Modul Pelatihan

Metodologi Data Science

Thematic Academy

Digital Talent Scholarship

Tahun 2021

Tujuan Pembelajaran

A. Tujuan Umum

Setelah mempelajari modul ini peserta latih diharapkan mampu menjelaskan metodologi pembuatan aplikasi intelijen menggunakan data science dan menerapkannya untuk suatu permasalahan data science.

B. Tujuan Khusus

Adapun tujuan mempelajari unit kompetensi melalui modul Metodologi Data science adalah untuk memfasilitasi peserta latih sehingga pada akhir pelatihan diharapkan memiliki kemampuan sebagai berikut:

1. Mampu menjelaskan berbagai kegagalan projek pengembangan sistem intelijensi menggunakan data
2. Mampu menjelaskan langkah-langkah yang diperlukan untuk menyelesaikan masalah dengan menggunakan data science.

Latar Belakang

Unit kompetensi ini dinilai berdasarkan tingkat kemampuan peserta dalam memahami metodologi data science. Adapun penilaian dilakukan dengan menggabungkan serangkaian metode untuk menilai kemampuan dan penerapan pengetahuan pendukung penting. Penilaian dilakukan dengan mengacu kepada Kriteria Unjuk Kerja (KUK) dan dilaksanakn di Tempat Uji Kompetensi (TUK), ruang simulasi atau workshop dengan cara:

1.1. Lisan
1.2. Wawancara
1.3. Tes tertulis
1.4. Metode lain yang relevan

Deskripsi Pelatihan

Tujuan utama dari modul pelatihan ini adalah untuk membantu para peserta memahami kebutuhan akan metodologi dan Langkah-langkah metodologi dasar dalam mengembangkan aplikasi intelijen menggunakan data science.

Kompetensi Dasar

A. Mampu menentukan objektif bisnis dari suatu kegiatan Data science
B. Mampu mengidentifikasi Langkah-langkah dalam pengembangan kegiatan Data science

Indikator Hasil Belajar

Peserta mampu memahami dan menggunakan langkah-langkah dasar dalam membuat aplikasi intelijen menggunakan pendekatan data science

INFORMASI PELATIHAN

Akademi	Thematic Academy
Mitra Pelatihan	Kementerian Komunikasi dan Informatika
Tema Pelatihan	**Data Scientist: Artificial Intelligence untuk Dosen dan Instruktur**
Sertifikasi	• *Certificate of Attainment;* • Sertifikat Kompetensi Associate Data Scientist
Persyaratan Sarana Peserta/spesifikasi device Tools/media ajar yang akan digunakan	Memiliki laptop/komputer dengan spesifikasi minimal : • RAM minimal 2 GB (disarankan 4 GB) • Laptop dengan 32/64-bit processor • Laptop dengan Operating System Windows 7, 8, 10, MacOS X atau Linux • Laptop dengan konektivitas WiFi dan memiliki Webcam • Akses Internet Dedicated 126 kbps per peserta per perangkat • Memiliki aplikasi Zoom • Memiliki akun Google Colab
Aplikasi yang akan digunakan selama pelatihan	• Google Colab • Jupyter notebook
Tim Penyusun	Windy Gambetta, Ir., MBA (ITB)

INFORMASI PEMBELAJARAN

Unit Kompetensi	Materi pembelajaran	Kegiatan pembelajaran	Durasi Pelatihan	Rasio Praktek : Teori	Sumber pembelajaran
-	Metodologi data science	Daring/Online	Live Class 2 JP LMS 4 JP @ 45 menit	70:30	LMS

Materi Pokok

Metodologi Data science

Sub Materi Pokok

- Mengapa Metodologi *Data science* diperlukan
- Berbagai Metodologi *Data science*
- Langkah Generik Pengembangan Aplikasi dengan *Data science*

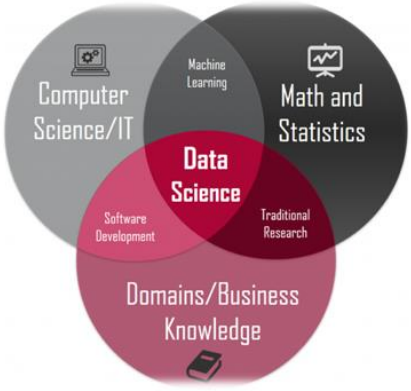
A. Materi Pelatihan

1. Mengapa Metodologi Data science diperlukan

1.1 Lingkup Data science

Pemanfaatan teknologi Kecerdasan Artifisial atau yang lebih dikenal sebagai Artificial Intelligence (AI) sudah semakin banyak untuk menyelesaikan suatu masalah melelui pengembangan solusi AI yang merupakan sistem/ aplikasi intelijen. Terdapat dua pendekatan umum untuk mengembangkan sistem intelijen, berdasar pakar atau sumber pengetahuan atau berdasar data yang menggunakan teknik Pembelajaran Mesin (*Machine Learning*).

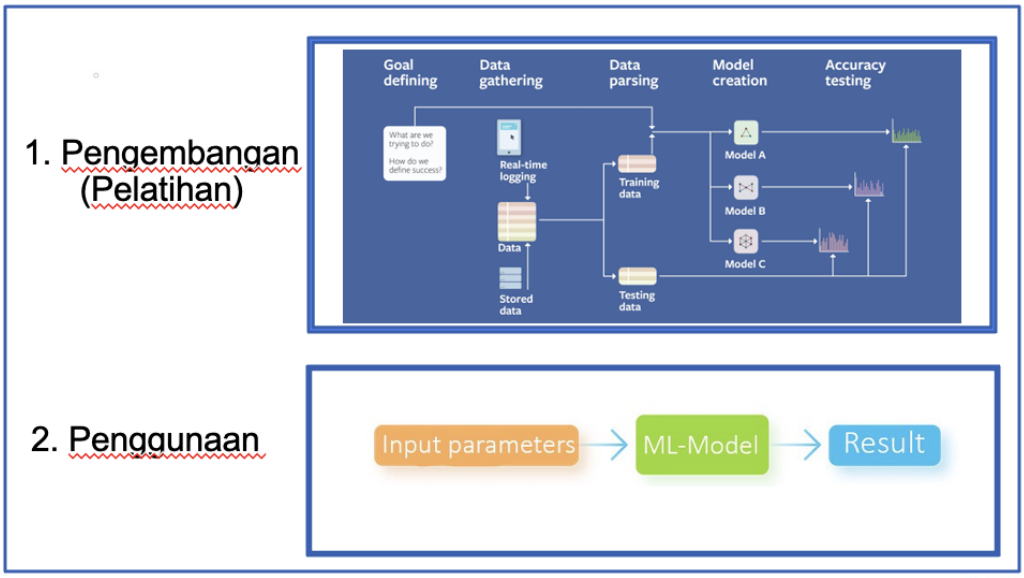
Pendekatan kedua merupakan pendekatan yang dilakukan oleh *data science* yang merupakan cabang ilmu yang memadukan paling tidak tiga disiplin keilmuan, yaitu Computer Science (bidang Artificial Intelligence terutama Machine Learning), Matematika/ Statistika serta *Domain Knowledge* (bidang terapannya).



Gambar 1. Data science sebagai interseksi tiga keilmuan

Untuk mewujudkan suatu sistem intelijen berdasar data maka dilakukan dua tahap:

- Pelatihan: Kegiatan untuk mencari pola/ pengetahuan dalam kumpulan data (basis data atapun big data) dan menggabungkan pengetahuan ke dalam suatu aplikasi intelijen
- Penggunaan: Pemanfaatan aplikasi intelijen untuk menjawab suatu pertanyaan.



Gambar 2. Tahapan Pemanfaatan Data untuk membuat Sistem Intelijen

Sistem yang dikembangkan memiliki kemampuan yang disesuaikan dengan tujuan dari tugas yang harus diselesaikan:

a. Deskriptif
b. Diagnostik
c. Prediktif
d. Preskriptif

Sistem deskriptif mencoba membuat suatu penjelasan keadaan saat ini. Dalam kasus untuk suatu organisasi adalah pemanfaatan sistem untuk melakukan penjelasan status/ keadaan keuangan berdasar data rasio keuangan. Sistem intelijen untuk tujuan ini dibentuk dari kumpulan data keuangan perusahaan (misalnya dari data Bursa Efek) dan setelah model terbentuk, sistem tersebut jika diberikan masukan data keuangan suatu organisasi tertentu akan memberikan penjelasan tentang kesehatan dari perusahaan tersebut. Pengetahuan yang dihasilkan proses pemodelan *data science* ini sering disebut sebagai model.

