PERTEMUAN 12

Teknik Sampling

Pengertian Sampling

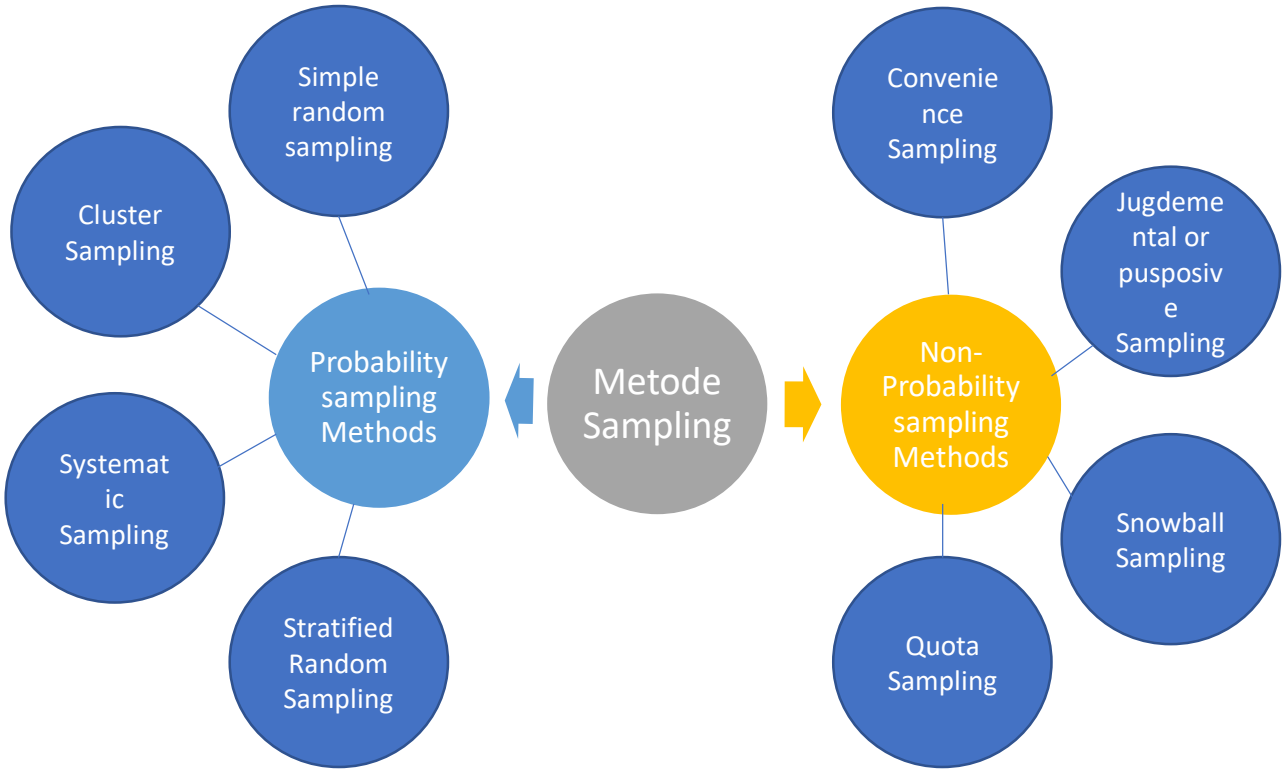
Sebelum melakukan tahapan dalam data preparation, terlebih dahulu adalah pemilihan/penentuan objek yang dapat dilakukan dengan menggunakan

Penentuan:

