KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 11:
Data Preprocessing

Sub CPMK

Mahasiswa mampu melakukan *data preprocessing* dengan benar.

Pokok Bahasan

- Kualitas data
- *Data preprocessing* (Prapemrosesan data)
- *Data cleaning* (Pembersihan data)
- *Data transformation* (Transformasi data)

Kualitas Data

Beberapa faktor yang mendefinisikan kualitas data :

- akurasi,
- kelengkapan,
- konsistensi,
- ketepatan waktu,
- kepercayaan (*believability*),
- kemampuan interpretasi (*interpretability*).

Kualitas Data: Akurasi

- Data Tidak Akurat adalah data mengandung kesalahan nilai atribut atau nilai menyimpang dari yang diharapkan.
- Ada banyak kemungkinan alasan untuk data yang tidak akurat:
  - Instrumen pengumpulan data yang digunakan mungkin salah.
  - Ada kesalahan manusia atau komputer yang terjadi pada entri data.
  - Pengguna mungkin dengan sengaja mengirimkan nilai data yang salah karena tidak ingin mengirimkan informasi pribadi.
  - Kesalahan dalam pengiriman data juga dapat terjadi karena ada batasan teknologi.
  - Ketidakkonsistenan dalam konvensi penamaan atau kode data, atau format yang tidak konsisten untuk bidang input (contoh: tanggal).
  - Tuple duplikat.

Kualitas Data: Kelengkapan

- Alasan data yang tidak lengkap :
  - Atribut yang menarik mungkin tidak selalu tersedia, seperti informasi pelanggan untuk data transaksi penjualan.
  - Data tidak dianggap penting pada saat entri (kesalahpahaman)
  - Malfungsi peralatan sehingga data tidak terekam
  - Data hanya berisi data agregat.

Kualitas Data: Konsistensi

- Data yang tidak konsisten, misalnya data yang memiliki perbedaan dalam kode yang digunakan untuk mengkategorikan item.

Kualitas Data:
Ketepatan Waktu

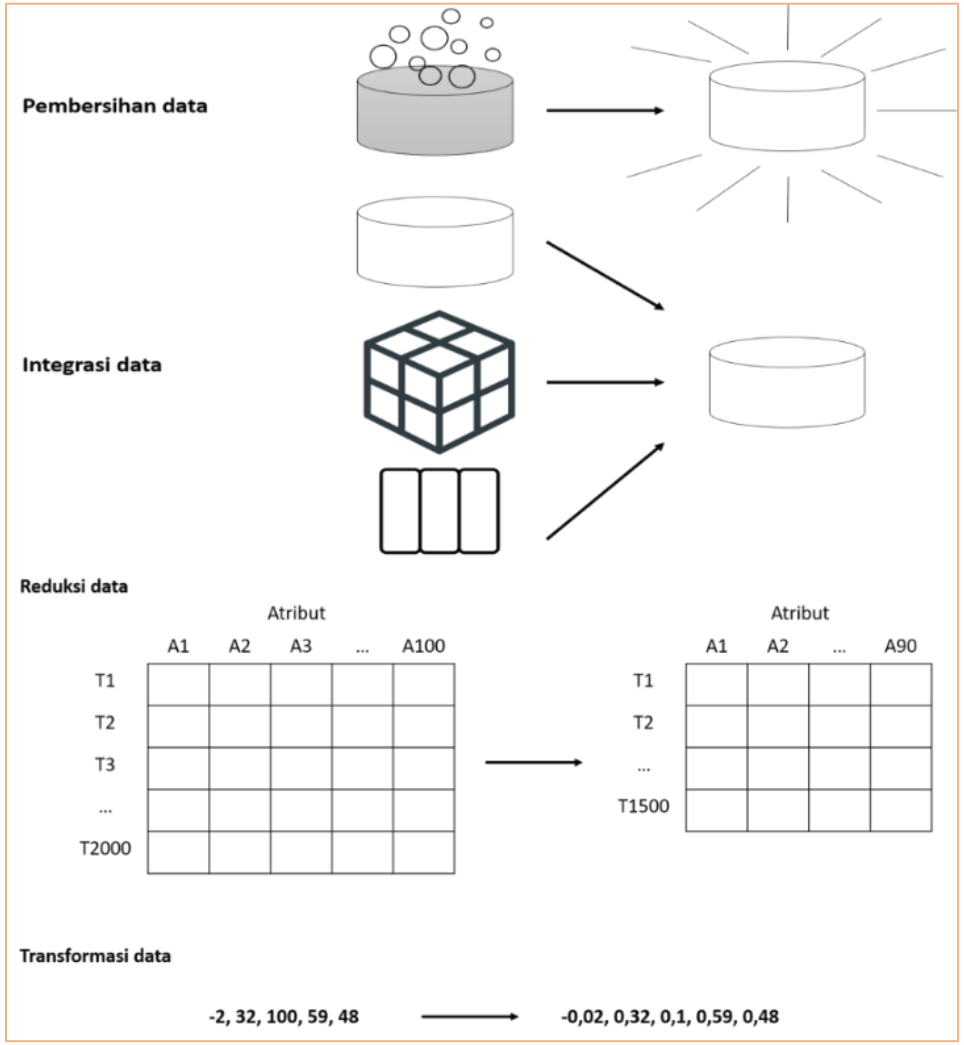
- Ketepatan waktu juga mempengaruhi kualitas data.
- Misalkan beberapa perwakilan divisi penjualan gagal menyerahkan catatan penjualan mereka tepat waktu pada akhir bulan. Dengan begitu, data yang disimpan dalam *database* tidak lengkap. Namun, setelah semua data diterima, data sudah benar.
- Jadi, data yang tidak diperbarui secara tepat waktu berdampak negatif pada kualitas data.

