Capaian Pembelajaran

Pada topik ini, kita akan mempelajari:

- Melakukan pengecekan kelengkapan data
- Membuat rekomendasi kelengkapan data
- Melakukan pengujian kualitas data

Tahapan Data Preparation: Pemilihan, Pembersihan & Validasi

1. **Pilih/ Select Data**
   - Pertimbangkan pemilihan data
   - Tentukan dataset yang akan digunakan
   - Kumpulkan data tambahan yang sesuai (internal atau eksternal)
   - Pertimbangkan penggunaan teknik pengambilan sampel
   - Jelaskan mengapa data tertentu dimasukkan atau dikecualikan
2. **Bersihkan/ Clean Data**
   - Perbaiki, hapus atau abaikan noise
   - Putuskan bagaimana menangani nilai-nilai khusus dan maknanya
   - Tingkat agregasi, nilai yang hilang (missing value), dll
   - Bersihkan atau manipulasi outlier
3. **Validasi Data**
   - Periksa/Nilai Kualitas Data
   - Periksa/Nilai Tingkat Kecukupan Data

Paramater/Daftar Isi Dokumentasi Data Validation

Laporan dokumentasi data cleaning, setidaknya memiliki parameter berikut:

- Validasi data
  - Kebenaran, misal di Indonesia isian Gender yang diakui hanya 2 P/W; Agama hanya 6 (Islam, Protestan, Katholik, Hindu, Budha, Konghucu)
  - Kelengkapan, misal data provinsi seluruh Indonesia (34 prov), namun hanya sebagian yg ada
  - Konsistensi, misal penulisan STM atau SMK;
- Kecukupan data → Perlukan diulang berikan justifikasi (Resampling)

https://dilbert.com/strip/2008-05-07

Verifikasi vs. Validasi

- Verifikasi
  - Apakah Anda membuat produk dengan benar?
  - Perangkat lunak harus sesuai dengan spesifikasi
- Validasi
  - Apakah Anda membuat produk yang benar?
  - Perangkat lunak harus dikembangkan sesuai yang diperlukan pengguna

Validasi Data

- Verifikasi vs Validasi
  - Verifikasi: Benar vs Salah (sesuai prosedur)
  - Validasi: Kuat vs Lemah (sesuai kenyataan)
- Validasi merupakan tahapan kritikal yang sering diabaikan DS pemula, karena memeriksa, diantaranya sbb:
  - Tipe Data (mis. integer, float, string)
  - Range Data
  - Uniqueness (mis. Kode Pos)
  - Consisten expression (mis. Jalan, Jl., Jln.)
  - Format Data (mis. utk tgl "YYYY-MM-DD" VS "DD-MM-YYYY.") → tmt (terhitung mulai tanggal)
  - Nilai Null/Missing Values
  - Misspelling/Type
  - Invalid Data (gender: L/P: L; Laki-laki; P: Pria/Perempuan?)

- Teknik Validasi Data dan Model:
  - Akurasi
  - Kelengkapan
  - Konsistensi
  - Ketepatan Waktu
  - Kepercayaan
  - Nilai Tambah
  - Penafsiran
  - Kemudahan Akses

Validasi Data

- Hasil operasi validasi data dapat menyediakan data yang digunakan untuk analisis data, intelijen bisnis, atau melatih model pembelajaran mesin.
- Data dapat diperiksa sebagai bagian dari proses validasi dalam berbagai cara, termasuk tipe data, batasan, terstruktur, konsistensi, dan validasi kode.
- Validasi data berkaitan dengan kualitas data. Validasi data dapat menjadi komponen untuk mengukur kualitas data, yang memastikan bahwa kumpulan data yang diberikan dilengkapi dengan sumber informasi yang berkualitas tinggi, otoritatif, dan akurat.
- Validasi data juga digunakan sebagai bagian dari alur kerja aplikasi, termasuk pemeriksaan ejaan dan aturan untuk pembuatan kata sandi yang kuat.

Urgensi Validasi Data

- Untuk data scientist, data analyst, dan orang lain yang bekerja dengan data, memvalidasi nya sangat penting. Output dari sistem apa pun hanya bisa sebaik data yang menjadi dasar operasi.
- Operasi ini dapat mencakup pembelajaran mesin atau model kecerdasan buatan, laporan analisis data, dan dasbor intelijen bisnis.
- Memvalidasi data memastikan bahwa data tersebut akurat, yang berarti semua sistem yang mengandalkan kumpulan data yang diberikan telah divalidasi juga.

Validasi Data

- Reliability
- Validity
- Reproducibility
- Repeatability
- Accuracy

Kapan harus melakukan validasi data?

- Saat indikator baru diimplementasikan
- Data akan dipublikasi pada website atau dengan cara lain.
- Ada perubahan terhadap indikator yang ada sebelumnya.
- Data yang dihasilkan dari indikator yang ada telah berubah tanpa dapat dijelaskan.
- Sumber data telah berubah.
- Subyek pengumpulan data telah berubah.

