Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

PERTEMUAN KE: 2

PENGUMPULAN DATA DAN PREPROCESSING

KULIAH

OLEH: NURJOKO

[12-09-2023] | [KODE MK: SD230411 | MATAKULIAH: MATEMATIKA | SKS: 4 | VERSI: 01

Learning Objectives

- Memahami Proses Pengumpulan Data
- Mengenal Jenis-jenis Data
- Mengenal Sumber Data
- Mengenal Teknik Pengumpulan Data
- Memahami Proses Preprocessing Data

Konsep Pengumpulan Data

Pengumpulan data adalah salah satu tahap penting dalam proses sains data yang melibatkan pengumpulan informasi atau pengambilan sampel dari berbagai sumber untuk analisis.

Beberapa konsep penting dalam pengumpulan data:

Beberapa Konsep Penting Dalam Pengumpulan Data:

1. Tujuan Pengumpulan Data

Mengapa Anda mengumpulkan data dan apa yang akan Anda lakukan dengan data tersebut

Ex: Evaluasi Kepuasan Pelanggan: Tujuan pengumpulan data adalah untuk mengukur tingkat kepuasan pelanggan terhadap produk atau layanan perusahaan. Data yang dikumpulkan melalui survei pelanggan dapat digunakan untuk mengidentifikasi area yang perlu diperbaiki atau ditingkatkan.

2. Desain Pengumpulan Data

Proses perencanaan pengumpulan data dimulai dengan merancang metode dan alat pengumpulan yang sesuai dengan tujuan yang mencakup pemilihan jenis data yang akan dikumpulkan (misalnya, data kuantitatif atau kualitatif), teknik sampling, dan instrumen pengumpulan data.

3. Sumber Data

Mengidentifikasi sumber data yang dapat digunakan untuk mengumpulkan informasi yang diperlukan. Sumber data dapat berupa survei, database, wawancara, observasi, atau bahkan data yang sudah ada.

4. **Metode Pengumpulan Data:** Metode pengumpulan data mencakup cara konkret untuk mendapatkan informasi dari sumber data. Ini bisa berupa survei online, wawancara langsung, pengukuran fisik, observasi lapangan, atau pengumpulan data dari sumber elektronik.
5. **Validitas dan Keandalan:** Data yang dikumpulkan harus valid (mengukur apa yang dimaksud) dan memiliki keandalan (konsisten dan dapat diandalkan). Dalam pengumpulan data, penting untuk mengurangi bias dan kesalahan sebanyak mungkin.

6. Etika Pengumpulan Data

Ketika mengumpulkan data dari individu atau organisasi, penting untuk mematuhi prinsip-prinsip etika, termasuk privasi data, izin partisipasi, dan perlindungan terhadap pengungkapan informasi pribadi.

7. Dokumentasi

Setiap tahap dalam pengumpulan data harus didokumentasikan dengan baik meliputi instruksi pengumpulan, catatan data yang diperoleh, dan catatan lainnya yang relevan.

- Data menurut sifatnya

- **Data kualitatif adalah data yang tidak berbentuk angka.**
- **Data kuantitatif adalah data yang dinyatakan dalam bentuk angka.**

- Data menurut sumbernya

- **Data Internal** adalah data yang bersumber dari keadaan atau kegiatan suatu kelompok atau organisasi.
- **Data Eksternal** adalah data yang bersumber dari luar kelompok atau organisasi.

- Data menurut cara memperolehnya

- **Data Primer** adalah data yang dikumpulkan dan diolah sendiri oleh suatu organisasi atau perorangan langsung dari obyeknya.
- **Data Sekunder** adalah data yang diperoleh dalam bentuk jadi dan telah diolah oleh pihak lain, biasanya dalam bentuk publikasi.

- Data menurut waktu pengumpulannya

- **Data cross section** adalah data yang dikumpulkan dalam suatu periode tertentu.
- **Data berkala (time series**) adalah data yang dikumpulkan dari waktu ke waktu. Data ini sering juga disebut sebagai data historis.