Kualitas Data:
Kepercayaan dan Kemampuan Interpretasi

- Kepercayaan mencerminkan seberapa besar data dipercaya oleh pengguna, sedangkan kemampuan interpretasi mencerminkan seberapa mudah data tersebut dipahami.
- Misalkan sebuah data, memiliki beberapa kesalahan pada masa lalu (tetapi sekarang sudah diperbaiki), pengguna tidak lagi mempercayainya. Selain itu, data memuat kode yang tidak dapat diinterpretasikan oleh departemen lain.
- Jadi, meski data sekarang akurat, lengkap, konsisten, dan tepat waktu, pengguna mungkin menganggapnya berkualitas rendah karena kepercayaan dan interpretasi yang buruk.

Data Preprocessing

- Data dunia nyata cenderung kotor, tidak lengkap, dan tidak konsisten.
- Teknik prapemrosesan data dapat meningkatkan kualitas data sehingga membantu meningkatkan akurasi dan efisiensi proses penambangan berikutnya.
- Prapemrosesan data merupakan langkah penting dalam proses penemuan pengetahuan karena keputusan yang berkualitas harus didasarkan pada data yang berkualitas.
- Mendeteksi anomali data, memperbaikinya lebih awal, dan mengurangi data yang akan dianalisis dapat menghasilkan keuntungan besar untuk pengambilan keputusan.



Prapemrosesan Data (*Data Preprocessing*)

Langkah-langkah utama yang terlibat dalam prapemrosesan data :

Pembersihan data

- membersihkan data dengan mengisi nilai yang hilang, menghaluskan *noisy data*, mengidentifikasi atau menghapus outlier, dan menyelesaikan inkonsistensi.

Integrasi data

- Memasukkan data dari berbagai sumber, melibatkan integrasi beberapa database, kubus data, atau file.
- Pembersihan data tambahan dapat dilakukan untuk mendeteksi dan menghapus redundansi yang mungkin dihasilkan dari integrasi data.

Reduksi data

- Memperoleh representasi tereduksi dari kumpulan data yang volumenya jauh lebih kecil, tetapi menghasilkan analisis yang sama (atau hampir sama).
- Strategi reduksi data termasuk pengurangan dimensi dan pengurangan jumlah.

Transformasi data

- Nilai data mentah untuk atribut diganti dengan rentang atau tingkat konseptual yang lebih tinggi, misalnya dengan normalisasi, diskritisasi data, dan pembuatan hierarki konsep.

Pembersihan Data

- Analis atau peneliti harus mencari anomali, memverifikasi data dengan pengetahuan domain, dan memutuskan pendekatan yang tepat untuk membersihkan data.
- Pembersihan data rutin mencoba untuk mengisi nilai yang hilang, menghaluskan *noise* saat mengidentifikasi outlier, dan memperbaiki inkonsistensi dalam data.
- Ada 2 jenis data yang hilang :
  - MCAR (*missing completely at random*) adalah skenario yang diinginkan jika ada data yang hilang.
  - MNAR (*missing not at random*) adalah masalah yang lebih serius. Jika terjadi MNAR, sebaiknya memeriksa proses pengumpulan data lebih lanjut dan mencoba memahami mengapa informasi tersebut hilang.