Sistem diagnostik berusaha menjelaskan mengapa suatu masalah tertjadi dengan melihat data historis. Misalkan suatu sistem medis dikembangkan supaya bisa memperkirakan mengapa seorang pasien mengalami berbagai gejala. Dnegan kata lain apa penyebab gejala tersebut terjadi pada pasien tersebut.

Sistem prediktif mencoba memproyeksikan/ memprediksi hasil di masa depan berdasarkan data historis. Sistem untuk memprediksi apakah suatu saham akan naik atau turun di masa depan merupakan salah satu contoh sistem prediktif.

Sistem preskriptif membawa aplikasi ke tahapan selanjutnya yaitu dengan memberi saran perbaikan. Jadi untuk sistem medis, tidak hanya mendiagnosa penyakit yang diderita tapi juga menyarankan penanganannya atau obat yang sebaiknya diberikan.

Sementara jenis tugas yang bisa dikembangkan adalah:

1. Regresi atau estimasi
2. Klasifikasi
3. Klustering
4. Asosiasi
5. Deteksi Anomali
6. Sequence Mining
7. Rekomendasi

Pembahasan tentang jenis-jenis tugas analitiks akan dijelaskan pada saat pembahasanan Business Understanding (sub bab 3.1).

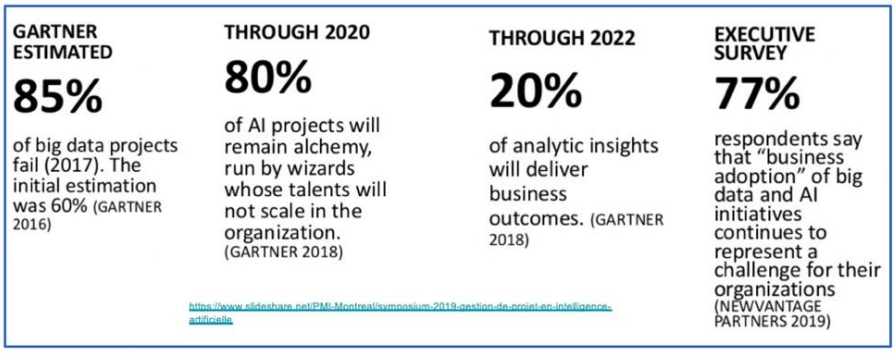
1.2 Masalah dalam Pengembangan Data science

Meskipun projek pengembangan Data science banyak dianggap projek yang dapat memberikan keuntungan pada suatu organisasi karena menggunakan berbagai teknologi AI terutama berbagai algoritma Pembelajaran Mesin (*Machine Learning)* dan dilengkapi dengan basis data atau *big* data namun dalam kenyataannya tujuan yang ingin dicapai tidak berhasil. Kegagalan atas projek Data science lebih besar dari kegagalan projek Teknologi Informasi lainnya. Gambar di bawah menjelaskan hal tersebut.

Alasan terjadinya kegagalan tersebut bervariasi. Beberapa penyebab utama adalah:

1. Lingkup Masalah

   Masalah yang ingin diselesaikan tidak jelas ataupun masalahnya bukan masalah yang dapat diselesaikan dengan menggunakan data. Di sisi lain, bidang AI menjadi ‘jargon’ yang dianggap bisa menyelesaikan segala masalah sehingga ‘over promise’.



Gambar 3. Kegagalan projek Data science sangat tinggi

2. Data

   Data yang diperlukan sering tidak ada, jumlah terbatas atau kualitasnya buruk (banyak error, tidak lengkap, dan lainnya). Pengembang (data scientist) tidak mengerti tentang arti data tersebut sehingga tidak tepat penggunaannya. Masalah lainnya adalah bahwa data bersifat bias (memihak) yang dapat menyebabkan model yang dihasilkan tidak tepat.

3. Model

   Model yang dihasilkan kurang tepat karena tidak mendapatkan model yang cukup akurat ataupun model yang dihasilkan tidak dapat dimengerti. Kedua hal ini meyebabkan model tidak mendapat dukungan untuk dipasang alias projek dianggap gagal.

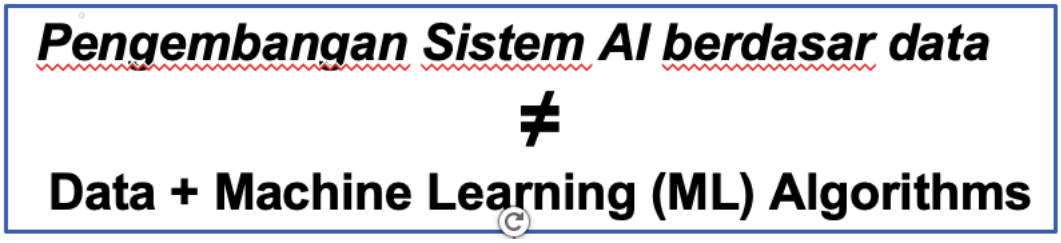
4. Algoritma kompleks

   Pemilihan algoritma pembelajaran mesin sangat mempengaruhi model yang dapat dibentuk. Terlalu banyaknya algoritma yang berbeda paradigmanya mempersulit data scientist. Akibatnya hanya algoritma tertentu yang dikuasai yang diimplementasikan, padahal model mungkin bisa lebih baik jika algotima lain yang dipergunakan.

5. Sumber Daya Manusia

   Kegiatan data science sering dianggap kegiatan seorang atau beberapa data scientist, padahal dalam kenyataannya berbagai pihak terlibat dalam pengembangan sistem intelijen ini.

Jadi pendekatan bahwa pengembangan sistem intelijen sebagai suatu kegiatan penggunaan data dan pemanfaatan algoritma Machine Learning pada data tersebut tidaklah tepat.



Gambar 4. Data dan Algoritma tidaklah cukup sebagai Kegiatan data science

Diperlukan suatu metodologi untuk membuat itu menjadi suatu sistem intelijen yang berhasil dimanfaatkan (dipasang/ dideploy dan dipergunakan). Metodologi Pengembangan didefinisikan sebagai

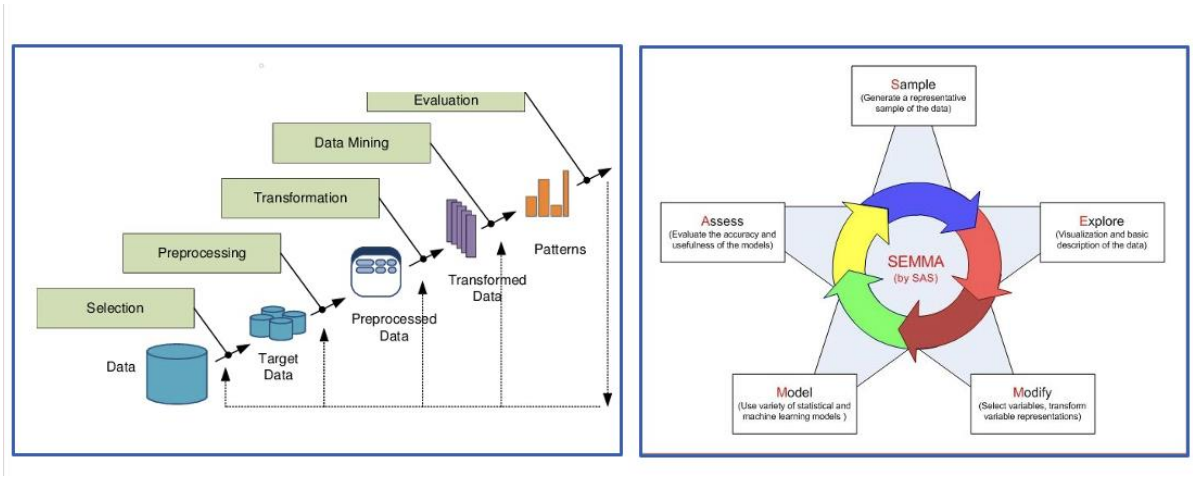
> Metoda iterative yang dipakai untuk menyelesaikan masalah dengan menggunakan data dan pendekatan data science melalui urutan langkah yang ditentukan.