Populasi

Sampel

Metode Sampling



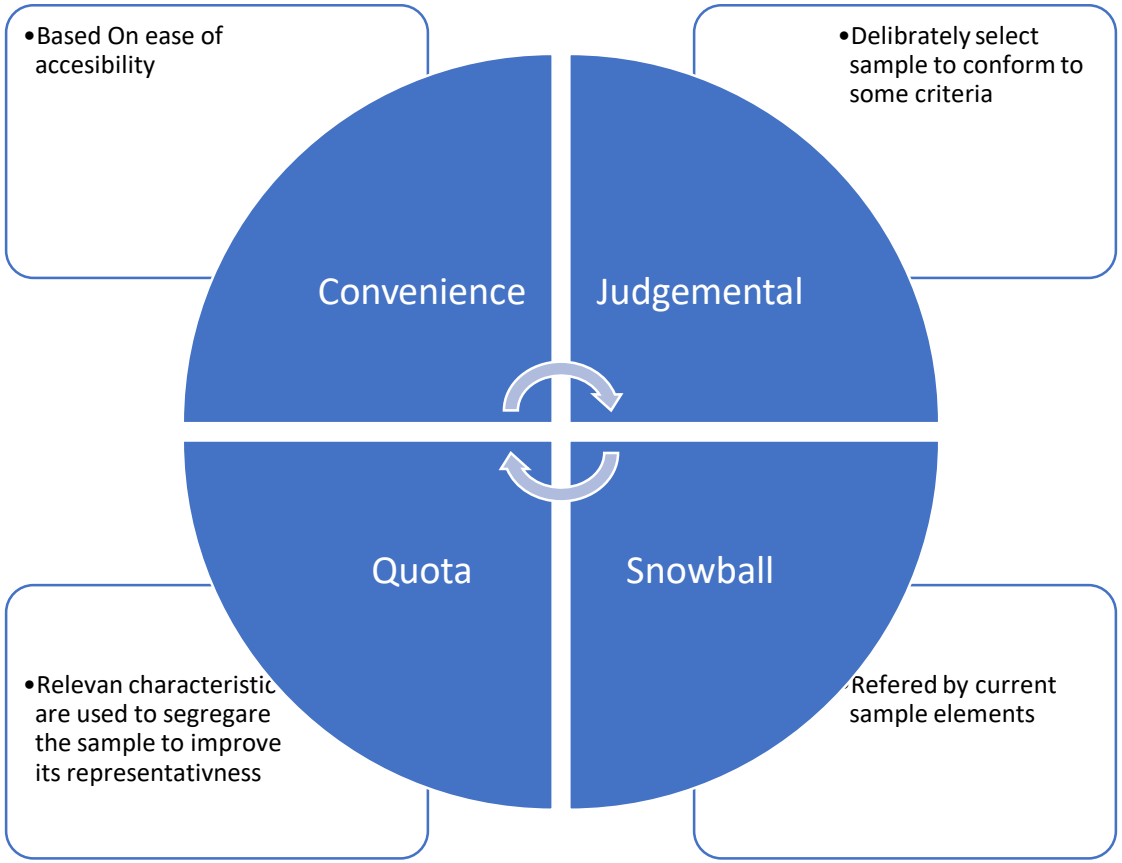
Kategori Metode Sampling

- Probability Sampling:
- Populasi diketahui
- Randomisasi/keteracakan: Ya
- Conclusiver
- Hasil: Unbiased
- Kesimpulan: Statistik

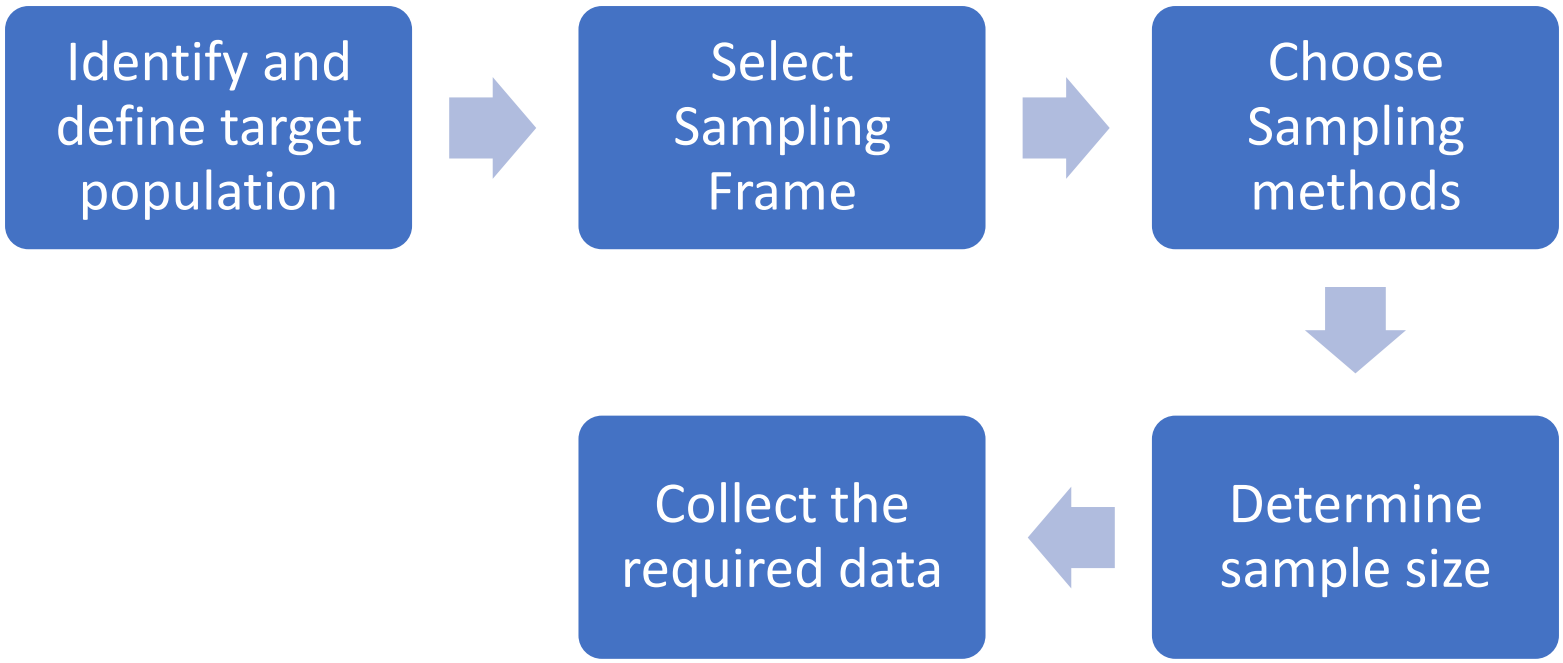
Non-Probability Samplimg

- Populasi tidak diketahui
- Keterbatasan penelitian
- Randomisasi/keteracakan: Tidak
- Exploratory
- Hasil: Biased
- Kesimpulan: Analitik

Non-Probability Methods



Tahapan Sampling



Resampling

Ini dilakukan setelah proses pemilihan, pembersihan dan rekayasa fitur dilakuan 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 detekis 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.