Dua Tipe Reliability Test

- Repeatability: diulangnya pengukuran hasil oleh orang yang sama atau alat yang sama pada catatan yang sama dan kondisi yang sama.
- Reproducibility: diulangnya pengukuran hasil oleh orang yang berbeda pada catatan yang sama dan kondisi yang sama.
    - Reproduced, or replicated, oleh seseorang yang bekerja secara independent.

Kenapa Kualitas Data Penting

- Berpengaruh terhadap keputusan organisasi
  - Data yang hilang atau salah dapat mengakibatkan kesalahan pembuatan keputusan
- Berpengaruh terhadap model machine learning
  - Data yang bersih dapat meningkatkan performansi model
- Berpotensi menghasilkan keputusan yang bias dalam sistem ML/AI.
- Stabilitas operasional: inkonsistensi data dapat menyebabkan malapetaka pada sistem produksi

Data: Akademis vs. Real-world

- Dataset akademis
  - Statis
  - Seringkali down-sampled, sudah bersih sebelum dipublikasikan
  - Fitur dapat dapat dipahami
  - Contoh: UCI ML datasets
- Real-world data
  - Secara konstan bisa berubah
  - Dapat terdiri dari banyak fitur yang tidak dipahami
  - Data berasal dari sumber yang berbeda
  - Seringkali sulit diakses (contoh: data terkompresi dan terpartisi dalam suatu sistem terdistribusi)

Perubahan dalam Strategi Pengumpulan Data

- Pre-internet Era
  - Data dikumpulkan dalam bentuk basisdata relasional
  - ETL (Extract, Transform, Load) mengekspor data ke data warehouse untuk dianalisis
  - Melakukan pemodelan data

- Internet Era: Kumpulkan datanya dulu, analisis kemudian
  - Kemajuan internet meningkatkan jumlah data semi-terstruktur yang banyak Karakteristik Big Data (Velocity, Veracity, Variety, Volume)
  - Data baru yang terkumpul sudah “mapan” (key-value stores, document databases, data lakes)
    - Diskalakan ke ukuran data yang makin besar
  - Insentif ekonomi
    - Menurunkan biaya storage
    - Data menjadi bernilai sebagai input untuk aplikasi berbasis ML

Sumber Kesalahan Data

- Kesalahan entri data
  - Kesalahan dalam pengisian atau tipe data
  - Perbedaan dalam pengejaan data
- Kesalahan pengukuran
  - Penempatan sensor yang tidak tepat
  - Gangguan dalam proses pengukuran
- Kesalahan integrasi data
  - Inkonsistensi data, duplikasi data
  - Kesalahan satuan pengukuran data
- Kesalahan penyaringan data
  - Bias pada editorial ringkasan data

Dimensi Kualitas Data

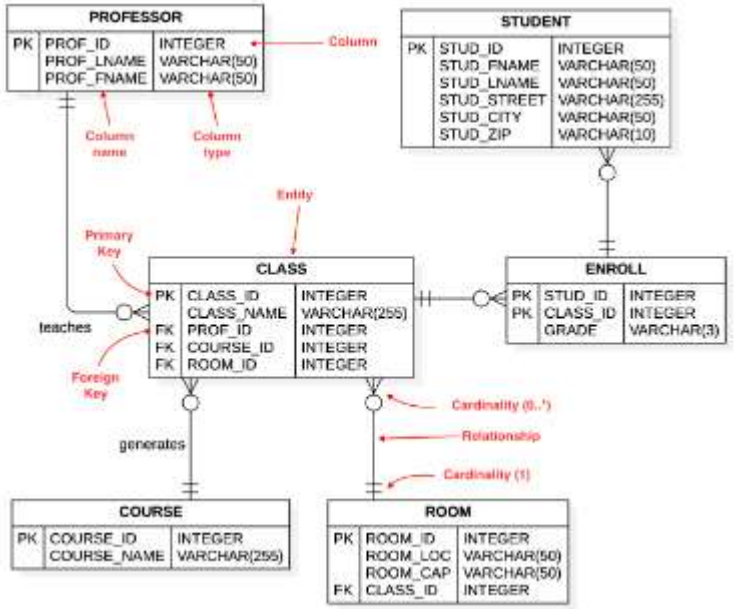
- Kelengkapan
  - Sejauh mana tersedianya data yang diperlukan untuk menggambarkan objek dunia nyata
- Konsistensi: Batasan intra-relasi (rentang nilai yang dapat diterima)
  - Tipe data spesifik, interval untuk kolom numerik, kumpulan nilai untuk kolom kategoris
- Konsistensi: Kendala antar-hubungan
  - Validitas referensi data ke entri data lain (misalnya, “primary keys” & “foreign keys" dalam database)

Pendekatan untuk Meningkatkan Kualitas Data

- Desain antarmuka entri data
  - Terapkan batasan integritas (misal, batasan pada nilai numerik, integritas referensial)
  - Dapat memaksa pengguna untuk “menemukan” data kotor
- Manajemen organisasi
  - Penyederhanaan proses untuk pengumpulan dan analisis data
  - Menangkap metadata
- Pengauditan data otomatis dan pembersihan data
  - Penerapan teknik otomatis untuk mengidentifikasi dan memperbaiki kesalahan data
- Analisis data eksplorasi dan pembersihan data
  - Pendekatan *human-in-the-loop* diperlukan pada sebagian besar proses
  - Interaksi antara visualisasi data dan pembersihan data
  - Iterasi proses

Pemeriksaan Tipe Data

- Pemeriksaan tipe data mengkonfirmasi bahwa data yang dimasukkan memiliki tipe data yang benar.
- Misalnya, field atau isian data yang hanya menerima data numerik.
- Sehingga hal yang harus diperhatikan adalah kesesuaian tipe data yang berkaitan dengan entitas relasinya *(Entity Relationship Diagram)*.