Pembersihan Data: data hilang

Cara mengisi nilai yang hilang (*missing value*):

1. Abaikan Tuple: Ini biasanya dilakukan ketika label kelas hilang (untuk klasifikasi).
2. Isi nilai yang hilang secara manual.
3. Gunakan konstanta global untuk mengisi nilai yang hilang: Ganti semua nilai atribut yang hilang dengan konstanta yang sama, misalnya Tidak Diketahui.
4. Gunakan ukuran pemusatan untuk atribut untuk mengisi nilai yang hilang. Untuk distribusi data normal atau simetris, nilai rata-rata (mean) dapat digunakan, sedangkan distribusi data miring (tidak simetris) harus menggunakan median.
5. Gunakan mean atau median untuk semua sampel yang termasuk dalam kelas yang sama dengan tupel yang diberikan.
6. Gunakan nilai yang paling mungkin untuk mengisi nilai yang hilang, misalnya dengan regresi, alat berbasis inferensi menggunakan formula Bayesian, atau induksi pohon keputusan. Metode ini menggunakan sebagian besar informasi dari data saat ini untuk memprediksi nilai yang hilang dibandingkan metode lainnya.

Pembersihan Data: penghalusan data

1. **Binning:** Metode binning menghaluskan nilai data yang diurutkan dengan berkonsultasi dengan nilai di sekitarnya. Nilai yang diurutkan didistribusikan ke sejumlah bin.

Data berurutan: 4, 6, 11, 22, 24, 26, 27, 29, 36

Partisi ke dalam bin (frekuensi-sama):
Bin 1: 4, 6, 11
Bin 2: 22, 24, 26
Bin 3: 27, 30, 36

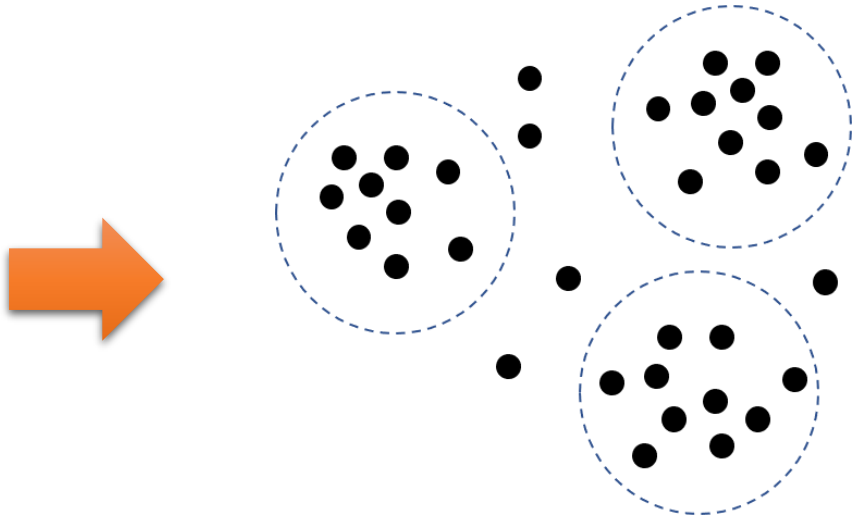
Smoothing dengan rata-rata bin:
Bin 1: 7, 7, 7
Bin 2: 24, 24, 24
Bin 3: 31, 31, 31

Smoothing dengan batasan bin:
Bin 1: 4, 4, 11
Bin 2: 22, 22, 26
Bin 3: 27, 27, 36

Pembersihan Data: penghalusan data

2 Regresi: suatu teknik yang menyesuaikan nilai data dengan suatu fungsi. Regresi linier melibatkan pencarian garis "terbaik" sehingga satu atribut dapat digunakan untuk memprediksi yang lain.

3 Analisis outlier: Pencilan (*outlier*) dapat dideteksi dengan pengelompokan, misalnya, di mana nilai-nilai serupa diatur ke dalam kelompok atau cluster. Secara intuitif, nilai yang berada di luar kumpulan cluster dapat dianggap outlier.