2. Berbagai Metodologi Data science

Secara umum terdapat dua kelompok metodologi, metodologi teknis dan metodologi bisnis.

2.1 Metodologi Teknis

Metodologi teknis dimulai dari data yang lalu diproses untuk mendapatkan pola yang berguna. Dua diantara metodologi ini adalah metodologi Knowledge Discovery and data Mining (KDD) dan Metodologi SEMMA.



Gambar 5. Metodologi KDD dan SEMMA

KDD merupakan proses pemanfaatan metoda Data Mining untuk mengekstraksi pengetahuan sesuai dengan ukuran atau threshold yang ditentukan. Proses dimulai dengan adanya sekumpulan data (dataset) yang akan mengalami serangkaian proses sebagai berikut:

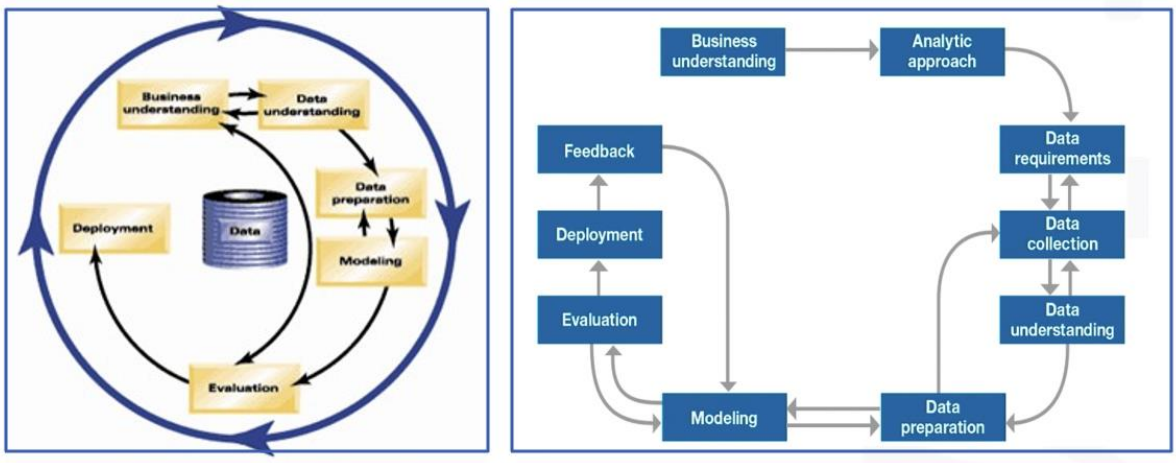
- Selection: Pemilihan data (data target) yang akan menjadi sampel untuk proses selanjutnya.
- Preprocessing data: Melakukan serangkaian proses untuk melengkapi data dan menjaga konsistensi data.
- Transformation: Mengubah representasi data untuk mempermudah dan memperbaiki agar sesuai dengan Teknik data mining yang akan dipergunakan
- Data Mining: Kegiatan pengembangan model untuk mencari pola dari data yang diberikan
- Evaluation: Proses interpretasi dan evaluasi pola yang diperoleh apakah pola yang menarik, berguna atau relevan.

Sementara metodologi SEMMA sesuai dengan namanya melakukan serangkaian kegiatan yang bersifat siklik (berulang) yaitu:

- *Sample*: Proses ekstraksi data untuk mendapatkan dataset yang cukup untuk mendapatkan informasi signifikan namun tidak terlalu besar sehingga mudah untuk diproses selanjutnya.
- *Explore*: Proses untuk mengeksplorasi data dengan mencari *trend* dan anomali untuk mendapatkan pemahaman tentang data
- *Modify*: Proses modifikasi data dengan membuat, memilih dan transformasi variable untuk proses pemodelan
- *Model*: Proses pemodelan dari data dengan mencari secara otomatis kombinasi data yang dapat dipakai untuk prediksi
- *Assess*: Mengevaluasi pola yang ditemukan apakah berguna dan cukup andal.

2.2 Metodologi Bisnis

Berbeda dengan metodologi sebelumnya, kelompok metodologi bisnis menempatkan kegiatan data science sebagai kegiatan yang berawal dari pemahaman masalah bisnis, yang disebut kegiatan *business understanding* atau *ideation*.



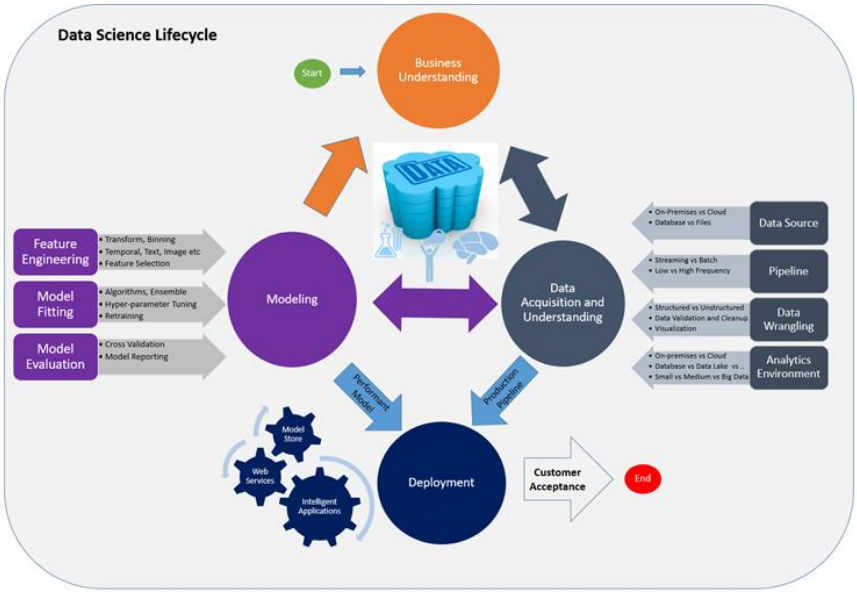
Gambar 6. Metodologi CRISP-DM dan IBM Data science

CRISP-DM dan Metodologi Data science dari IBM diawali dengan kegiatan Business Understanding yang merupakan proses pemahaman terhadap masalah yang akan diselesaikan. Di dalam kegiatan tersebut juga dilakukan proses pemetaan antara masalah bisnis dengan tugas analitiks (tugas *data science* yang sesuai). Pada Metodologi IBM hal tersebut dipisah menjadi proses *Analytic Approach*.

Berikutnya kegiatan pemahaman terhadap data (*Data Understanding*) yang meliputi penentuan kebutuhan data, pengumpulan data dan eksplorasi data. Pada Metodologi IBM masing-masing sub kegiatan dijadikan proses tersendiri.

Langkah berikutnya adalah *Data Preparation* yang dilakukan untuk memperbaiki kualitas data agar sesuai dengan proses Modeling yang akan dilakukan berikutnya. Kualitas model yang dihasilkan di evaluasi (*Evaluation*) sebelum dideploy menjadi sistem operasional. Rangkaian kegiatan diakhiri dengan proses *feedback* dan pelaporan.