Pemeriksaan Format Data

- Banyak tipe data mengikuti format standar tertentu. Kasus penggunaan yang umum adalah kolom tanggal yang disimpan dalam format tetap seperti YYYY-MM-DD atau DD-MM-YYYY.
- Prosedur validasi data yang memastikan tanggal dalam format yang tepat membantu menjaga konsistensi data dan waktu.
- Format data sangat penting dalam pengolahan data, sehingga hal ini akan menjadi krusial dan harus diperhatikan pula oleh seorang data scientist dalam kasus menangani data *time-series*.

Pemeriksaan Konsistensi dan Pemeriksaan Keunikan

- Pemeriksaan konsistensi adalah jenis pemeriksaan logis yang mengonfirmasi bahwa data telah dimasukkan dengan cara yang konsisten secara logis.
    - Contohnya adalah memeriksa apakah tanggal pengiriman setelah tanggal pengiriman untuk sebuah paket.
- Sementara untuk keunikan, beberapa data seperti ID atau *primary key*. Database kemungkinan harus memiliki entri unik di bidang ini. Pemeriksaan keunikan memastikan bahwa item tidak dimasukkan beberapa kali ke dalam database.

Pembersihan Data: Tipe dan Teknik

- Data kuantitatif
  - Bilangan bulat atau bilangan floating point dalam berbagai bentuk (set, tensor, deret waktu)
  - Tantangan: konversi unit (terutama untuk unit yang mudah berubah seperti mata uang)
  - Dasar teknik pembersihan: deteksi outlier
- Data kategori
  - Nama atau kode untuk menetapkan data ke dalam grup, tidak ada urutan atau jarak yang ditentukan
  - Masalah umum: salah mengeja saat entri data
  - Dasar teknik pembersihan: normalisasi / deduplikasi
- Free-text Entry
  - Kasus khusus dari data kategorikal, biasanya dimasukkan sebagai teks bebas
  - Tantangan utama: deduplikasi
- Pengidentifikasi / Kunci
  - Pengidentifikasi unik untuk objek data (misalnya, kode produk, nomor telepon, ID)
  - Tantangan: mendeteksi penggunaan kembali pengidentifikasi di seluruh objek yang berbeda
  - Tantangan: Pastikan integritas data

Robust Univariate Outlier Detection

- Analisis univariat
  - Pendekatan sederhana: selidiki kumpulan nilai dari satu atribut dari kumpulan data
  - Perspektif statistik: nilai yang dianggap sebagai sampel dari beberapa proses pembuatan data
- Center & Dispersion
  - Kumpulan nilai memiliki pusat yang mendefinisikan apa yang menjadi nilai "rata-rata"
  - Himpunan nilai memiliki dispersi yang mendefinisikan apa yang "jauh dari rata-rata"
- Deteksi outlier
  - Asumsi: nilai yang salah jauh dari distribusi normal nilai dalam himpunan
  - Pendekatan: mengidentifikasi outlier menggunakan teknik statistik
  - Masalah: Bagaimana cara menghitungnya dengan andal ketika datanya kotor / salah?

Sampel Data Usia

- Kumpulan data usia pegawai dalam suatu perusahaan:

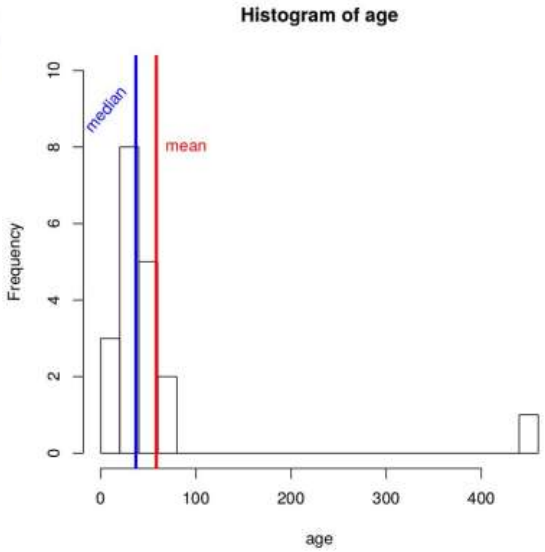
12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

Sampel Data Usia

- Kumpulan data usia pegawai dalam suatu perusahaan:

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450

- Pendekatan potensial:
  - Asumsikan distribusi normal dari nilai umur
  - Hitung rata-rata dan simpangan baku
  - Tandai nilai 2 standar deviasi dari rata-rata
  - Intervalnya adalah [96 – 2 * 59, 9 + 2 * 59] = [-22, 127]



