Panduan Praktikum
Praproses Data

Durasi: 3 × 110 menit

Capaian Pembelajaran Mata Kuliah

Setelah menyelesaikakn 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

- Pengertian Praproses Data

Praproses data adalah tahap penting dalam data mining yang bertujuan untuk mempersiapkan data mentah menjadi format yang sesuai untuk analisis lebih lanjut. Langkah ini mencakup pembersihan data, transformasi, integrasi, dan reduksi data.

- Tujuan Praproses Data

1. Membersihkan data dari kesalahan, duplikasi, atau nilai yang hilang.
2. Meningkatkan kualitas data untuk meningkatkan kinerja algoritma.
3. Menstandarkan format data agar seragam.
4. Mengurangi dimensi data untuk efisiensi proses analisis.

- Tahapan Praproses Data

1. Pembersihan Data (Data Cleaning)

- Tujuan: Mengatasi data yang tidak lengkap, salah, atau tidak konsisten.

- Teknik:

  - Mengisi nilai yang hilang dengan:

    - Nilai rata-rata, median, atau modus.

    - Algoritma prediktif.

  - Menghapus duplikasi data.

  - Menstandarkan nilai yang tidak konsisten (contoh: "Y" dan "Yes" menjadi satu format).

2. Integrasi Data (Data Integration)

- Tujuan: Menggabungkan data dari berbagai sumber menjadi satu kesatuan yang kohesif.

- Teknik:

  - Resolusi konflik skema (contoh: menyamakan satuan seperti "kg" dan "g").

  - Penyatuan data yang serupa dari tabel atau database berbeda.

3. Transformasi Data (Data Transformation)

- Tujuan: Mengubah data ke dalam format yang sesuai untuk analisis.

- Teknik:

  - Normalisasi: Mengubah nilai ke rentang tertentu, misalnya 0-1.

  - Standardisasi: Mengubah data agar memiliki rata-rata 0 dan standar deviasi 1.

  - Encoding: Mengonversi data kategoris menjadi numerik, seperti one-hot encoding atau label encoding.

4. Reduksi Data (Data Reduction)

- Tujuan: Mengurangi dimensi data tanpa kehilangan informasi penting.

- Teknik:

  - Seleksi atribut (memilih atribut yang relevan).

  - Ekstraksi fitur (menggabungkan beberapa atribut menjadi satu).

  - Sampling (mengambil subset data yang representatif).

- Permasalahan yang Umum Ditemui dalam Praproses Data

1. Data yang Tidak Lengkap:

- Penyebab: Kesalahan input, data tidak tersedia.

- Solusi: Mengisi nilai yang hilang atau menghapus entri yang tidak lengkap.

2. Data yang Tidak Konsisten:

- Penyebab: Perbedaan format atau konflik antar sumber data.

- Solusi: Standarisasi format data.

3. Dimensi Data yang Terlalu Tinggi:

- Penyebab: Terlalu banyak atribut yang tidak relevan.

- Solusi: Menggunakan seleksi atribut atau algoritma reduksi dimensi seperti PCA (Principal Component Analysis).

- Pentingnya Praproses Data

1. Data yang tidak diproses dengan baik dapat menghasilkan analisis yang salah.

2. Memastikan bahwa data memiliki kualitas tinggi untuk menghasilkan model yang akurat.

3. Mengurangi waktu dan biaya pemrosesan data.

- Studi Kasus Sederhana

- Dataset: Informasi transaksi toko online.

- Langkah Praproses:

1. Menghapus entri duplikat pada data pelanggan.

2. Mengisi nilai hilang pada atribut "kategori produk" menggunakan nilai modus.

3. Menormalisasi atribut "harga" ke dalam rentang 0-1.

4. Menggabungkan data transaksi dari cabang toko yang berbeda.

Panduan Praktikum

Download lalu pelajari dan running code pada link berikut:
https://github.com/aldinata/Modul-Praktikum-Data-Mining/blob/main/Materi/2%20-%20Praproses%20Data.ipynb