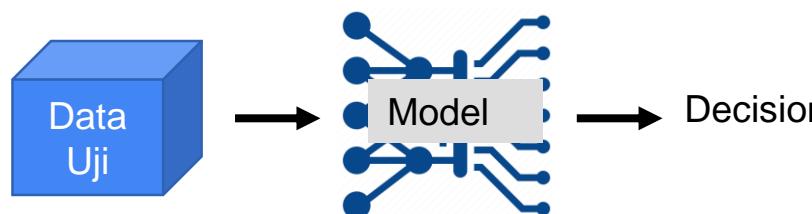
| Variable | Type | Definition |
|----------|------|---|
| BAD | Num | BAD: 1 = applicant defaulted on loan or seriously delinquent; 0 = applicant paid loan |
| LOAN | Num | LOAN: Number of loans in loan request |
| MORTDUE | Num | MORTDUE: Amount due on existing mortgage |
| VALUE | Num | VALUE: Value 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 |



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 |

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

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

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



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 Proyek, dan Pakar TI dalam satu Tim



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/>
- Togaf
 - <https://pubs.opengroup.org/architecture/togaf9-doc/arch/>
- SDLC:
 - <https://glints.com/id/lowongan/software-development-life-cycle/#.YQ0hM4gzZPZ>
 - <https://www.dicoding.com/blog/metode-sdlc/>
 - http://www.id.w3ki.com/sdlc/sdlc_rad_model.html
 - <https://medium.com/@purwanto.dev/metodologi-system-development-life-cycle-sdlc-2f0349df1364>
- Proses Bisnis:
 - <https://camunda.com/bpmn/>
 - <https://www.bpmn.org/>
 - <http://ccg.co.id/blog/2017/04/28/pemodelan-proses-bisnis-dengan-bpmn/>



Quiz / Tugas

- Quiz dapat diakses melalui <https://spadadikti.id/>



Terima kasih