Normalisasi Data String

- Fingerprint keying: hapus tanda baca dan sensitivitas huruf besar-kecil:
  - Hapus spasi di sekitar string
  - Temukan padanan karakter ASCII
  - Urutkan fragments dan menghapus duplikatnya
    - ACT, INC □ act inc
    - ACT INC □ act inc
    - ACT,Inc □ act inc
    - Act Inc □ act inc
- Entri teks bebas dari atribut kategori sangat rawan kesalahan:
  - Ejaan yang berbeda (Jérôme vs Jerome)
  - Tanda baca yang berbeda (ACME Inc. vs ACME, Inc)
  - Kesalahan ketik (Alice → Ailce)
  - Kesalahpahaman (Rupert → Robert)

Missing Value Imputation

- Data yang hilang adalah masalah kualitas data utama
  - Hilang karena berbagai alasan
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR)
- Berbagai cara untuk menangani data yang hilang untuk aplikasi ML
  - Analisis kasus lengkap (hapus contoh dengan atribut yang hilang)
  - Tambahkan simbol placeholder untuk nilai yang hilang
  - Hitung nilai yang hilang
    - Sering diimplementasikan dengan teknik dari library ML popular, seperti mean dan mode imputasi
    - ML: supervised learning untuk imputasi nilai yang hilang

Categorical Data Imputation

- Asumsikan data tabular
  - Ingin memasukkan nilai yang hilang dalam kolom dengan data kategoris
- Ide: menerapkan teknik dari supervised learning
- Contoh: katalog produk, colors missing
- Perlakukan masalah imputasi sebagai masalah klasifikasi multi-kelas

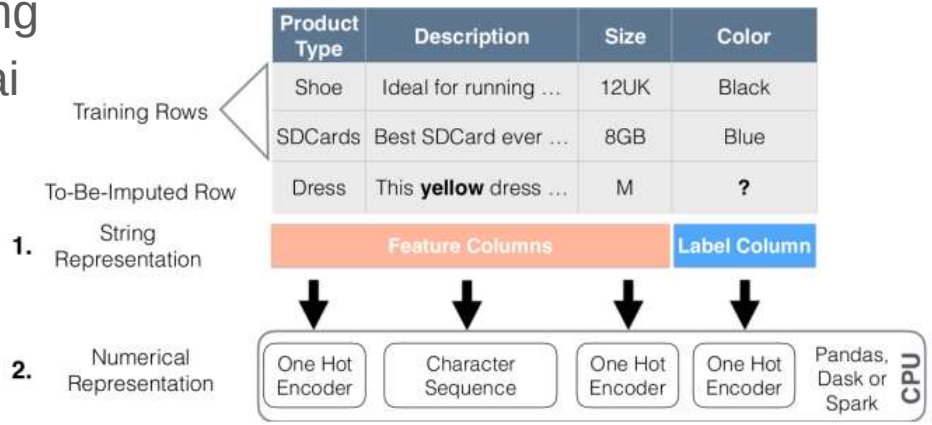
Product Type	Description	Size	Color
Shoe	Ideal for running …	12UK	Black
SDCards	Best SDCard ever …	8GB	Blue
Dress	This **yellow** dress …	M	**?**

Categorical Data Imputation

- Asumsikan data tabular
  - Ingin memasukkan nilai yang hilang dalam kolom dengan data kategoris
- Ide: menerapkan teknik dari supervised learning
- Contoh: katalog produk, colors missing
- Perlakukan masalah imputasi sebagai masalah klasifikasi multi-kelas

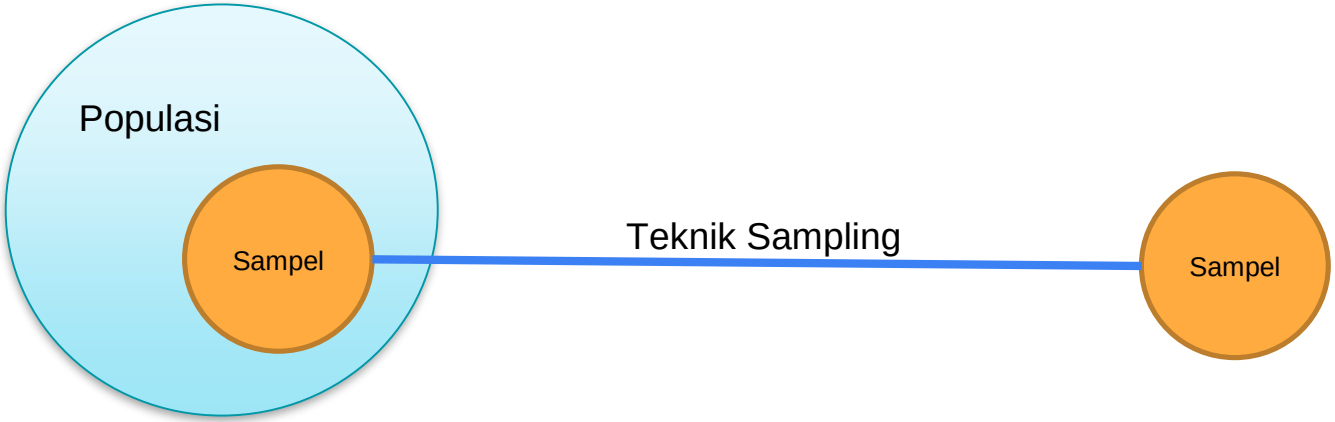
Product Type	Description	Size	Color
Shoe	Ideal for running …	12UK	Black
SDCards	Best SDCard ever …	8GB	Blue
Dress	This **yellow** dress …	M	?