Resampling (Lanjut)

Berikut adalah bbrp cara utk mengatasi imbalance dataset:

Gunakan pengukuran (metrik) yang tepat, misal dengan menggunakan:

- Precision/Spesikasi: berapa banyak instance yang relevan
- Recall/Sensitifitas: berapa banyak instance yang dipilih
- F1 score: harmonisasi mean dari precision dan recall
- MCC: koefisien korelasi antara klasifikasi biner antara observasi vs prediksi
- 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 tidakmencukupi

Pemilihan (Seleksi Fitur) Data

- Setelah menentukan sampling atas data yang akan diambil nanti, selanjutnya adalah melakukan seleksi fitur (feature selection) atas data sampling tsb

- Seleksi fitur merupakan konsep inti dalam ML yang berdampak besar bagi kinerja model prediksi

Pemilihan (seleksi fitur) data

- Fitur data yang tidak/sebagian saja relevan damapak berdampak negatif thdp kinerja model

- Definisi Seleksi Fitur: proses otomatis atau manual memilih fitur data yang paling berkontribusi thdp variabel prediksi atau output yang diinginkan

Manfaat dan jenis seleksi Fitur

Manfaat:

Reduksi Overfitting: semakin kecil data redundant maka keputusan berdasarkan noise semakin berkurang Meningkatkan Akurasi: semakin kecil data misleading maka akurasi model lebih baik

Reduksi Waktu Training: semakin kecil titik data (data point) maka kompleksitas algortima berkurang dan latih algoritma lebih cepat

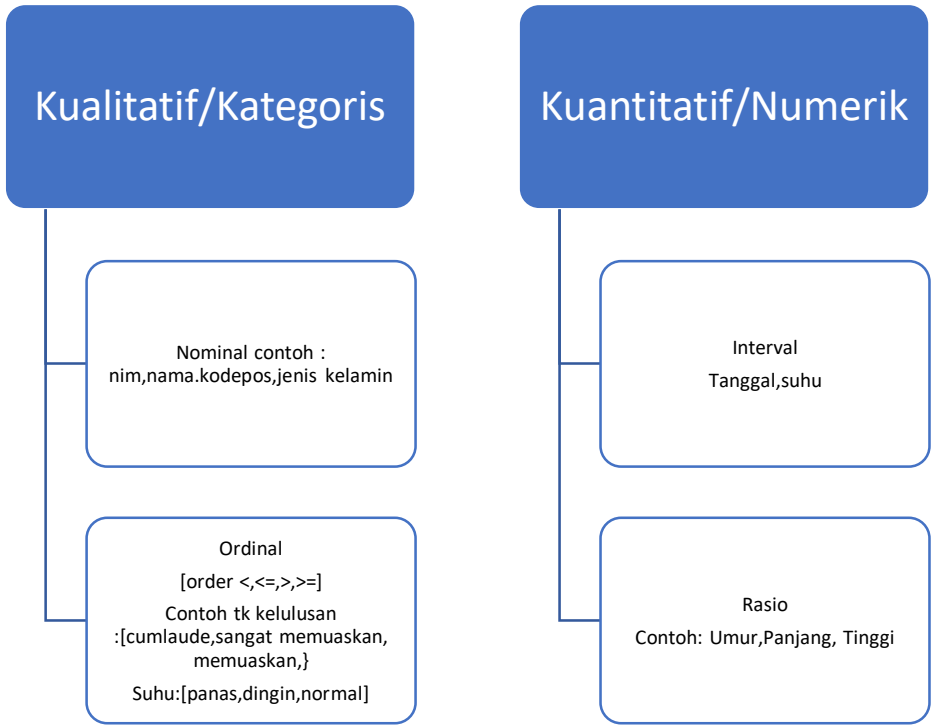
Jenis:

Unsupervised: metode yang mengabaikan variabel target, seperti menghapus variabel yang berlebihan menggunakan korelasi

Supervised: metode yang menggunakan variabel target, seperti menghapus variabel yang tidak relevan

Membedakan Jenis Data

Numerik vs Kategorik



Validasi data

Verifikasi vs Validasi

- Verifikasi: Benar vs Salah
- Validasi: Kuat vs Lemah

Validasi merupakan tahapan kritikal yang sering diabaikan DS-tist 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.”)
- Nilai Null/Missing Values
- Misspelling/Type
- Invalid Data (gender: L/P: L; Laki-laki; P: Pria/Perempuan?)

Teknik validasi data dan model

Teknik Validasi Data dan Model:

- Akurasi
- Kelengkapan
- Konsistensi
- Ketepatan Waktu
- Kepercayaan
- Nilai Tambah
- Penafsiran
- Kemudahan Akses