Metodologi lain adalah dari Microsoft. Sama dengan sebelumnya, proses diawali dengan kegiatan *Business Understanding*. Daftar proses utamanya adalah sebagai berikut

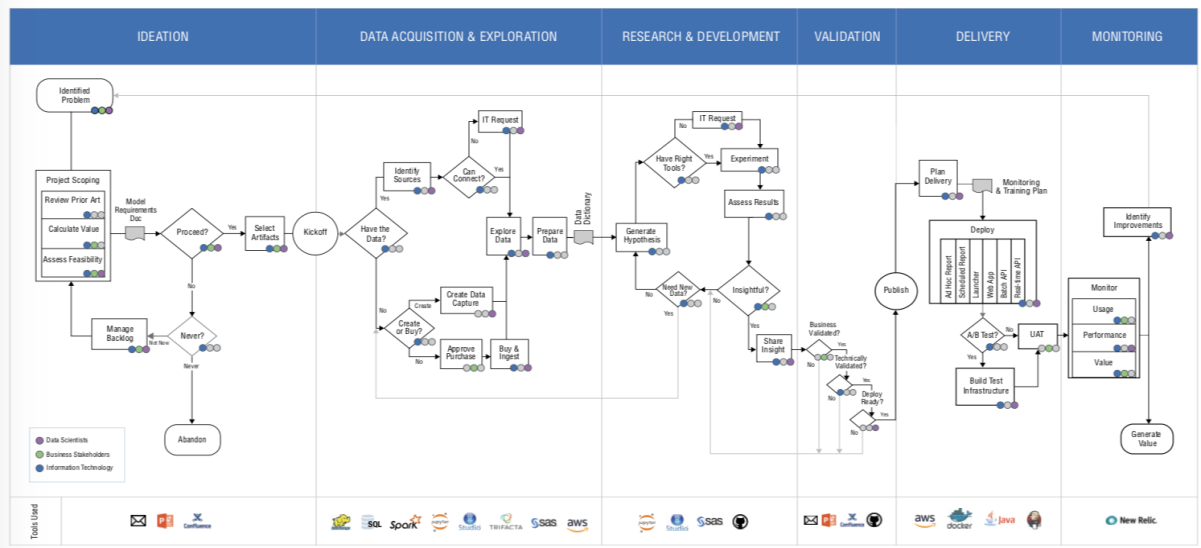


Gambar 7. Metodologi *Data science Life Cycle* dari Microsoft

- *Business Understanding*: Kegiatan untuk memahami masalah yang dihadapi
- *Data Acquisition and Understanding:* Kegiatan yang meliputi proses pengumpuilan dan eksplorasi data. Data bisa diambil dari data internal (on promise) ataupun dari cloud dan bisa berupa database ataupun file flat. Proses dilakukan melalui *pipeline*, yang dapat berupa proses *batch* atau *streaming*. Eksplorasi (*data wrangling*) meliputi pembersihan data, validasi dan visualisasi.
- *Modeling*: Pengembangan model yang meliputi *feature engineering*, *model fitting*, dan *model evaluation*.
- *Deployment*: Pemasangan model ke dalam aplikasi intelijen, suatu web service atau objek pada model store. Proses diakhiri dengan UAT (Customer Acceptance)

Metodologi lainnya adalah Metodologi dari Domino (Domino DataLab Methodology). Proses utama pada metodologi ini adalah:

- Ideation adalah pemahaman terhadap masalah pada proses bisnis serta identifikasi objektif bisnisnya. Langkah berikutnya adalah melakukan perhitungan terhadap objektif bisnis tersebut beserta Cost-Benefit Analysis.
- Data Acquisition and Preparation: Menentukan data yang diperlukan baik yang berasal dari sistem internal ataupun eksternal. Setelah proses akuisisi dilakukan eksplorasi terhadap data dan juga proses persiapan data.



Gambar 8. Metodologi *Domino DataLab*

- Research and Development: Pemodelan dilakukan sebagai suatu kegiatan pembuktian hipotesa dan pemodelan. Jika hasil sudah dianggap cukup maka dilakukan kegiatan berikutnya sementara jika belum dilakukan perbaikan data atau perubahan hipotesa. Dalam proses eksperimen, selaain metrik statistic dipergunhakan juga KPI organisasi.
- Validation: Model yang sudah dibuat divalidasi dari sudut bisnis dan teknis sebelum dipasang (deployment)
- Delivery: Deployment yang dimulai dengan perencanaan, lalu pemasangan dan perawatan sistem. Dalam proses ini juga dilakukan UAT (User Acceptance Testing).

Metodologi Domino juga dilengkapi daftar personal yang terlibat pada setiap langkah baik *data scientist, business people*, dan petugas *Information technology Division*. Juga dilengkapi daftar tools yang bisa dipergunakan dalam setiap langkah metodologi.

Hal yang perlu diperhatikan terkait dengan metodologi data science dalah adanya standard kompetensi kerja nasional di bidang data science yaitu Standard Kompetensi Kerja Nasional Indonesia (SKKNI): KepMen Ketenagakerjaan No 299 thn 2020. Di dalam SKKNI tersebut terdapat 21 (dua puluh satu) unit kompetensi yang diperlukan dalam membuat aplikasi intelijen menggunakan data science. SKKNI ini menjadi dsar pelatihan yang sekarang kita lakukan.

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

Gambar 9. Unit Kompetensi Data science sesuai SKKNI No 299 tahun 2020

3. Langkah Generik Pengembangan Aplikasi dengan *Data science*

Langkah generic di sini menggunakan standard SKKNI No 299 Tahun 2020. Terdapat tujuh kelompok kegiatan sebagai berikut.

3.1 Business Understanding: Menentukan Masalah Bisnis

Langkah pengembangan sistem intelijen dari data diawali dengan penentuan masalah yang ingin diselesaikan. Masalah bisnis yang akan diselesaikan harus jelas dan terukur. Artinya masalahnya dapat dimengerti oleh semua yang terlibat (data scientist, manajemen, project/product manager, staf IT, dll.) dan status pencapaiannya dapat diukur. Terdapat empat langkah utama yang harus dilakukan.

3.1.1 Menentukan Masalah Bisnis

Misalkan dalam suatu perbankan terdapat masalah bahwa jumlah pinjaman yang gagal bayar sangat tinggi, rasio Non-Performing Loan (NPL) tinggi. Masalah ini harus dipertegas agar dapat diselesaikan dengan *data science* atau tidak. Biasanya, pemberian pinjaman diberikan berdasarkan ranking *Credit Score* dari calon peminjam. Jika *Credit scorenya* baik maka pinjaman akan diberikan sementara jika tidak baik maka permohonan akan ditolak. Kegagalan bayar dapat diartikan bahwa penentuan *credit score* kurang tepat. Jadi permasalahan NPL ini didefinisikan sebagai perbaikan penentuan *Credit Score*. Perbaikan akan dicapai jika setelah diimplementasikan maka nilai NPL menurun.

3.1.2 Menentukan Tugas Analitiks

Setelah masalah secara jelas didefinisikan dan ukuran kesuksesan bisnisnya juga sudah ditentukan maka masalah tersebut harus diubah menjadi suatu tugas analitik. Jadi masalah bisnis diubah menjadi suatu masalah analitik yang akan dikembangkan aplikasinya.

Terdapat beberapa tugas yang dapat dipilih sesuai dengan masalah yang dihadapi.

3.1.2.1 Regresi

Regresi adalah kegiatan untuk menentukan suatu nilai kontinyu dari kasus yang diberikan. Masalah NPL di atas dapat dirumuskan sebagai masalah regresi jika Credit

Score yang harus ditaksir adalah suatu nilai (misalnya dalam skala 0 yeng menunjukkan status tidak dapat dipercaya hingga 100 yang menyatakan status dapat dipercaya secara absolut). Dengan demikian, sistem yang akan dibuat akan memberikan suatu nilai (antara 0 dan 100) yang menyatakan credit score seorang calon peminjam.

Contoh lain adalah masalah prediksi harga saham. Masalah ini merupakan masalah regresi karena tugas yang harus diselesaikan adalah membuat suatu sistem yang bisa memberikan suatu nilai (kontinyu) yang merupakan prediksi harga saham. Contoh lain adalah kegiatan menaksir harga rumah oleh seorang agen perumahan berdasarkan berbagai atribut yang diberikan dari suatu rumah.

3.1.2.2 Klasifikasi

Berbeda dengan regresi, klasifikasi adalah kegiatan memilih kelas atau kategori dari suatu kasus. Jumlah kelas atau kategori bisa dua atau lebih, dan sistem akan memilih salah satu dari kelas tersebut. Sebagai contoh masalah NPL di atas dapat dibuat menjadi maasalah klasifikasi jika credit score dinyatakan dalam 4 kelas: 1 Tidak dapat dipercaya, 2 agak dapat dipercaya, 3 bisa dipercaya, dan 4 sangat dipercaya. Jika didefinisikan seperti ini maka sistem intelijen klasifikasi yang akan dibuat akan memberi salah satu nilai jika diberi masukan mengenai satu calon peminjam.