Imputation of attribute `color`



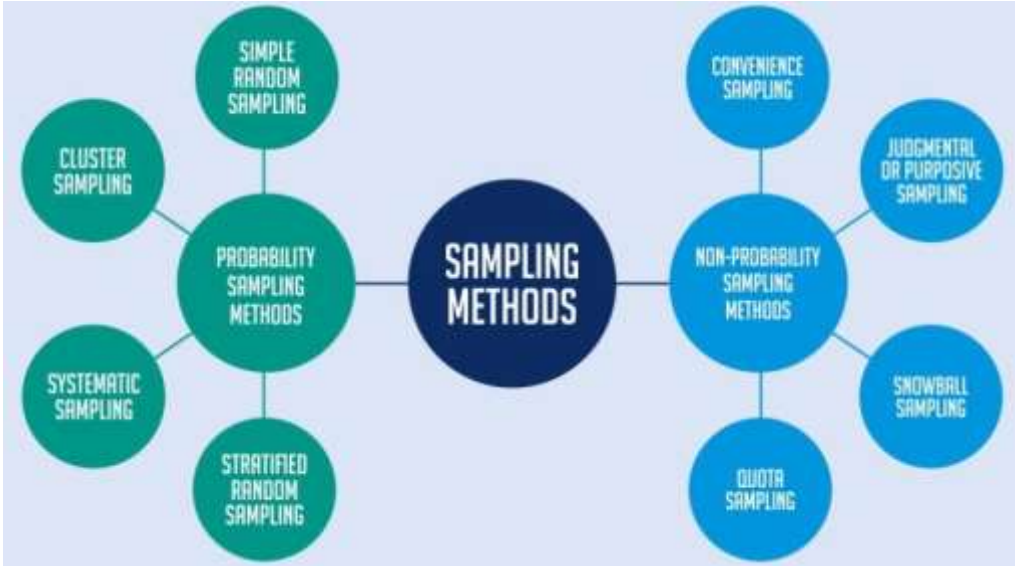
Sampling Data: Pengertian Sampling

- Sebelum melakukan tahapan dalam data preparation, terlebih dahulu adalah pemilihan/penentuan objek yang dapat dilakukan dengan menggunakan penentuan:
  - Populasi
  - Sampel



