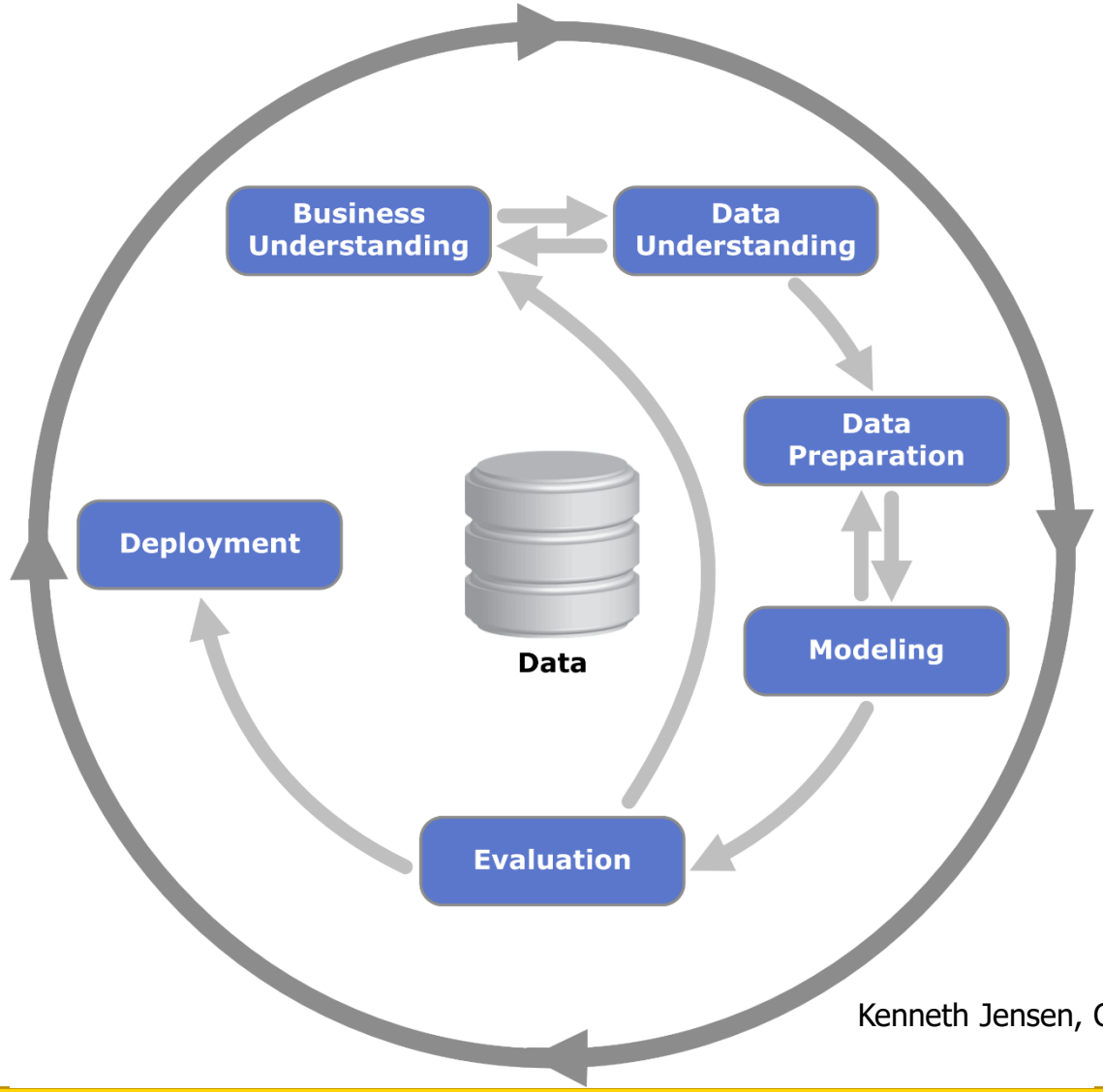
Metodologi Sains Data

Adila Alfa Krisnadhi*, Siti Aminah, Aruni Yasmin Azizah,
Dina Chahyati, Fariz Darari