Contoh lain adalah masalah penentuan apakah suatu perusahaan akan bangkrut atau tidak di masa depan (satu atau dua tahun lagi) berdasarkan data keuangan perusahaan tersebut saat ini. Jadi output dari sistem adarah bangkrut atau tidak. Ini adalah kegiatan klasifikasi.

3.1.2.3 Klastering

Berbeda dengan dua tugas sebelumnya, klastering tidak memberi nilai atau kelas tapi hanya mengelompokkan kumpulan kasus berdasarkan kemiripan. Sebagai contoh kita bisa membuat pengelompokkan customer suatu perusahaan menjadi beberapa kelompok berdasarkan sekumpulan atribut. Jika pengelompokkan sudah dilakukan maka ketika ada data baru maka dapat ditentukan kelompok mana yang paling tepat untuk data tersebut.

Dalam bidang medis misalkan untuk mengelompokkan pasien berdasarkan kemiripan kasusnya; di bidang perbankan misalkan untk melakukan segmentasi nasabah.

Kelompok yang dihasilkan dalam klastering tidak memiliki label. Biasanya pemberian label dilakukan manual setelah proses klastering selesai dilakukan.

3.1.2.4 Asosiasi

Asosiasi adalah tugas untuk memprediksi kumpulan kejadian yang terjadi bersama. Misalnya dalam bidang *retail*, menentukan kumpulan barang apa yang biasanya dibeli oleh seorang konsumen. Dengan mengetahui daftar tersebut maka dapat dilakukan berbagai kegiatan lain seperti *upselling/crossselling* ataupun strategi penempatan barang di suatu toko retail.

3.1.2.5 Anomaly Detection

Tugas ini menemukan kasus anomali atau kasus yang tidak biasa terjadi. Misalnya dalam pemrosesan transaksi pembayaran kartu kredit dalam transaksi ecommerce, anomaly detection dilakukan untuk mencari transaksi illegal/ tidak sah. Juga mendeteksi penerobosan jaringan merupakan kegiatan yang termasuk dalam tugas anomaly detection.

3.1.2.6 Sequence Mining

Dalam tugas ini akan diprediksi apa yang akan terjadi dari suatu urutan kejadian/ keadaan saat ini. Misalnya menentukan *next action* dari suatu kumpulan aksi seorang pembeli di ecommerce, apakah calon pembeli ini akan *exit* (tidak membeli) atau akan meneruskan langkah ke checkout.

Contoh lain adalah untuk menentukan apakah seorang pelanggan akan terus berlangganan atau akan berhenti dari berbagai interaksi yang diidentifikasi.

3.1.2.7 Rekomendasi

Sistem rekomendasi memberikan usulan atau rekomendasi bagi pengguna berdasar asosiasi preferensi dengan pengguna lain yang memiliki kesamaan sifat/ *taste* dengan

pengguna tersebut. Sebagai contoh rekomendasi film yang akan ditonton bisa diberikan setelah mempelajari film-film yang ditonton oleh pengguna lain yang perilaku menontonnya mirip. Sistem rekomendasi menjadi salah satu jenis sistem berbasis Artificial Intelligence yang paling banyak dipergunakan. E-commerce menggunakannya untuk memberi rekomendasi pembelian barang; Netflix untuk rekomendasi film, spotify untuk rekomendasi lagu, sementara berbagai situs berita menggunakannya untuk memberi rekomendasi berita yang perlu dibaca.

Setelah tugas tersebut diidentifikasi maka perlu dikenali metrik performansi yang sesuai sebagai ukuran yang harus dicapai oleh model yang akan dikembangkan. Setiap tugas memiliki metrik yang berbeda. Beberapa metrik yang biasa dipergunakan adalah

- Root mean Square error (RMSE)
- R-Square
- Jackard Index
- Log-loss
- Precision
- Recall
- F1-Score

3.1.3 Menentukan kebutuhan data

Langkah berikut yang harus dilakukan setelah menentukan tugas analitiks yang harus dipenuhi adalah menentukan data apa yang diperlukan dan darimana data tersebut dapat diperoleh. Tiga yang perlu dipikirkan adalah:

- Struktur data: Atribut atau konten apa saja yang perlu dikumpulkan.
- Jumlah data: Berapa banyak data yang diperlukan (jumlah rekordnya ataupun jumlah dokumen).
- Sumber data: Data bisa diperoleh dari sistem eksternal (ERP, database), eksternal (Web API, web scraping), dataset public ataupun open data.

3.1.4 Merencanakan manajemen projek

Setelah jelas apa yang akan dilakukan dan darimana data yang diperlukan dapat diperoleh maka kegiatan selanjutnya adalah membuat rencana pelaksanaan projek agar kegiatan *data science* ini dapat mencapai objektif bisnis yang telah ditentukan.

Dalam kegiatan ini paling tidak ada tiga hal yang perlu dilakukan.

- Pertama, Cost Benefit Analysis, untuk melihat apakah tujuan kegiatan akan menguntungkan bagi organisasi. Perkiraan biaya yang meliputi biaya personil tim pengembang (data scientist, projek manager, dan lainnya), biaya peralatan, biaya komputasi, biaya pengumpulan data, dan lain-lain; akan dibandingkan dengan potensi keuntungan yang akan diperoleh baik keuntungan finansial (pertambahan omset penjualan, pengurangan biaya operasi, dll.) maupun keuntungan non finansial (reputasi, dll.)
- Kedua, *Situation Assesment*, keadaan organisasi. Di tahap ini akan dianalisa berbagai aspek dari organsiasi seperti sumber daya yang dimiliki, asumsi-asumsi yang perlu diambil, risiko dan mitigasi, dan lain-lain.
- Ketiga, Project Plan seperti lingkup kegiatan dalam bentuk deskripsi *Work Breakdown Structure* (WBS), skedul, serta tim pengembang yang terlibat.

3.2 Data Understanding

Kegiatan *data understanding* merupakan kegiatan untuk memahami data lebih dalam. Tiga kegiatan utama adalah sebagai berikut:

3.2.1 Mengumpulkan data

Berdasarkan spesifikasi yang telah ditentukan dalam Langkah *Business Understanding*, langkah berikutnya adalah mengumpulkan data dari sumber yang telah ditentukan. Biasanya data disimpan dalam bentuk *big data* ataupun *dataset* tertentu sesuai dengan sistem pengolahan yang akan dipergunakan.

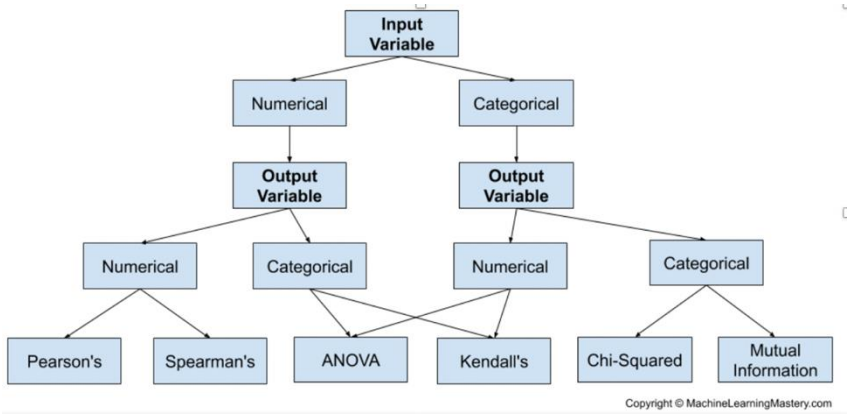
Data dikumpulkan bersama dengan deskripsi dari setiap atribut untuk memperjelas artinya. Deskripsi yang perlu dituliskan adalah arti dari dataset, sumbernya, waktu pengambilan data; arti dari satu rekor atau dokumen serta arti dari setiap atribut (melingkupi jenis data, rentang data, artinya, dan lain-lain).

3.2.2 Menelaah data

Kegiatan ini terkait dengan analisa data secara eksploratif (EDA- Exploratory Data Analysis) baik secara statistik dan juga visualisasi untuk mempermudah pemahaman tentang data.