Pembersihan Data: ketidakkonsistenan

Deteksi perbedaan. Perbedaan dapat disebabkan oleh beberapa faktor:

- formulir entri data dirancang dengan buruk (misal memiliki banyak bidang opsional),
- kesalahan manusia dalam entri data,
- kesalahan yang disengaja,
- data using,
- kesalahan perangkat instrumentasi
- kesalahan sistem
- representasi data dan penggunaan kode yang tidak konsisten,
- kesalahan integrasi data.

Alat komersial untuk membantu dalam langkah deteksi perbedaan:

- Alat penggosok data (*data scrubbing tools*) menggunakan pengetahuan domain sederhana untuk mendeteksi kesalahan dan melakukan koreksi pada data. Alat-alat ini mengandalkan teknik penguraian dan pencocokan fuzzy saat membersihkan data dari berbagai sumber.
- Alat audit data (*data auditing tools*) menemukan perbedaan dengan menganalisis data untuk menemukan aturan dan hubungan, dan mendeteksi data yang melanggar kondisi tersebut.

Proses dua langkah deteksi perbedaan dan transformasi data (untuk memperbaiki perbedaan) dilakukan secara berulang.

Transformasi Data

Strategi transformasi data:

1. Smoothing, untuk menghilangkan *noise* dari data. Teknik termasuk binning, regresi, dan clustering.
2. Konstruksi Atribut, atribut baru dibangun dan ditambahkan dari kumpulan atribut yang diberikan untuk membantu proses penambangan.
3. Agregasi, operasi ringkasan yang diterapkan pada data. Misalnya, data penjualan harian dapat digabungkan untuk menghitung jumlah total bulanan dan tahunan.
4. Normalisasi Data, data atribut diskalakan sehingga berada dalam rentang yang lebih kecil, seperti -1,0 hingga 1,0 atau 0,0 hingga 1,0.
5. Diskritisasi Data, nilai mentah dari atribut numerik (misalnya, usia) diganti dengan label interval (misalnya, 0-10, 11-20, dll.) atau label konseptual (misalnya, pemuda, dewasa, senior).
6. Pembentukan hierarki konsep untuk data nominal, yaitu atribut seperti "jalan" dapat digeneralisasikan ke konsep tingkat yang lebih tinggi, seperti "kota" atau "negara".

Transformasi Data: Normalisasi

Normalisasi min-max melakukan transformasi linier pada data asli. Nilai v_i menunjukkan nilai ke-i dari data atribut A yang dinormalisasi menjadi v_i', yaitu

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Normalisasi dengan penskalaan desimal dinormalisasi dengan memindahkan titik desimal dari nilai atribut A. Nilai v_i dari A dinormalisasi ke v_i' dengan menghitung

$$v_i' = \frac{v_i}{10^j}$$

j bilangan bulat terkecil sehingga $\max(|v_i'|) < 1$.

Normalisasi skor-z, nilai untuk atribut A dinormalisasi berdasarkan nilai mean dan deviasi standar dari A. Nilai v_i dari A dinormalisasi ke v_i' dengan menghitung

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

Variasi normalisasi z-score di atas

$$v_i' = \frac{v_i - \bar{A}}{s_A}$$

dengan $s_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|)$.

Deviasi absolut rata-rata lebih kuat (*robust*) terhadap outlier daripada deviasi standar.

Percobaan Praktikum di Modul

TUGAS

TUGAS KELOMPOK:

1. Mengunduh dataset yang memuat nilai yang hilang (*missing value*), kemudian menjelaskan kualitas data yang diperoleh.
2. Melakukan *data preprocessing*, diantaranya pembersihan data dan transformasi data.
3. Membuat visualisasi data sebelum dilakukan imputasi data dan setelah imputasi data, kemudian menginterpretasikan hasilnya.

Cara Pengumpulan:

- File yang dikumpulkan adalah laporan yang dibuat dengan R Markdown (.pdf atau .html yang memuat chunk syntax) dan dataset (.csv)
- Pengumpulan tugas maksimal sebelum pertemuan depan, pukul 23.59 WIB
- Beri nama file : Tugas 11_Kelompok XX

Referensi

Pathak, M. A., (2014), *Beginning Data Science with R*, Springer International Publishing, Switzerland.

Han J., Micheline K., & Jian P., (2012), *Data Mining Concepts and Techniques Third Edition*, Elsevier, United States of America.

https://datascienceplus.com/imputing-missing-data-with-r-mice-package/

https://www.statology.org/how-to-normalize-data-in-r/

KMMI 2021
Eksplorasi dan Visualisasi Data

Pertemuan 11:
Data Preprocessing