CSGE603130 - Kecerdasan Artifisial dan Sains Data Dasar

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

CRISP-DM: Cross-industry standard process for data mining (Bisnis & Teknis)



Kenneth Jensen, CC BY-SA 3.0, via Wikimedia Commons

Business understanding

- Fokus pada definisi problem bisnis: "What problem are you trying to solve?". Contoh:
  - Business owner: "How can we reduce the costs of performing an activity?"
  - Interpretasi: "Is it to improve the efficiency of the activity?" or "Is it to increase business profitability?"
  - Lalu tentukan tujuan spesifik yang mendukung.
  - Pastikan dukungan dari semua stakeholder.
- Lalu tentukan pendekatan analitik yang tepat:
  - Descriptive → Current status; "Show relationships in the data" (clustering)
  - Diagnostic → What and why did it happen?
  - Predictive → What if this trend continues? What will happen next? (regresi jika perlu menentukan nilai bilangan riil dan probabilitas; klasifikasi jika perlu menjawab pertanyaan yes/no)
  - Prescriptive → How to solve it?

Contoh: Congestive Heart Failure Readmission

- What is the best way to allocate the limited healthcare budget to maximize its use in providing quality care?
- As public funding for readmissions was decreasing, this insurance company was at risk of having to make up for the cost difference, which could potentially increase rates for its customers.
- Knowing that raising insurance rates was not going to be a popular move, the insurance company sat down with the health care authorities in its region and brought in data scientists to see how data science could be applied to the question at hand.
- Classification problem: Choose a predictive model such as decision tree, etc.
  - "Given a profile of patient with congestive heart failure, is he/she going to be readmitted to a hospital within N month?"

Data understanding

- Tentukan prasyarat data yang sesuai dengan problem bisnis
  - Konten, format, representasi, dll
- Lakukan pengumpulan data (bisa terstruktur, takterstruktur, semi terstruktur)
- Pahami data menggunakan statistika dan visualisasi
- Iterasi jika perlu

Contoh: Congestive Heart Failure Readmission

Data requirement:

- Selecting a suitable patient cohort from the health insurance providers member base.
  - A patient needed to be admitted as in-patient within the provider service area, so they'd have access to the necessary information.
  - They focused on patients with a primary diagnosis of congestive heart failure during one full year.
  - A patient must have had continuous enrollment for at least six months, prior to the primary admission for congestive heart failure, so that complete medical history could be compiled.
- Congestive heart failure patients who also had been diagnosed as having other significant medical conditions, were excluded from the cohort because those conditions would cause higher-than-average readmission rates and, thus, could skew the results.
- Then the content, format, and representations of the data needed for decision tree classification were defined.

Contoh: Congestive Heart Failure Readmission

Data requirement:

- This modeling technique requires one record per patient, with columns representing the variables in the model.
- To model the readmission outcome, there needed to be data covering all aspects of the patient's clinical history.
- This content would include admissions, primary, secondary, and tertiary diagnoses, procedures, prescriptions, and other services provided either during hospitalization or throughout patient/doctor visits.
- Transformation required!

Contoh: Congestive Heart Failure Readmission

Data collection:

- Available data source: Corporate data warehouse (source of medical & claims, eligibility, provider, member information), in-patient record system, claim payment system, disease management program information
- After the initial data collection is performed, an **assessment by the data scientist** takes place to determine whether or not they have what they need.
- In this phase the data **requirements are revised** and decisions are made as to whether or not the collection requires more or less data.
- Once the data ingredients are collected, the data scientist will have a good understanding of what they will be working with.
- Techniques such as descriptive statistics & visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.
  - Univariate statistics on each variable (mean, median, mode, standard deviation, etc.)
  - Pairwise correlation to determine redundant variables and histogram to understand the variables' distribution.

Data preparation

- Mencakup semua aktivitas menyiapkan dataset untuk modeling
  - Cleaning, integrating from multiple sources, transforming,
  - Feature engineering untuk mendapatkan variabel prediktor baru, dan meningkatkan kinerja
- Paling memakan banyak waktu: 70-90% waktu proyek (bisa kurang jika sumber data sudah terkelola dengan baik).