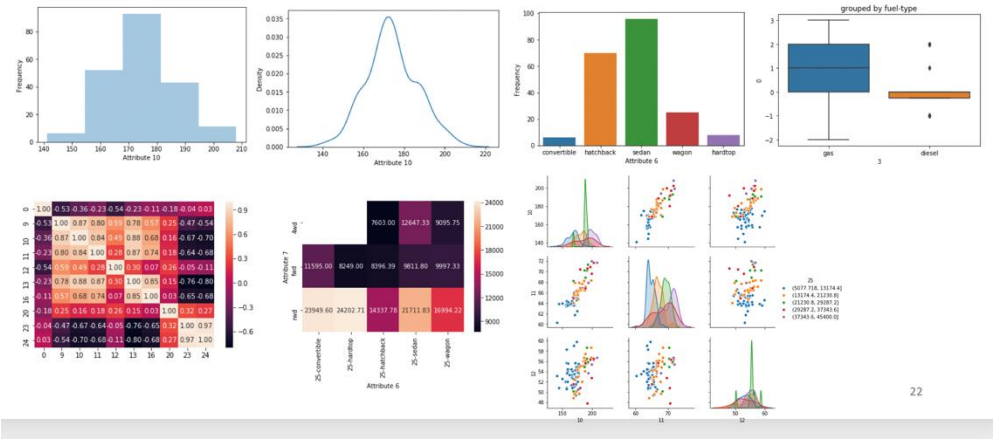
Analisa univariate dipergunakan untuk melihat deksripsi dari setiap atribut ataupun label yang menyatakan sebaran dari atribut tersebut. Selain itu juga dianalisa mengenai ketidaklengkapan data (atribut tanpa isi), kesalahan data (atribut dengan data illegal atau data yang mencurigakan), dan lain-lain.

Analisa multi atribut untuk melihat keterhubungan antara atribut dan antara atribut dengan label. Hubungan ini akan dipergunakan dalam langkah berikut yaitu pemilihan atribut/ fitur.



Gambar 10. Teknik Analisa multivariabel yang dapat dipergunakan

Teknik visualisasi dipergunakan untuk mempermudah analisa terhadap data baik visualisasi atribut tertentu ataupun visualisasi terhadap hubungan dua atribut. Masing-masing teknik visualisasi memiliki tujuan dan batasannya. Gambar di bawah menunjukkan beberapa visualisasi dasar yang biasa dipergunakan.



Gambar 11. Teknik Visualisasi Data

3.2.3 Memvalidasi data

Langkah validasi data menilai kesesuaian data dengan masalah yang akan diselesaikan. Dari analisa data yang sudah dilakukan dapat dinilai apakah data yang sudah dikumpulkan sesuai dengan ekspektasi, baik dari jumlah, atribut maupun kualitasnya. Salah satu kualitas yang perlu dianalisa adalah tentang berbagai *imbalanced* data (jumlah data tidak seimbang).

Pada kasus deteksi transaksi penggunaan kartu kredit illegal, biasanya jumlah data illegal jauh di bawah jumlah data legal. Perbandingan bisa 1:200. Hal ini perlu diperhatikan dan diselesaikan pada tahapan *data preparation* karena bisa memberikan hasil yang menyesatkan. Masalah *imbalanced*, juga bisa menyebabkan sistem AI bersifat *bias*/ memihak dan tidak fair. Hal ini dapat terjadi jika data yang dipergunakan tidak seimbang jumlahnya untuk atribut-atribut tertentu, misalnya atribut *gender*, atribut penganut agama, suku ataupun ras.

3.3 Data Preparation

Data preparation menggunakan hasil analisa yang telah dilakukan untuk mengubah data agar kualitas data meningkat dan bisa memperbaiki proses pemodelan. Setiap perubahan yang dilakukann terhadap data perlu didokumentasikan sebagai *audit trail*. Terdapat empat langkah utama dalam persiapan data yang dibahas di bawah ini.

3.3.1 Memilih dan memilah data

Tidak semua data yang telah dikumpulkan akan dipergunakan. Data dipilih berdasar kesesuaian kualitas data dan tujuan tugas analitiks yang telah ditentukan. Pemilihan data dilakukan berdasarkan rekord ataupun atribut (fitur). Pemilihan berdasarkan atribut dilakukan dengan menghilangkan atribut yang dianggap tidak layak untuk dipergunakan karena sifat atribut tersebut ataupun karena mirip sifatnya dengan atribut lain ataupun tidak ada kaitannya dengan fitur label.

Atribut nama pasien misalnya ataupun NIK bisa dihilangkan karena tidak ada hubungannya dengan fitur risiko operasi yang menjadi label. Tidak ada kaitan antara nama ataupun NIK dengan tingkat risiko operasi. Sementara atribut tanggal lahir dan umur tidak perlu dua-duanya dipergunakan karena keduanya berarti sama. Selain itu, jika ada dua atribut yang tidak berkorelasi dengan label maka tidak perlu dipergunakan.

Pemilihan data berdasarkan rekord dilakukan jika ada data duplikasi atau data yang terlalu banyak kosong (tidak lengkap) ataupun banyak errornya. Dapat juga terjadi bahwa rekor tidak dipergunakan karena tidak ada relevansinya dengan tujuan pemodelan. Misalnya, data pasien pria tidak perlu dipergunakan untk pemodelan risiko operasi Caesar karena hanya untuk pasien wanita.

3.3.2 Membersihkan data

Proses pembersihan data dilakukan untuk memperbaiki kualitas data. Pengisian atribut yang tidak lengkap dilakukan jika dapat diperkirakan isinya, misalnya dengan melihat pada rata-rata, median dari data atribut tersebut; ataupun dengan data dari hubungan dengan atribut lain. Pada kasus pertama, jika tipe data tersebut numerik maka data kosong bisa digantikan dengan median atau mean dari atribut tersebut atau menggunakan data yang paling sering muncul untk data kategorikal.

Untuk kasus kedua bisa menggunakan data median atau mean dari data lain yang memiliki nilai atribut sama. Sebagai contoh jika ada data pegawai yang kolom gaji kosong maka bisa diisi dengan rata-rata gaji dari data yang memiliki nilai atribut lain

yang sama, misalnya atribut golongan. Jadi data gaji dapat diisi dengan rata-rata data gaji untuk karyawan yang golongannya sama dengan data yang kosong tersebut.

3.3.3 Merekonstruksi data

Mengubah struktur data agar lebih cocok dengan algoritma. Prosesnya adalah normalisasi, transformasi data dan feature engineering

- Normalisasi, pengubahan skala dari atribut/ fitur. Skala suatu atribut diubah agar mirip dengan skala atribut lain. Hal ini untuk menghilangkan efek perbedaan skala yang bisa mempengaruhi kemampuan pemodelan.
- Transformasi data adalah proses pengubahan suatu data ke jenis lain. Data numerik bisa diubah ke dalam bucket (kelompok data), pengubahan ke data kategorikal, atau data kategorikal dubah menjadi berbagai atribut dummy, dll.
- Rekayasa fitur (feature engineering) mempergunakan algoritma seperti PCA (Principal Component Analysis) untuk melakukan reduksi dimensi data untuk memperbaiki pemodelan.

3.3.4 Integrasi data

Melakukan penggabungan data menjadi satu dataset jika sumber data berasal dari beberapa sumber. Yang perlu diperhatikan adalah arti dari setiap atribut dan arti setiap record awal. Rekord digabungkan jika artinya sama, misalkan ada dua data penjualan yang berasal dari Cabang A dan Cabang B. Selain itu perlu dilihat arti setiap atribut agar tidak ada duplikasi (dua atribut berbeda memiliki arti yang sama) dan juga konsistensi antar atribut.