Teknik Pengumpulan Data

Data diperoleh melalui **pengamatan** dan juga **pengukuran.**

1.Berdasarkan Jenis Cara Pengumpulannya

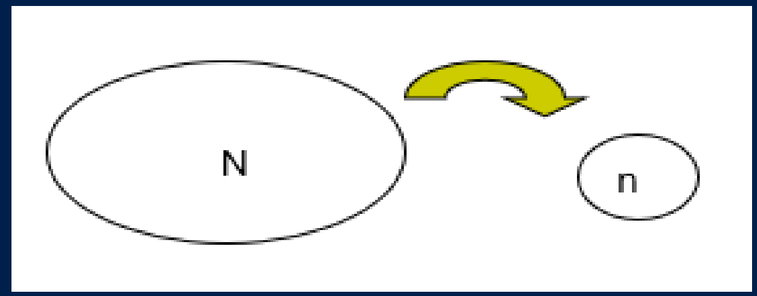
a. Pengamatan (Observasi)
b. Penelusuran Literatur
c. Penggunaan Kuesioner (angket)
d. Wawancara (interview)

2.Berdasarkan Banyaknya Data yang diambil

a. Sensus
b. Sampling

Populasi dan Sampel

- **Populasi**
  Keseluruhan objek yang ada dalam ruang lingkup yang diteliti (N)
- **Sampel**
  Bagian dari populasi (n)
  Dengan sampel diharapkan karakteristik dari populasi dapat diketahui



Teknik Sampling

Teknik Sampling

Probability Sampling

1. Simple random sampling
2. Proportionate stratified random sampling
3. Disproportiate strafied random sampling
4. Area (cluster) sampling

Non Probability Sampling

1. Sampling sistematis
2. Sampling kuota
3. Sampling aksidental
4. Porposive sampling
5. Sampling jenuh
6. Snowball sampling

Probability Sampling

a. **Sampling acak sederhana (Simple Random Sampling)**

   Pengambilan smpel dari populasi dilakukan secara acak tanpa memperhatikan strata yang ada dalam populasi.

b. **Sampling acak bertingkat proporsional (Proportionate Stratified Random Sampling)**

   Populasi memiliki anggota yang tidak homogen dan berstrata secara proprsioanal.

c. **Sampling acak bertingkat tidak proporsional (Dispoprtianate Stratified Random Sampling)**

   Populasi berstrata tetapi tidak proporsional

d. **Area Sampling (Cluster Sampling)**

   Obyek yang diteliti sangat luas misal penduduk suatu negara, propinsi atau kabupaten.

Nonprobability Sampling

a. **Sampling Sistematis**

   Penentuan sampling berdasarkan urutan dari anggota populasi yang telah diberi nomor urut

b. **Sampling Kuota**

   Sampel dari populasi yang mempunyai ciri-ciri tertentu sampai jumlah (kuota) yang diingiinakan.

c. **Sampling Aksidental**

   Teknik sampling berdasarkan kebetulan saja misalkan orang yang lewat dan dipandang cocok sebagai sumber data

Nonprobability Sampling

d. Sampling Purposive

Teknik penentuan sampel untuk pertimbangan tertentu

e. Sampling Jenuh

Semua anggota populasi digunakan sebagai sampel,biasanya kurang dari 30 anggota

f. Snowball Sampling

Penentuan sampel mula-mula kecil kemudian sampel tersebut disuruh memilih anggota lain untuk dijadikan sampel, begitu selanjutnya.

Pembersihan Data

Pembersihan data (data cleaning) adalah tahap penting dalam analisis data yang bertujuan untuk mengidentifikasi, memperbaiki, atau menghapus kesalahan, inkonsistensi, atau anomali dalam data.