Contoh: Congestive Heart Failure Readmission

Data preparation: Define congestive heart failure

- The set of diagnosis-related group codes needed to be identified, as congestive heart failure implies certain kinds of fluid buildup.
- Next, define the re-admission criteria for the same condition.
  - The timing of events needed to be evaluated in order to define whether a particular congestive heart failure admission was an initial event, which is called an index admission, or a congestive heart failure-related re-admission.
  - Based on clinical expertise, a time period of 30 days was set as the window for readmission relevant for congestive heart failure patients, following the discharge from the initial admission.

Contoh: Congestive Heart Failure Readmission

Data preparation: Prepare dataset

- Transactional records
  - Claims: professional provider, facility claims submitted for physician, laboratory, hospital, and clinical services.
  - Inpatient & Outpatient: diagnoses, procedures, prescriptions, etc
  - Possibly thousands of the records per patient
- Then, aggregate all transactional record to patient level:
  - Roll up to 1 record per patient
  - Create new columns representing the transaction: Outpatient visits/inpatient episodes, frequency, diagnoses, procedure, prescriptions, comorbidities with CHF
- Completing the dataset:
  - One record per patient
  - List of variables used in modeling. Target: CHF readmission with 30 days (Yes/No), following discharge from CHF hospitalization

Modeling

- Dimulai dari dataset yang sudah disiapkan (dari tahap data preparation), model prediktif atau deskriptif dikembangkan sesuai pendekatan yang dipilih dari tahap business understanding
- Bersifat iteratif: perlu melakukan beberapa kali eksperimen hingga kinerja memuaskan.

Contoh: Congestive Heart Failure Readmission

Modeling: Decision tree classification

- The best parameter to adjust is the relative cost of misclassified yes and no outcomes.
  - When a true, non-readmission is misclassified, and action is taken to reduce that patient's risk, the cost of that error is the wasted intervention → Type I error (false positive)
  - When a true readmission is misclassified, and no action is taken to reduce that risk, the cost of that error is the readmission and all its attended costs, plus the trauma to the patient →Type II error (false negative).

Evaluation

- Ilmuwan data:
  - mengevaluasi kualitas model
  - memeriksa apakah model sudah mengatasi masalah bisnis secara baik dan menyeluruh
- Membutuhkan penghitungan macam-macam ukuran kinerja serta sajian visual (tabel, graf) pada dataset uji.

Contoh: Congestive Heart Failure Readmission

Evaluation: Does the model used really answer the initial question, or does it need to be adjusted?

- While a data science model will provide an answer, the key to making the answer relevant and useful to address the initial question, involves getting the stakeholders familiar with the tool produced.
- In a business scenario, stakeholders have different specialties that will help make this happen, such as the solution owner, marketing, application developers, and IT administration.
- Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test.

Deployment

- Setelah model sudah memuaskan business stakeholder, model di-deploy pada lingkungan produksi atau ruang uji yang menyerupainya.
- Deployment awal difokuskan pada evaluasi kinerja awal pada lingkungan riil.
  - A/B testing banyak dipakai di sini.
- Deployment lebih luas pada proses bisnis secara operasional membutuhkan keterlibatan banyak pihak/bagian dalam organisasi.
- Setelah deployment diinisiasi, proses tidak berhenti.
  - Perlu evaluasi secara berkesinambungan untuk terus memperbaiki model.
  - Prinsip: "Machine learning models always get worse over time" (due to the presence of new data).

Contoh: Congestive Heart Failure Readmission

Deployment:

- Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test
  - Actual, real time use in the field
- In preparation for solution deployment, the next step was to assimilate the knowledge for the business group who would be designing and managing the intervention program to reduce readmission risk.
- In this scenario, the business people translated the model results so that the clinical staff could understand how to identify high-risk patients and design suitable intervention actions.
- Feedback is gathered from users to help refining the model and assessing its performance and impacts.

Credits

- Siti Aminah, Dhimas Arief Dharmawan, "Data Science Methodology", Salindia Kelas Sains Data, Semester Genap 2020/2021, Fasilkom UI.
- Gambar dan tangkapan layar hanya untuk kebutuhan penjelasan
  - Hak cipta tetap ada pada pemilik aslinya.