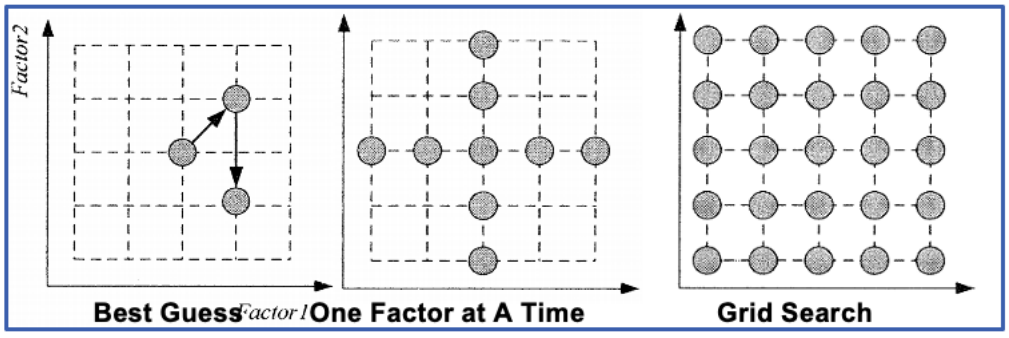
3.4 Modeling

Tahap berikut adalah pemanfaatan algoritma Machine Learning (ML) untuk membentuk model dari data yang sudah diperbaiki/ disiapkan. Target pemodelan adalah menemukan model terbaik yang dapat ditemukan menggunakan data yang sudah diproses. Masalahnya adalah teknik algoritma mana yang akan dipergunakan dan kombinasi parameter apa dari algoritma tersebut yang akan memberi hasil terbaik meski jumklah eksperimen dilakukan sesedikit mungkin. Untuk itu dilakukan beberapa langkah pada tahap pemodelan ini.

3.4.1 Membangun scenario pemodelan

Pembangunan skenario pemodelan dilakukan agar model terbaik bisa ditemukan dengan usaha sekecil mungkin. Detil langkahnya adalah sebagai berikut:

- Melakukan pemilihan algoritma yang sesuai: Dari ratusan algoritma yang ada dipilih beberapa algoritma sesuai dengan data dan tugas analitiks yang telah ditentukan. Untuk itu pengetahuan tentang karakteristik setiap algoritma utama perlu dikuasai agar dapat melakukan pemilihan algoritma setepat mungkin. Beberapa algoritma utama adalah sebagai berikut.
    1. **k-Nearest Neighbor (k-NN)**
    2. **Naïve Bayes**
    3. **Regression Techniques**
    4. **Support Vector Machines (SVMs)**
    5. **Decision Trees dan Random Forest**
    6. **Random Forests**
    7. **Deep Learning Algorithms**
- Membagi data: Data yang sudah diperbaiki dibagi menjadi data latih (dipergunakan untuk membentuk model) dan data uji (dipergunakan untuk mengukur performansi dari model yang sudah dibuat). Perbandingan 80: 20 atau 70:30 antara data latih dan data uji dapat dipergunakan.
- Melakukan Langkah eksperimen: Langkah eksperimen ditentukan sebagai strategi penemuan kombinasi parameter terbaik yang memberi model dengan performansi terbaik. Beberapa strategi eksperimen adalah *best guess*, *one factor at a time* atau *grid search*.

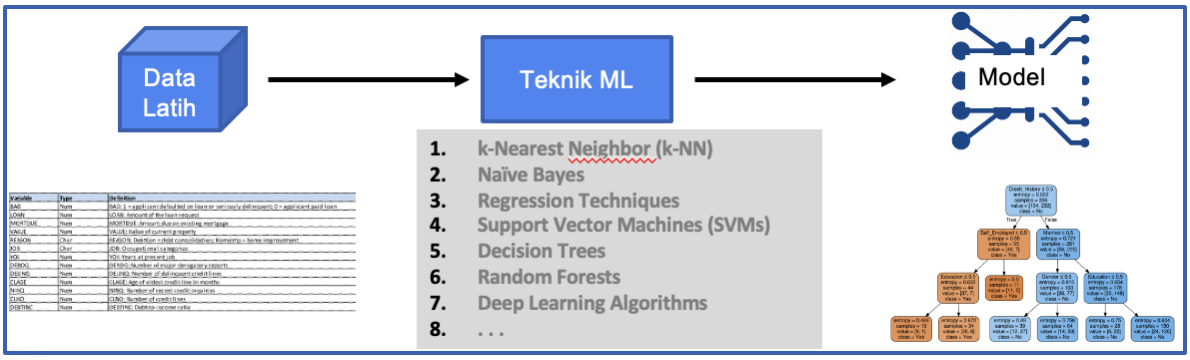


Gambar 12. Teknik Pengembangan Skenario Eksperimen

3.4.2 Membangun model

Dengan mengikuti strategi eksperimen, berbagai model dibangun untuk mencari model terbaik. Selama performansi belum dicapai secara maksimal (masih memungkinkan ditingkatkan) atau batasan tertentu belum dicapai (misalkan batasan

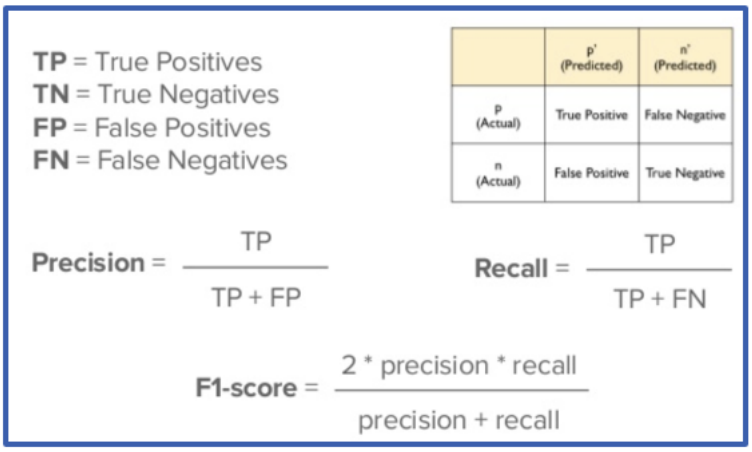
waktu, batasan biaya komputasi yang dipergunakan) maka pemodelan dilakukan dengan mengubah parameter sesuai dengan skenario eksperimen.



Gambar 13. Pemodelan

Dua kegiatan yang dilakukan dalam tahap ini yaitu membangun model dan menguji model:

- Membangun model dilakukan dengan menggunakan data latih
- Menguji model menggunakan model yang dibangun dan data uji. Metrik performansi yang dipergunakan disesuaikan dengan tugas analitik yang dilakukan. Artinya Ketika kita melakukan regresi maka metrik performansi yang dipergunakan akan berbeda dibandingkan ketika kita melakukan proses klastering.



Gambar 14. Metriks Performansi yang biasa dipakai

3.5 Model Evaluation

3.5.1 Mengevaluasi Model

Di dalam langkah ini dilakukan pengukuran performansi model yang sudah diperoleh dan analisa apakah model tersebut sudah cukup baik dari sudut teknis dan sudut bisnis

(domain) untuk dipergunakan. Karena jumlah model yang dikembangkan bukan hanya satu maka diperlukan suatu cara memilih model mana yang terbaik. Analisis ini dilakukan secara teknis ataupun bisnis.

Model terbaik secara teknis dianalisa menggunakan metrik performansi yang telah ditentukan. Selain itu perlu dilakukan uji hipotesis untuk mengukur signifikansi perbedaan performansi tersebut. Berbagai metoda uji hipotesa dapat dipergunakan. Sementara dari sisi bisnis, perbandingan dilakukan dengan berbagai kriteria bisnis misalnya aspek eksplainabilitas (kemampuan menjelaskan bagaimana model mencapai kesimpulan/ keputusan), metrik fairness (apakah sistem cenderung memihak pada salah satu golongan/ agama/ ras atau atribut spesial tertentu), ataupun dampak performansi terhadap bisnis (sebagai contoh perbedaan akurasi model bisa berarti perbedaan pendapatan bagi perusahaan), dan lainnya.