Berikut adalah beberapa langkah yang umumnya dilakukan dalam pembersihan data:

Pembersihan Data

1.Identifikasi dan Penanganan Duplikasi:

Identifikasi dan penanganan data ganda atau duplikasi adalah langkah pertama. Data yang sama terduplikasi dalam dataset harus diidentifikasi dan dihapus atau digabungkan agar hanya ada satu entri.

2.Penanganan Nilai yang Hilang (Missing Values):

Identifikasi dan penanganan nilai yang hilang adalah langkah kunci. Data yang tidak lengkap dapat merusak analisis. Anda dapat memilih untuk mengisi nilai yang hilang dengan nilai rerata, median, atau metode lain yang sesuai, atau Anda dapat memutuskan untuk menghapus baris atau kolom yang mengandung nilai yang hilang terlalu banyak.

Pembersihan Data

3. Penyelarasan Format Data:

Pastikan bahwa semua data memiliki format yang seragam. Misalnya, pastikan tanggal dalam format yang benar, angka dalam format yang benar (desimal atau bulat), dan teks atau kategori dalam format yang konsisten.

4. Penanganan Outlier:

Outlier adalah nilai yang jauh berbeda dari sebagian besar nilai dalam dataset. Anda dapat memutuskan untuk menghapus outlier jika mereka mengganggu analisis atau memilih untuk mempertahankannya jika mereka memiliki arti statistik atau bisnis yang sah.

Teknik Transformasi Data

Teknik transformasi data adalah serangkaian metode yang digunakan untuk mengubah, menggabungkan, atau mengolah data menjadi bentuk yang lebih sesuai atau lebih bermanfaat untuk analisis atau pemodelan.

Beberapa pengertian

1. Transformasi → proses konversi data ke dalam skala baru agar memenuhi homogenitas ragam dan sebaran data menjadi normal.
2. Data yang perlu ditransformasi adalah data yang akan dianalisis varian, namun data tersebut tidak memenuhi syarat untuk dilakukan analisis.
3. Data hasil hitungan umumnya termasuk salah satu contoh data yang tidak memenuhi syarat untuk dianalisis.
4. Data hasil pengukuran adalah data dapat langsung dianalisis varian

Transformasi Logaritmik:

- Transformasi logaritmik (log transformation) sering digunakan untuk mengubah data yang memiliki distribusi miring (skewed) ke kanan menjadi lebih simetris. Ini berguna ketika perbedaan antara nilai-nilai rendah dan tinggi sangat besar.
- Contohnya adalah harga saham atau penghasilan.
- Rumus transformasi logaritmik: log(x) atau log10(x).

Transformasi Akar Kuadrat:

- Transformasi akar kuadrat (square root transformation) digunakan untuk mengurangi perbedaan antara nilai-nilai tinggi dan rendah dalam data. Ini dapat berguna ketika data memiliki skew positif.
- Rumus transformasi akar kuadrat: sqrt(x).

Transformasi Box-Cox:

- Transformasi Box-Cox adalah transformasi yang lebih umum yang dapat digunakan untuk mengubah data menjadi distribusi normal. Ini berguna ketika asumsi normalitas dibutuhkan untuk analisis statistik.

Kesimpulan

- Pengumpulan data yang akurat dan preprocessing yang baik adalah kunci untuk menghasilkan hasil analisis yang andal dan bermakna.
- Data yang tidak diproses dengan baik dapat mengarah pada kesalahan interpretasi dan kesimpulan yang keliru.
- Memahami karakteristik data, merencanakan dengan baik, dan memilih teknik preprocessing yang sesuai adalah langkah-langkah penting dalam pengambilan keputusan berdasarkan data yang kuat.
- Ilmu data yang baik melibatkan proses berulang dari pengumpulan data, preprocessing, analisis, hingga interpretasi, dan pengambilan keputusan yang berkelanjutan.

REFERENCES

Fill in IEEE Style

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”