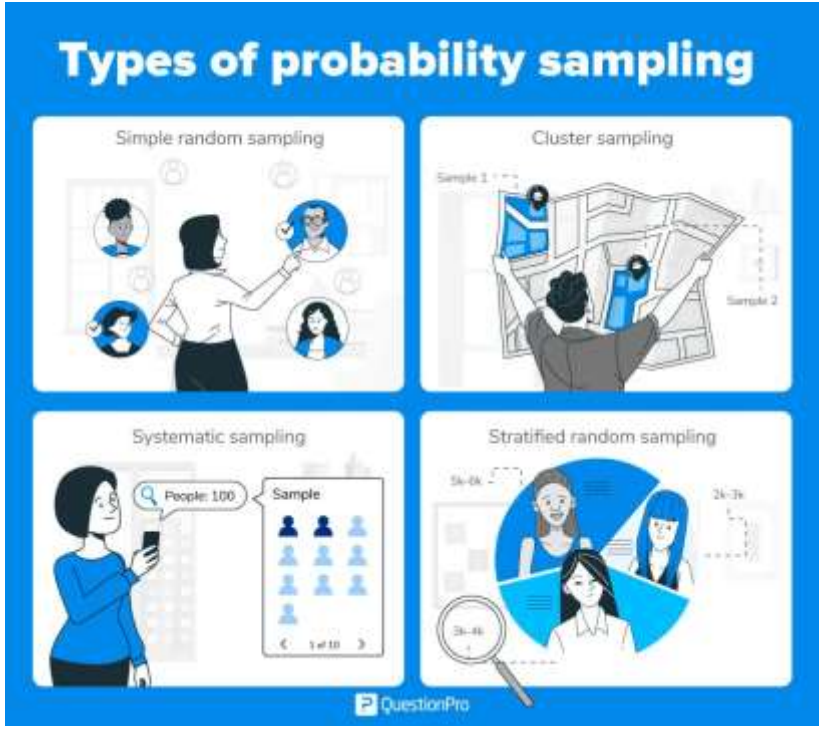
Sampling Data: Metode Sampling

- Kategori Metode Sampling

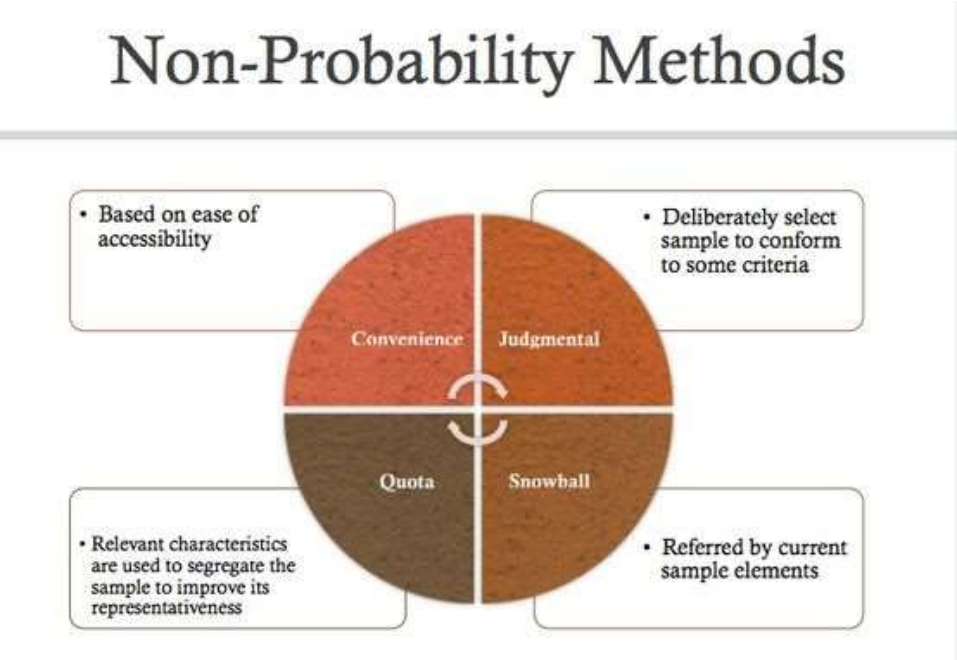


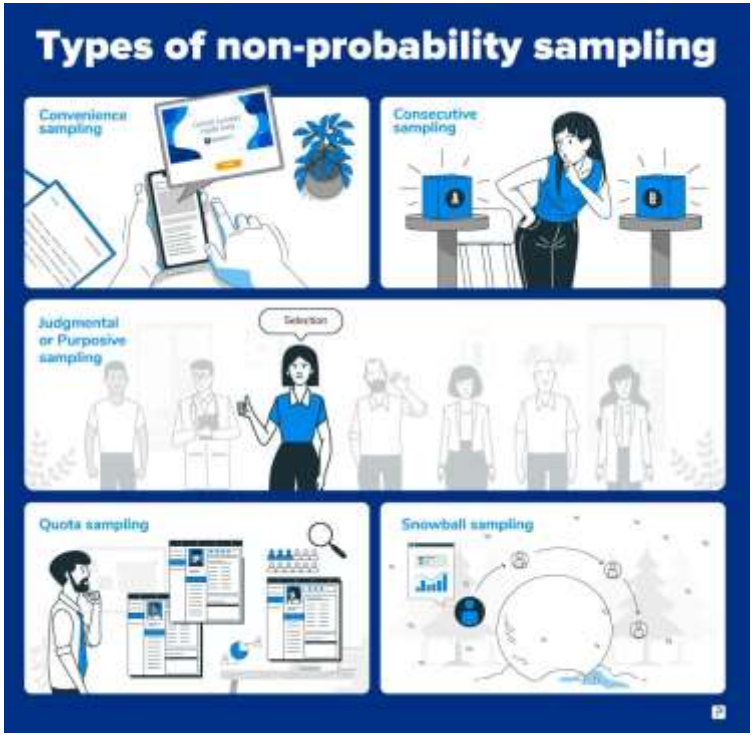
- Probability Sampling:
  - Populasi diketahui
  - Randomisasi/keteracakan: Ya
  - Conclusiver
  - Hasil: Unbiased
  - Kesimpulan: Statistik
- Non-Probability Sampling
  - Populasi tidak diketahui
  - Keterbatasan penelitian
  - Randomisasi/keteracakan: Tidak
  - Exploratory
  - Hasil: Biased
  - Kesimpulan: Analitik