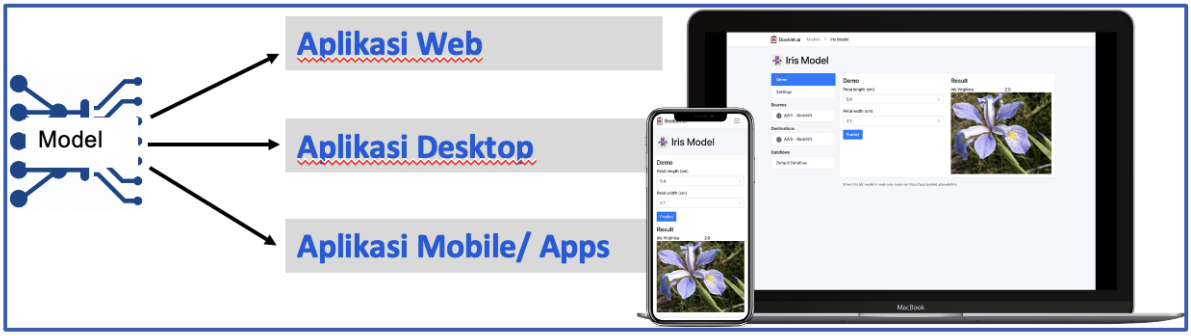
3.5.2 Mengevaluasi Proses

Melakukan review terhadap apa yang sudah dilakukan untuk mencari kemungkinan perbaikan dari sudut proses agar model lebih baik lagi. Setiap langkah yang sudah dilakukan direview untuk melihat apakah ada langkah yang belum optimal dilakukan. Jika dianggap sudah tidak ada yang terlewat berarti model yang dibentuk sudah maksimal dari data yang diperoleh.

3.6 Deployment

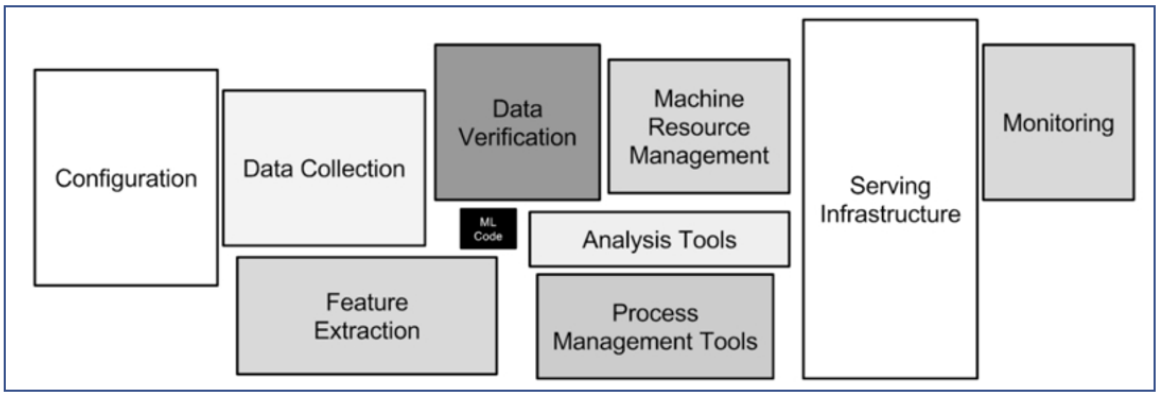
3.6.1 Memasang Model

Memasang model menjadi bagian dari aplikasi (sistem) intelijen untuk dipergunakan sesuai dengan task dan tujuan pembuatannya. Pemasangan model di lingkungan operasional, bisa dalam bentuk aplikasi web, mobile, API, dll.



Gambar 15. Deployment

Secara umum, aplikasi di-save ke dalam bentuk tertentu dan dalam aplikasi model diload dan diaktifkan serta digabungkan dengan interaksi terhadap komponen aplikasi lain. Biasanya pemasangan model dilakuikan oleh AI/ML Engineer bukan oleh seorang Data scientist.



Gambar 16. Model (ML Code) merupakan bagian kecil dari sistem intelijen

3.6.2 Memonitor Performansi Sistem

Model dikembangkan berdasar data (masa silam). Dalam operasi mungkin saja performansi aplikasi berbeda dengan performansi ketika dikembangkan atau diuji. Hal ini dikarenakan keadaan operasional sudah berbeda dengan keadaan pengembangan, ada berbagai data yang tidak bisa dikumpulkan dan lain-lain. Akibatnya performansi sistem harus dimonitor. Jika performansi sudah turun melewati threshold tertentu maka sistejm perlu dirawat.

Beberapa penyebab perbedaan performansi adalah:

- Data skews
    - Kekurangtepatan mendesain data latih: Distribusi data pelatihan tidak sama dengan distribusi data operasional.
    - Fitur yang diperlukan tidak dapat diperoleh Ketika dioperasikan sehingga harus diisi NULL yang menyebabkan keputusan yang dihasilkan tidak tepat.
    - Keterkaitan data: Model menggunakan data yang disimpan atau diproses dalam sistem lain yang tidak selalu bisa terhubung.
- Model Staleness
    - Perubahan lingkungan: Perubahan sebaran data karena perubahan populasi. Pola belanja berubah, barang yang diminati berubah, merupakan

dua contoh perubahan lingkungan yang akan mempengaruhi ketepatan model.

- Skenario adversarial: Serangan (fraud, criminal) akan mengaktifkan kelemahan model.

3.6.3 Melakukan Perawatan Model

Jika hasil monitoring memberi indikasi perlunya model dirawat maka modelpun diperbaiki, baik dengan mengubah beberapa parameter ataupun mengubah sisi aplikasinya. Setelah dirawat, aplikasi bisa dioperasionalkan lagi.

3.6.4 Memberhentikan Pemakaian Model

Jika proses perawatan terhadap model tidak bisa meningkatkan performansi maka ini berarti perlunya model diberhentikan penggunaaanya dan model baru perlu dikembangkan. Proses pengembangan model kembali dilakukan dari awal.

3.7 Project Evaluation

3.7.1 Mereview Project

Melakukan kilas balik terhadap apa yang sudah dilakukan. Apakah objektif projek sudah dicapai. Bagaimana performansi sistem dalam pengembangan dan operasional. Apakah perbedaannya signifikan. Apa lesson learned yang diperoleh dalam proses pengembangan aplikasi intelijen ini.

3.7.2 Melakukan Pelaporan Akhir

Menguraikan apa yang sudah dilakukan dari sudut business case denga apa yang sudah diperoleh dalam suatu laporan yang diberikan kepada pihak manajemen.

B. Keterampilan yang diperlukan dalam mengimplementasikan metodologi proyek data science

1. Mampu menganalisis permasalahan yang ada
2. Mampu mengidentifikasi tugas data science yang sesuai dengan masalah
3. Mampu mengidentifikasi langkah-langkah pengembangan aplikasi sesuai dengan tugas data science yang sudah diidentifikasi.

C. Sikap yang diperlukan dalam mengimplementasikan metodologi proyek data science

1. Disiplin
2. Analitik
3. Rasa ingin tahu yang tinggi
4. Bertanggung jawab
5. Kerja sama dalam tim

Tugas Dan Proyek Pelatihan
1. Membaca detil berbagai metodologi pengembangan data science 2. Melakukan proses pembandingan langkah antara berbagai metodologi yang sudah diibahas 3. Mendiskusikan faktor kesuksesan dan kegagalan pada tahapan generik metodologi data science. 4. Memberikan contoh untuk ketujuh jenis task analitiks yang dapat diselesaikan dengan data science. Misalnya untuk regresi dipilih satu permasalah, begitu pula untuk task yang lain. Pergunakan konteks perusahaan/ organisasi tempat bekerja untuk memudahkan.

Link Referensi Modul Ketiga
1. Standard Kompetensi Kerja Nasional Indonesia No 299 Tahun 2020 Bidang Keahlian Artificial Intelligence sub bidang Data science: https://skkni.kemnaker.go.id/tentang-skkni/dokumen 2. CRISP-DM: http://crisp-dm.eu/ 3. IBM Data science Methodology: https://www.slideshare.net/JohnBRollinsPhD/foundational-methodology-for-data-science 4. Microsoft Methodology: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview 5. Domino Methodology: https://www.dominodatalab.com/

Link Pertanyaan Modul Ketiga

Bahan Tayang
Power Point

Link room Pelatihan dan Jadwal live sesi bersama instruktur
Zoom

Penilaian
Komposisi penilaian Tugas Metodologi Data Science: Nilai 100

Target Penyelesaian Modul Ketiga
1 hari / sampai 6 JP

KOMINFO

Badan Penelitian dan Pengembangan SDM
Kementerian Komunikasi dan Informatika