

# Panduan Praktikum Eksplorasi Data

Durasi: 3 × 110 menit

## Capaian Pembelajaran Mata Kuliah

Setelah menyelesaikan praktikum ini diharapkan mahasiswa mampu menjelaskan tahapan data mining, karakteristik data, eksplorasi data, dan praproses data serta penerapannya pada suatu permasalahan.

## Tools

- Google Colab
- Jupyter Notebook
- PyCharm
- Spyder
- Python IDE yang lain

## Materi Praktikum

Eksplorasi data dilakukan sebagai langkah awal untuk memahami data dari sebelum dilakukan praproses. Tahapan ini bertujuan untuk menyeleksi teknik pemrosesan dan analisis data yang sesuai dengan dataset yang dimiliki. Dalam eksplorasi data, hal yang harus diperhatikan yaitu:

- Tipe data
- Kualitas data
- Statistika ringkasan
- Visualisasi

### 1. Tipe Data

Sebuah data set dapat dipandang sebagai sebuah koleksi dari objek-objek data. Nama lain dari sebuah objek data adalah record, titik, vector, pola, event, case, sample, observasi atau entitas. Objek digambarkan dengan sejumlah atribut yang menerangkan sifat atau karakteristik dari objek tersebut. Kumpulan dari objek data dan atributnya.

Skala pengukuran adalah aturan (fungsi) yang menghubungkan nilai numerik atau simbolik dengan sebuah atribut dari sebuah objek. Proses pengukuran adalah penggunaan skala pengukuran untuk menghubungkan sebuah nilai dengan sebuah atribut tertentu dari sebuah objek.

Atribut juga sering disebut variabel, field, fitur, atau dimensi. Atribut adalah sifat/properti/karakteristik objek yang nilainya bisa bermacam-macam dari satu objek dengan objek lainnya, dari satu waktu ke waktu yang lainnya. Atribut adalah sebuah sifat atau karakteristik dari sebuah objek yang dapat bervariasi, baik dari satu objek ke objek lain atau dari satu waktu ke waktu yang lain. Sebuah atribut adalah sifat atau karakteristik dari sebuah objek.

## 2. Kualitas Data

Data Mining adalah proses menemukan pengetahuan dari sekian banyak data yang tersimpan dalam database; data warehouse; atau repositori lain. Usaha yang diperlukan pada pengolahan data yaitu data understanding sebesar 20%, 60% untuk data preparation, dan hanya 20% untuk data mining serta menganalisis knowledge. Ini menunjukkan bahwa data preparation membutuhkan usaha terbesar dalam pengolahan data. Data yang ada di dunia ini cenderung tidak lengkap dikarenakan adanya beberapa missing value. Hal ini bisa saja terjadi karena valuenya tidak relevan pada masalah, tidak terekam saat pengumpulan data, atau memang tidak dijawab oleh responden dikarenakan alasan privasi. Apabila rata-rata missing value kurang dari 1 %, data yang missing ini tidak akan menimbulkan masalah untuk proses Knowledge Discovery in Database (KDD), 1-5% masih bias diolah, 5-15% dibutuhkan metode untuk menanganinya dan jika lebih dari 15% dapat menimbulkan interpretasi yang berbeda.

Contoh dari masalah kualitas data:

- a) Noise dan outlier
- b) Missing Value
- c) Duplicate data

## 3. Statistika Ringkasan

Nilai Tengah : dari sekumpulan data (distribusi), ada beberapa harga/nilai yang dapat kita anggap sebagai wakil dari kelompok data tersebut. Nilai-nilai yang biasa digunakan untuk mewakili data tersebut adalah mean, median, dan modus. nilai-nilai tersebut disebut sebagai nilai tengah.

## 4. Visualisasi

Visualisasi adalah konversi dari data menjadi sebuah format visual atau tabular sehingga karakteristik data dan hubungan antar data atau atribut dapat dianalisis. Visualisasi dari data adalah salah satu teknik yang tepat untuk eksplorasi data. Dapat mendeteksi pola umum dan trend data. Dapat mendeteksi outlier dan pola yang tidak biasa. Informasi berupa data numerik memang diperlukan, tetapi terkadang sulit memahami atau menggambarkannya. Sehingga, seringkali visualisasi data juga dilakukan untuk lebih mudah memahami data.

Baik histogram dan stem-and-leaf plots berguna untuk memberikan gambaran ukuran tendensi sentral dan kesimetrisan data pengamatan. Penyajian grafis lainnya yang bisa merangkum informasi lebih detail mengenai distribusi nilai-nilai data pengamatan adalah Box and Whisker Plots atau lebih sering disebut dengan BoxPlot atau Box-Plot (kotak-plot) saja. Seperti namanya, Box and Whisker, bentuknya terdiri dari Box (kotak) dan whisker.

## Panduan Praktikum

<https://github.com/alদিনata/Modul-Praktikum-Data-Mining/blob/main/Materi/1%20-%20Eksplorasi%20Data.ipynb>