Sampling Data: Metode Sampling

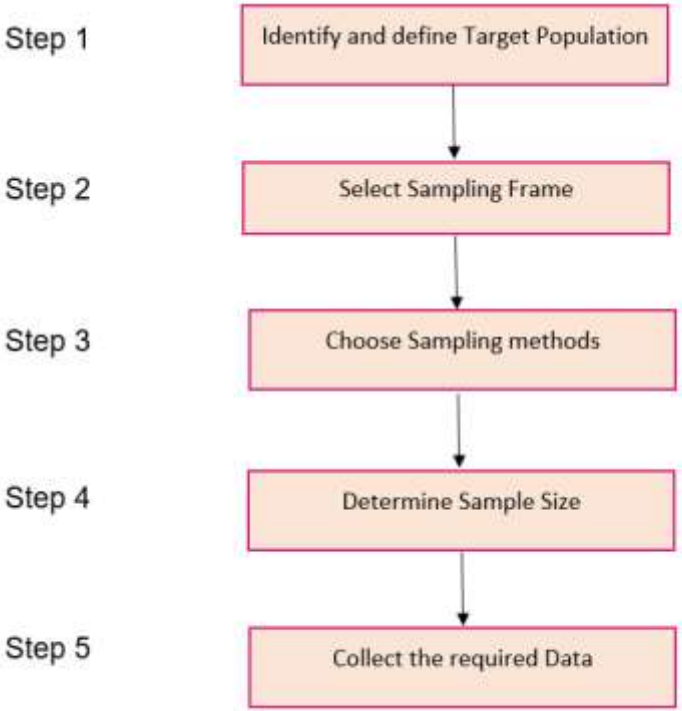


Sampling Data: Teknik Sampling



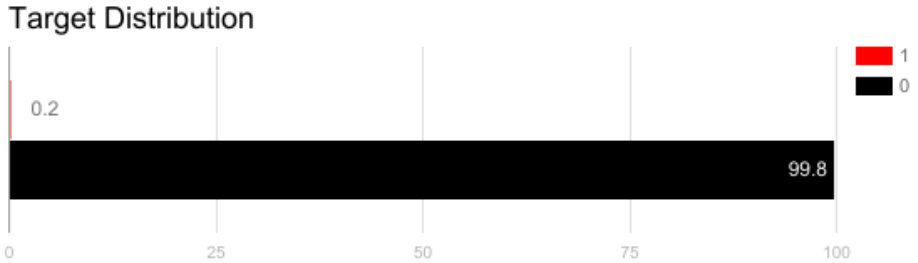


Sampling Data: Tahapan Sampling



Imbalance Dataset: Resampling

- Ini dilakukan setelah proses pemilihan, pembersihan dan rekayasa fitur dilakukan atas pertanyaan:
  - Tanya: apakah kelas target data yang kita inginkan telah secara sama terdistribusi di seluruh dataset?
  - Jawab: Di banyak kasus tidak/belum tentu. Biasanya terjadi imbalance (ketidakseimbangan) antara dua kelas. Misal utk dataset tentang deteksi fraud di perbankan, lelang real-time, atau deteksi intrusi di network! Biasanya data dari dataset tersebut berukuran sangat kecil atau kurang dari 1%, namun sangat signifikan. Kebanyakan algoritma ML tidak bekerja baik utk dataset imbalance tsb.

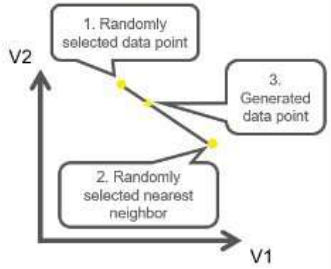
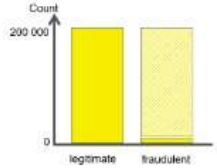


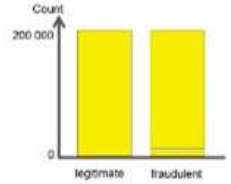
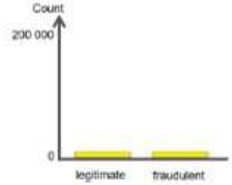
Imbalance Dataset: Resampling

- Berikut adalah bbrp cara utk mengatasi imbalance dataset:
  - Gunakan pengukuran (metrik) yang tepat, misal dengan menggunakan:
    - **Precision/Spesikasi**: berapa banyak instance yang relevan
    - **Recall/Sensitivitas**: berapa banyak instance yang dipilih
    - **F1 score**: harmonisasi mean dari precision dan recall
    - **Matthews correlation coefficient (MCC)**: koefisien korelasi antara klasifikasi biner antara observasi vs prediksi
    - **Area under the ROC curve (AUC)**: relasi antara tingkat true-positive vs false-positive
  - Resample data training, dengan dua metode:
    - **Undersampling**: menyeimbangkan dataset dengan mereduksi ukuran kelas yang melimpah. Dilakukan jika kuantitas data mencukupi
    - **Oversampling**: Kebalikan dari undersampling, dilakukan jika kuantitas data tidak mencukupi

Imbalance Dataset: Resampling

- Teknik Resampling:
  - oversampling (SMOTE)
  - oversampling (Bootstrap)
  - undersampling (Bootstrap)

Resampling method	Description	Target class distribution after resampling
Oversampling (SMOTE)	Generate new synthetic fraudulent transactions until the number of fraudulent transactions is ca. equal to the number of legitimate transactions: 1. Select one of the fraudulent transactions in the training data randomly 2. Select one of its *n* nearest neighbors in the same fraudulent class randomly 3. Select a random point between the existing fraudulent transaction and its nearest neighbor 	• Original data in yellow • New synthetic data in light patterned yellow 

Oversampling (Bootstrap)	Randomly draw with replacement a sample of fraudulent transactions until the number of fraudulent transactions is ca equal to the number of legitimate transactions	
Undersampling (Bootstrap)	Randomly draw with replacement as many legitimate transactions as there are fraudulent transactions	

Membandingkan Model

- *Model yang telah dibangun diharapkan memiliki akurasi yang lebih baik.*
- *Contoh:*

Dua buah, yaitu model 1 dan model 2, dilakukan pengujian terhadap 10 data dengan hasil seperti di samping. Dengan menghitung ratio jumlah percobaan valid terhadap jumlah percobaan diperoleh:

Data Uji	Model 1	Model 2
1	valid	tidak valid
2	tidak valid	tidak valid
3	valid	valid
4	valid	valid
5	tidak valid	tidak valid
6	valid	valid
7	valid	tidak valid
8	tidak valid	tidak valid
9	valid	valid
10	tidak valid	valid

Akurasi model 1 : 6/10= 60%

Akurasi model 2 : 5/10 = 50%

Membandingkan Model

- *Akurasi model yang lebih tinggi belum cukup untuk dapat diklaim bahwa model tersebut **secara statistik** signifikan berbeda (dan lebih baik) dari dari model lainnya.*

- *Untuk mendukung klaim bahwa model 1 lebih baik dari model 2 perlu pengujian secara statistik dengan membuat dua hipotesis yang berlawanan:*

  H_0 *: Kedua model memiliki akurasi yang sama*

  H_1*: Kedua model memiliki akurasi yang berbeda*

- *Pengujian statistik yang sederhana dapat dilakukan dengan McNemar's Test*
- *Untuk pengujian lainnya yang lebih detail (5 cv test dsb.) silakan dilanjutkan ke pengayaan.*

McNemar's Test

- *Data pengujian disusun menjadi tabel contingency sebagai berikut (perhatikan pasangan dalam model 1/model 2)*

	Model 2 valid	Model 2 tidak valid
Model 1 valid	valid/valid	valid/tidak valid
Model 1 tidak valid	tidak valid/valid	tidak valid/tidak valid

- *Sehingga dari tabel sebelumnya diperoleh :*

	Model 2 valid	Model 2 tidak valid
Model 1 valid	4	2
Model 1 tidak valid	1	3

McNemar's Test

- *McNemar's test statistic dihitung dengan*

- $S = (\textbf{valid/tidak valid} - \textbf{tidak valid/valid})^2 / (\textbf{valid / tidak valid} + \textbf{tidak valid/valid})$

- *Hal penting dari S diatas adalah klaim statistik konsen kepada perbedaan valid dan tidak valid pada kedua model, bukan pada akurasi maupun tingkat error*

- *Melalui perhitungan statistik lebih lanjut, perlu memperhatikan masing masing nilai dalam tabel contingency. Distribusi χ^2 mengasumsikan nilai nilai lbesar untuk nilai elemen-elemen tabel contingency. Untuk nilai kecil, digunakan distribusi Binomial. Dalam praktikal, nilai S di atas dilakukan koreksi. Perhitungan detail statistik ini dapat dibaca di referensi.*

Parameter penting dalam McNemar's Test

- *Parameter dalam McNemar's Test, selain s adalah p*
- *Dalam penggunaan praktikal, dapat digunakan perintah (dalam python) untuk mendapatkan dua nilai ini, dengan memperhatikan apakah nilai elemen tabel contingency besar atau kecil*
- *Contoh : dari table contingency sebelumnya, dapat dituliskan:*

  T = [[4,2],
  [1,3]]

- *Untuk case nilai-nilai kecil (misalnya tabel contingency T di atas), dapat digunakan perintah:*

```
s, p = mcnemar(T, exact=True)
```

- *Parameter lain adalah ambang batas p untuk threshold, yaitu α, misalnya $\alpha = 0.05$*

Penolakan / Penerimaan hipotesis

- *Berdasarkan nilai p dan ambang α dapat ditentukan:*

- *Jika $p > \alpha$, hipotesis H_0 gagal untuk ditolak, kedua model secara statistik tidak ada perbedaan*
- *Jika $p \leq \alpha$, hipotesis H_0 ditolak, kedua model scara statistik secara signifikan ada perbedaan*

- *McNemar's adalah pengujian yang sederhana dan telah berkembang diantaranya 5xcv t-test beserta pengembangannya. Detail teori pengujian ini dapat dilihat di referensi.*

Ringkasan

- Kualitas data penting untuk: pengambilan keputusan, kepatuhan terhadap kewajiban hukum, peningkatan kinerja model ML, pengoperasian sistem pemrosesan data
- Data real selalu berantakan dan sulit ditangani, maka proses validasi data menjadi langkah yang penting sebelum digunakan dalam pemodelan.
- Dimensi kualitas data: kelengkapan dan konsistensi data
- Deteksi kesalahan sudah menjadi masalah yang sangat sulit: biasanya memerlukan pembersihan berulang, visualisasi dan *human in the loop*

Referensi

- Sebastian Schelter. *Data Validation and Data Cleaning*. Moore-Sloan Fellow, Center for Data Science, New York University. data-validation-and-data-cleaning.pdf (dataresponsibly.github.io).
- Methodology for data validation 2.0 Revised edition 2018. Methodology for data validation 1.0 (europa.eu)
- F. Biessmann, J. Golebiowski, T. Rukat, D. Lange, P. Schmidt. Automated Data Validation in Machine Learning Systems. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.
- John W. Creswell, Research Design: Pendekatan Kualitatif, Kuantitatif, dan Mixed Edisi Ketiga, diterjemahkan oleh Achmad Fawaid, (Yogyakarta: Pustaka Belajar, 2013).
- Ram Mohan Vadavalasa. Data Validation Process in Machine Learning Pipeline. International Journal for Scientific Research & Development, Vol 8, Issue 4, 2020.

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih