**MATERI MODUL ONLINE DATA MINING**
**PRAPROSES DATA**
**SESI ONLINE 6**
**Syefira Salsabila**

## 1. PENDAHULUAN

Teknologi informasi yang berkembang menciptakan sekumpulan data dan informasi yang semakin besar. Hal ini merupakan dampak dari peningkatan kebutuhan teknologi terhadap data, media penyimpanan, penggunaan database, penggunaan otomatisasi data via sensor, penelitian terkait sistem monitoring dan aplikasi *mobile* atau *smartphone*. Untuk menganalisa data dan membuat suatu pola dari data yang ada, maka data harus disusun, ditransformasi, diproses,dan dianalisa.

*Data mining* atau eksplorasi data yang dikenal dengan nama *Knowledge Discovery Database* (KDD) adalah suatu proses komputasi di dalam aplikasi dengan menggunakan algoritma tertentu untuk menemukan, mengekstraksi pola-pola dan informasi dari sekumpulan data yang ada. Semakin besar data yang disimpan maka semakin kaya hasil ekstraksi data yang didapat, sehingga semakin banyak pembuktian hipotesis yang dihasilkan. Melalui data mining dapat dilakukan ekstraksi pengetahuan dan analisa data untuk menemukan hubungan tiap data, struktur data, pola, dan *regularities*. Teknik metode statistik banyak dipakai sebagai alat utama menganalisa data untuk mengidentifikasi hubungan sebab akibat.

Teknik data mining memberikan hasil yang lebih daripada sebab akibat karena kemampuannya dalam menemukan, menganalisa pola dan hubungan dari data-data penelitian yang ada. *Data mining* dan analisa statistik konvensional memiliki perbedaan tujuan. Metode statistik klasik memiliki fokus utama yaitu memverifikasi hipotesis yang dibuat, sedangkan fokus utama metode *data mining* yaitu mencari secara menyeluruh hubungan antar data untuk semua kemungkinan hipotesis termasuk hipotesis yang belum diketahui atau belum dibuat. *Data mining* juga memiliki kelebihan dalam mengurangi data-data *noise* pada sekumpulan data yang besar.

Melalui *data mining* dapat dilakukan eksplorasi data secara menyeluruh. Proses eksplorasi dilakukan untuk mencari informasi dari data termasuk di dalamnya menjangkau sekumpulan *massive* data dengan efektif dan efisien. Informasi dari data tersebut akan disimpan, diakses atau digunakan kembali tanpa melalui proses sebelumnya, sehingga memudahkan dalam menemukan interaksi dari data yang ada untuk dimodelkan dan diinterpretasi.

Tantangan *data mining* untuk masa yang akan datang adalah sebagai berikut:

*a.* Meningkatkan teknik automatisasi data yang terkait data-data penelitian yang tidak lengkap seperti data *time series.*

b. Meningkatkan teknik dalam proses penggunaan data kembali (*reuse*) sehingga data dapat digunakan kembali secara otomatis contohnya untuk analisa tren.
c. Mengembangkan standar prosedur untuk pengujian percobaan dan validasi dari teknik *data mining*
d. Melibatkan *end user* dalam melakukan desain algoritma dan menterjemahkan hasil percobaan yang lebih baik
e. Mengembangkan dan mengimplementasi metode *data mining* dengan mengkombinasikan teknik yang telah ada untuk hasil yang lebih baik
f. Meningkatkan teknik *data mining* secara *online* dan dapat melibatkan *database* penelitian yang lebih beragam
g. Mengembangkan alat yang dapat menjelaskan secara eksplisit dalam memberikan penjelasan secara detail terhadap penemuan hasil data agar dapat lebih mudah dimengerti.
h. Mendesain dan menggunakan teknik spasial *data mining*.

Proses *data mining* akan mengurangi waktu dalam proses analisa data, mengurangi kesalahan akibat *human error* akibat data penelitian yang besar, dan mendapatkan hasil akhir bervariasi yang lebih dari hipotesis-hipotesis yang telah dibuat. Analisa data yang banyak sulit dilakukan secara manual. Untuk itulah *data mining* digunakan, sehingga menghasilkan informasi yang ingin diketahui dan informasi yang belum diketahui ketika proses akhir dilakukan terutama penelitian dengan menggunakan data ekologi.

Ada 7 (tujuh) tahapan proses data mining, dimana 4 (empat) tahap pertama disebut juga dengan data preprocessing (terdiri dari *data cleaning*, *data integration*, *data selection*, dan data transformation), yang dalam implementasinya membutuhkan. waktu sekitar 60% dari keseluruhan proses.

*Data mining* mempunyai lima fungsi yaitu:

a. *Classification*, yaitu menyimpulkan definisi-definisi karakteristik sebuah grup
b. *Clustering*, yaitu mengindentifikasi kelompok-kelompok dari data–datayang mempunyai karakteristik khusus (*clustering* berbeda dengan *classification*, dimana pada clustering tidak terdapat definisi-definisi karakteristik awal yang diberikan pada waktu *classification*)
c. *Association*, yaitu mengidentifikasi hubungan antara kejadian-kejadian yang terjadi pada suatu waktu
d. *Sequencing*, hampir sama dengan *association*, *sequencing* mengidentifikasi hubungan-hubungan yang berbeda pada suatu periode waktu tertentu
e. *Forecasting*, yaitu memperkirakan nilai pada masa yang akan datang berdasarkan pola-pola unik yang ada.

Setiap langkah dalam melakukan proses *data mining* membutuhkan ketelitian yang cukup. Dari banyak proses yang terdapat pada *data mining*, proses yang perlu dilakukan dengan sangat hati-hati adalah *data preprocessing*. *Data preprocessing*

merupakan langkah yang dilakukan sebelum masuk pada proses mining pada data. *Data preprocessing* berisi beberapa kegiatan yang tujuan utamanya adalah melakukan pengenalan dan perbaikan pada data yang akan diteliti. Perlunya perbaikan pada data yang akan diteliti disebabkan karena data mentah cenderung tidak siap untuk di-*mining*. Contoh kasus yang paling banyak terjadi adalah adanya *missing values* pada data. Missing value biasanya disebabkan karena nilai tidak relevan dengan kasus yang sebenarnya, terlewat pada waktu pengumpulan data, atau ada pengabaian pada waktu pengumpulan data.

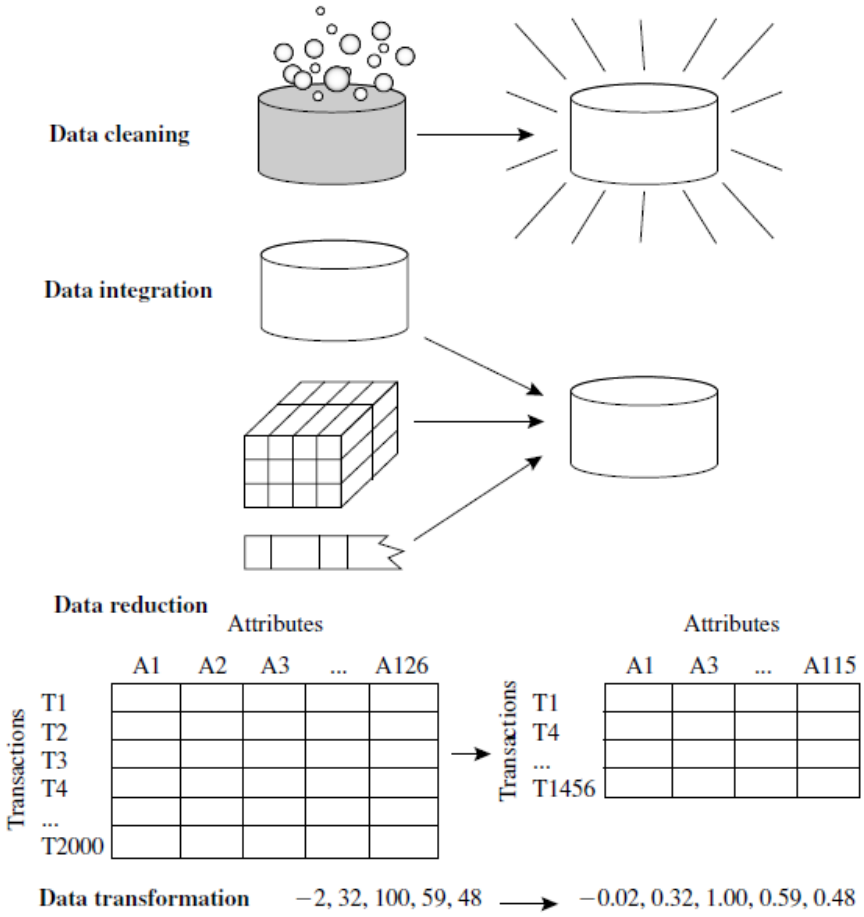
## 2. PRAPROSES DATA

Data dalam dunia nyata kotor data yang tidak lengkapnya nilai-nilai atribut kurang, atribut tertentu yang dipentingkan tidak disertakan, atau hanya memuat data agregasi. Sedangkan ada juga noisy yaitu yang memuat error atau memuat outliers (data yang secara nyata berbeda dengan data-data yang lain). Praproses data diadakan karena data yang tidak konsisten memuat perbedaan dalam kode atau nama. Sehingga data yang lebih baik akan menghasilkan data mining yang lebih baik. Data preprocessing membantu didalam memperbaiki presisi dan kinerja data mining dan mencegah kesalahan didalam data mining.

Data memiliki suatu kualitas jika data tersebut memenuhi kebutuhannya. Banyak faktor yang dapat mempengaruhi kualitas data, seperti dari sisi akurasinya, kelengkapannya, konsistensinya, ketepatan waktu, kepercayaan, dan interpretabilitas. Bayangkan kamu adalah seorang Kepala Departemen disuatu perusahaan dan ditugaskan untuk melakukan analisis data perusahaan untuk kebutuhan sales di kantor cabang. Secara langsung kamu akan memeriksa database dan data warehouse dalam perusahaan kamu, mengindentifikasi dan memilih atribut atau dimensi (contohnya jumlah barang, harga barang, dan yang akan dimasukan dalam analisis. Setelah itu Anda akan menyadari bahwa ada beberapa atribut yang tidak ada nilai recordnya. Sedangkan laporan yang kamu inginkan harus memasukan data purchasing semuanya, akan tetapi Anda menyadari bahwa datanya tidak lengkap. Untuk selanjutnya,pengguna dari sistem database akan melaporkan bahwa adanya eror, nilai yang tidak bisasa dan data yang tidak konsisten. Dengan kata lain data yang akan kamu lakukan analisis dengan teknik data mining tidak lengkap, tidak akurat (noisy) dan tidak konsisten.

Tahapan dari praproses data antara lain;

a. Pembersihan Data
b. Integrasi Data
c. Transformasi Data
d. Reduksi Data
e. Diskritisasi Data




Gambar 1. Ilustrasi dari Praproses Data

Pada data mentah sering ditemukan banyaknya nilai yang hilang (*missing value),* distorsi nilai, tidak tersimpannya nilai (*misrecording),* sampling yang tidak cukup bagus dan sebagainya. Penyebab kurang baiknya kualitas data mentah adalah karena adanya kesalahan dalam penyimpanan dan pengukuran, tapi bisa juga karena tidak adanya nilai mewakili yang tersedia. Outlier atau adanya nilai yang tidak biasa (lain dari umumnya) muncul karena banyak hal, antara lain kesalahan pada entri data dan adanya data yang tidak tersimpan sehingga nilai default otomatis tersimpan.

Data pasien merupakan data yang sangat penting dalam dunia kesehatan. Data pasien yang disimpan secara terstruktur dapat memberikan informasi tentang riwayat penyakit pasien. Namun demikian, ada beberapa kendala yang dihadapi untuk memperoleh informasi tersebut terkait dengan *dataset* pasien yang cukup besar. Diantara permasalahan yang ditemui dalam *dataset* pasien tersebut adalah adanya duplikasi data dan adanya *missing value*.

Duplikasi data dalam *data mining* dapat terjadi karena dua hal yaitu adanya *record* yang berulang dan adanya perbedaan identifikasi antara entitas yang sama dalam du-nia nyata. Adanya duplikasi data pada *dataset* dapat mempengaruhi kualitas perfor-ma *data mining*. Duplikasi data pada *dataset* pasien dimungkinkan terjadi karena *input* data pasien dilakukan oleh orang atau waktu yang berbeda. Selain itu, duplikasi juga bisa terjadi karena adanya kesalahan penulisan pada saat proses input data se hingga terdapat beberapa data yang memiliki kemiripan atau bahkan sama dalam *dataset*.

*Missing value* dalam *dataset* pasien berasal dari data-data yang atributnya tidak memiliki nilai. Informasi ini tidak diperoleh dimungkinkan karena adanya data pasien yang tidak lengkap seperti jenis kelamin, nama belakang pasien, tanggal lahir pasien, dan sebagainya. Keberadaan *missing value* ini juga dapat menyebabkan duplikasi data karena ada lebih dari satu data dengan nama yang sama dan memiliki kelengkapan data yang berbeda.

Pra-proses dilakukan karena dimungkinkan *data set* yang tidak lengkap, mengandung *noise* atau *outlier,* data tidak konsisten, atau ada data yang berulang. Tujuan penting dari pra-proses data adalah untuk meningkatkan kualitas data, sehingga proses data mining juga menghasilkan pengetahuan baru yang lebih baik. Tugas utama dalam pra-proses data adalah pembersihan data, integrasi data, transformasi data, reduksi data dan diskretisasi data. Tujuan dari praproses data antara lain;

a. Menghasilkan hasil *mining* yang berkualitas
b. Data *warehouse* membutuhkan integrasi yang konsisten
c. Data extraction, cleaning, and transformation merupakan salah satu tahapan untuk membangun gudang data

## 3. PEMBERSIHAN DATA (*DATA CLEANING)*

Pembersihan data yang kotor merupakan proses untuk mengisi nilai-nilai yang hilang, menghaluskan noisy data, mengenali atau menghilangkan outlier, dan memecahkan ketak-konsistenan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa data mining yang kita miliki. Data-data yang tidak relevan itu juga lebih baik dibuang karena keberadaannya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya. Garbage in garbage out (hanya sampah yang akan dihasilkan bila yang dimasukkan juga sampah) merupakan istilah yang sering dipakai untuk menggambarkan tahap ini. Pembersihan data juga akan mempengaruhi performasi dari sistem data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

Data *Cleaning* merupakan tahap pembersihan data merupakan tahap awal dari proses KDD. Seluruh atribut yang ada pada *dataset* di atas selanjutnya akan diseleksi untuk mendapatkan atribut-atribut yang berisi nilai yang relavan. Tidak *redundant* dan tidak *missing value*, dimana syarat tersebut merupakan syarat awal yang harus dikerjakan dalam *data mining* sehingga akan diperoleh *dataset* yang bersih untuk digunakan pada tahap *mining* data. Data dikatakan *missing value* bila atribut-atribut

dalam *dataset* tidak berisi nilai atau kosong, sedangkan data dikatakan *redundant* jika dalam satu *dataset* yang sama terdapat lebih dari satu *record* yang berisi nilai yang sama.

*Data* cleaning terdapat proses pembersihan data yang kosong atau data-data lain yang dapat mengakibatkan *noise*/*error* dengan memperkecil adanya data *outlier*. *Data Cleaning* menjadi proses yang sangat penting dikarenakan data perlu dibersihkan agar analisa menjadi lebih akurat. Beberapa metode dalam melakukan pembersihan data;

### 3.1 Mengisi *missing value*

Metode data mining seringkali mensyaratkan semua dinilai data lengkap atau tidak ada yang hilang. Padahal pada kenyataannya banyak atribut atau field dari beberapa record yang tidak diketahui nilainya. Solusi paling sederhana adalah dengan menghapus semua record yang berisi nilai yang kosong. Untuk data yang besar mungkin cara ini tidak berpengaruh terhadap model data mining yang dihasilkannya. Akan tetepi lain hasilnya jika data-data yang dihapus ini memiliki potensi yang sangat besar.

*Missing value* adalah kondisi di mana terdapat data atau informasi yang tidak ditemukan pada suatu atribut tertentu dalam *dataset*9. *Missing value* dapat terjadi karena nilainya tidak relevan untuk kasus tertentu, tidak bisa dicatat pada saat data dikumpulkan, atau disebabkan adanya privasi. *Missing Value* atau data tidak selalu tersedia maksutnya ada saja data yang memiliki banyak tuple atau record tidak memiliki nilai yang tercatat untuk beberapa atribut, seperti customer income dalam data sales. Hilangnya data bisa karena;

a) Kegagalan pemakaian peralatan
b) Tidak konsisten dengan data tercatat lainnya dan karenanya dihapus
c) Data tidak dimasukkan karena salah pengertian
d) Data tertentu bisa tidak dipandang penting pada saat entry
e) Tidak mencatat history atau tidak mencatat perubahan data
f) kehilangan data perlu disimpulkan

*Missing values* bisa dibagi menjadi 3 kelas berbeda berdasarkan karakteristik antar variabelnya :

a. *Missing Completely at Random*(MCAR) : *Missing values* tidak bergantung pada data lain
b. *Missing at Random*(MAR) : *Missing values* bergantung pada data lain, namun tidak bergantung pada data itu sendiri
c. *Not Missing at Random*(NMAR) : Peluang adanya *missing values* bergantung pada nilai atribut tersebut.

Pada umumnya, untuk menangani adanya *missing* data dapat dilakukan tiga kategori :

a. Mengabaikan atau menghapus *missing values* : Ada dua pendekatan pada kategori ini, yakni *complete case analysis* dan *discarding instances or attributes*. Hal ini mudah untuk dilakukan akan tetapi tidak efektif, dan merupakan metoda terakhir. Biasanya dilakukan saat label kelas hilang. Tidak efektif bila persentasi dari nilai-nilai yang hilang per atribut sungguh-sungguh bervariasi.
   Proses ini lebih cocok untuk data yang memiliki *missing value* sedikit. Cara lain dengan mengganti nilainya dengan nilai konstan seperti 999 atau ”unknown” atau dengan nilai rata-rata dari atribut. Namun ini dirasakan tidak efektif dan bisa menyebabkan keakuratan data berkurang. Oleh karena itu dibutuhkan suatu metode yang mampu melakukan imputasi terhadap *missing value* dengan nilai yang masuk akal

b. Estimasi parameter : Melakukan estimasi parameter dengan menggunkan *Maximum Likelihood*.
c. Imputasi : Mengisi *missing values* dengan menggunakan berbagi pendekatan.

## 3.2 *Meminimumkan noise*

Merupakan komponen random dari suatu error pengukuran. Noise berkaitan dengan modifikasi dari nilai asli. Contoh: distorsi atau penyimpangan dari suara orang saat berbicara di telepon yang jaringannya buruk. Noise bisa dikatakan juga error acak atau variansi dalam suatu variabel terukur. Terdapatnya nilai-nilai atribut tak benar mungkin karena:

a. Kegagalan instrumen pengumpulan data
b. Problem pemasukan data
c. Problem transmisi data
d. Keterbatasan teknologi
e. Ketak-konsistenan dalam konvensi penamaan

Cara untuk menangani Noise Data antara lain;

a. Metoda Binning:
   a) Pertama urutkan data dan partisi kedalam (kedalaman yang sama) bin-bin
   b) Kemudian noisy data itu bisa dihaluskan dengan rata-rata bin, median bin, atau batas bin.
b. Clustering
   a) Medeteksi dan membuang outliers
c. Inspeksi kombinasi komputer dan manusia
   a) Mendeteksi nilai-nilai yang mencurigakan dan memeriksa dengan manusia(misal, berurusan dengan outlier yang mungkin)
d. Regresi
   a) Menghaluskan dengan memasukkan data kedalam fungsi regresi

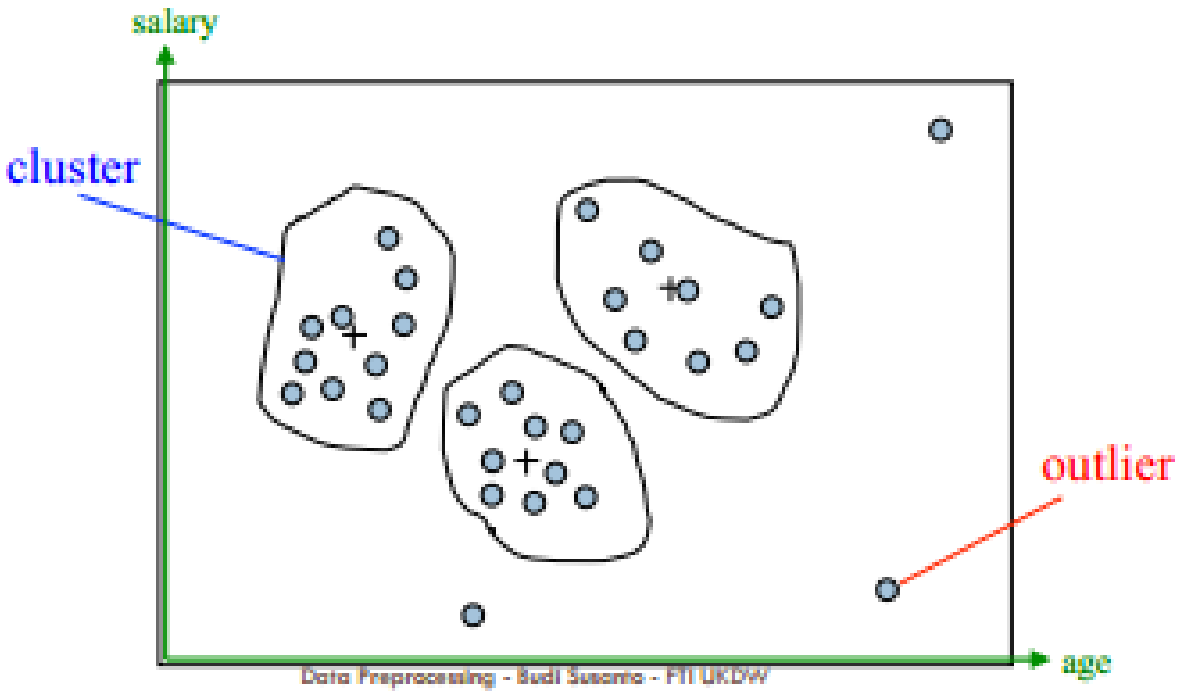
### 3.3 *Mendeteksi outlier*

Seringkali pada data set, terdapat suatu nilai yang berbeda dari biasanya dan tidak mencerminkan karakteristik data secara umum. Nilai yang tidak konsisten itu dnamakan outlier. Outlier/anomali adalah sehimpunan data yang dianggap memiliki sifat yang berbeda dibandingkan dengan kebanyakan data lainnya. Analisis outlier dikenal juga dengan analisis anomali atau deteksi anomali atau deteksi deviasi (nilai atributnya objek tsb, signifikan berbeda dengan nilai atribut objek lainnya ) atau *exception mining.* Beberapa hal penyebab yang menyebabkan data outlier antara lain;

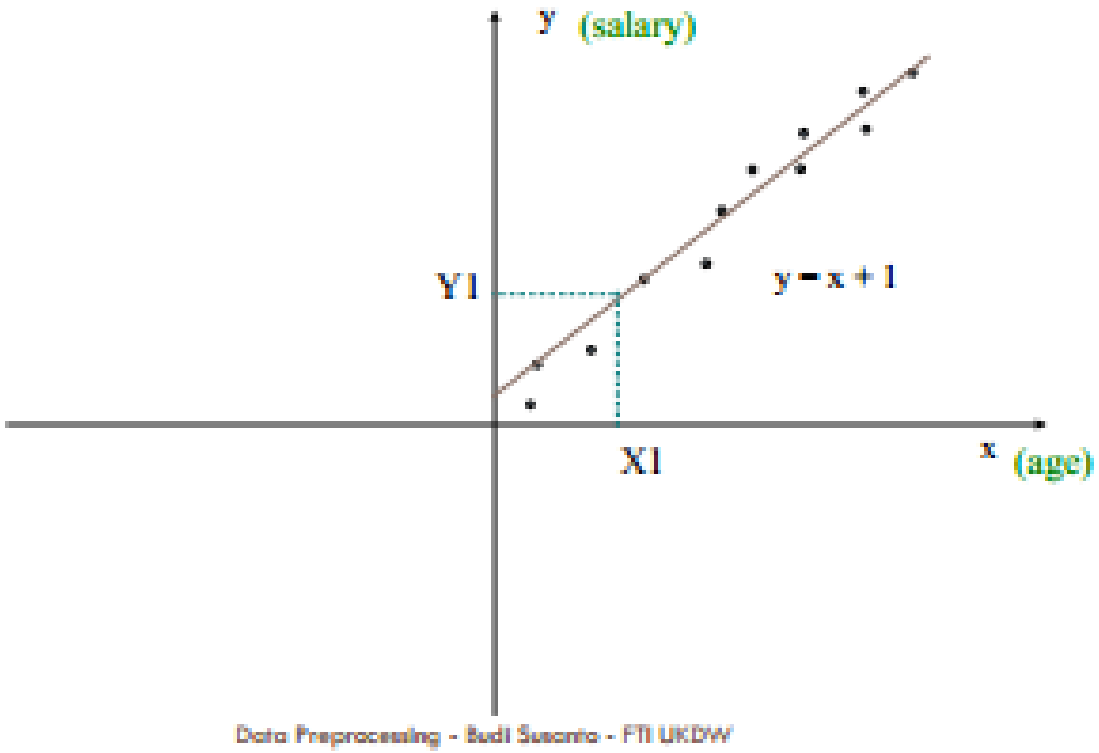
a. Data berasal dari kelas yang berbeda
b. Variasi natural data itu sendiri
c. Error pada saat pengukuran atau pengumpulan data

Cara mendeteksi Outlier bisa dengan dilakukan;

*a.* Clustering



*b.* Regresi linier




## 4. INTEGRASI DATA (*DATA INTEGRATION)*

Integrasi data merupakan suatu metode untuk mengkombinasikan data dari banyak sumber kedalam suatu simpanan terpadu. Dalam pendeteksian dan pemecahan konflik nilai data dapat dilakukan Untuk entitas dunia nyata yang sama, nilai-nilai atribut dari sumber-sumber berbeda adalah berbeda. Alasan yang mungkin: representasi berbeda, skala berbeda, misal berat bisa dalam pound atau kilogram.

Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dsb. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada. Dalam integrasi data ini juga perlu dilakukan transformasi dan pembersihan data karena seringkali data dari dua database berbeda tidak sama cara penulisannya atau bahkan data yang ada di satu database ternyata tidak ada di database lainnya.

### 4.1 Atribut Redudan

Data redundan sering terjadi saat integrasi dari banyak database seperti saat Atribut yang sama bisa memiliki nama berbeda dalam database berbeda selain itu

bisa juga saat Atribut yang satu bisa merupakan suatu atribut “turunan” dalam tabel lainnya, misal, annual revenue. Integrasi data hati-hati dari banyak sumber bisa membantu mengurangi/mencegah redundansi dan ketak-konsistenan dan memperbaiki kecepatan dan kualitas mining. Suatu atribut dikatakan redundan jika atribut tersebut bisa diperoleh dari atribut lainnya. Penyebab terjadinya redundansi antara lain ;
a) Atribut yang sama mempunyai nama yang berbeda pada database yang berbeda
b) Satu atribut merupakan turunan dari atribut lainnya

Cara mengatasi redundansi pada integrasi data dapat dilakukan dengan mencari hubungan korelasi antar variabel dapat dilihat menggunakan rumus korelasi. Jika data numerik, hubungan korelasinya seperti dibawah ini:

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\,\overline{A}\,\overline{B}}{(n-1)\sigma_A \sigma_B}$$

Semakin besar hasil perhitungan tersebut, semakin tinggi korelasi. Jika hasil perhitungan tersebut =0 berarti independen. Jika kurang dari nol tidak independen. Jika data kategorik, hubungan korelasinya seperti dibawah ini menggunakan chi-square:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Semakin besar chi-square, semakin tinggi korelasi. Jika hasil perhitungan tersebut =0 berarti independen. Jika kurang dari nol tidak independen

## 4.2 Duplikasi

Set data mungkin terdiri dari objek data yang ganda (duplikat), atau hampir selalu terjadi duplikasi antara satu dengan yang lainnya. Persoalan utama ketika menggabungkan data dari sumber-sumber yang bervariasi (heterogen). Contoh: orang yang sama dengan alamat email yang lebih dari satu. Pembersihan data (*data cleaning*) merupakan proses yang berkaitan dengan permasalahan data yang duplikat.

## 5 TRANFORMASI DATA (*DATA TRANSFORMATION)*

Proses Tranformasi data merupakan pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai. Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

Beberapa teknik data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa teknik standar seperti analisis asosiasi dan klastering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut binning. Disini juga dilakukan pemilihan data yang diperlukan oleh teknik data mining yang dipakai. Transformasi dan pemilihan data ini juga menentukan kualitas dari hasil data mining nantinya karena ada beberapa karakteristik dari teknik-teknik data mining tertentu yang tergantung pada tahapan ini.

Transformasi data merupakan proses penghalusan untuk menghilangkan noise dari data. Hal yang termasuk dalam proses transformasi antara lain:

a. *Smoothing* : menghapus noise dari data

Smoothing dilakukan jika data mengandung noise/nilai yang tidak valid terhadap data yang di-mining. Untuk mengatasinya harus dilakukan smoothing (dengan memperhatikan nilai-nilai tetangga). Berikut teknik atau metode untuk smoothing:

*a) Binning*

Metode binning dilakukan dengan memeriksa “nilai tetangga”, yaitu nilai-nilai yang ada disekelilingnya. Berikut adalah langkah-langkah metode binning:

1. Data diurutkan dari yang terkecil sampai dengan yang terbesar.
2. Data yang sudah urut kemudian dipartisi ke dalam beberapa bin. Teknik partisi ke dalam bin ada 2 (dua) cara: *equal-width* (*distance*) *partitioning* dan *equaldepth* (*frequency*) partitioning.
3. 3. Dilakukan smoothing dengan tiga macam teknik, yaitu: smoothing *by binmeans*, smoothing *by bin-medians*, dan smoothing *by bin-boundaries*.

*b) Clustering*

Digunakan untuk menyingkirkan *outliers* (keluar jauh-jauh dari cluster/*centroid*), data yang memiliki noise. Algoritma *k*-Means yang merupakan kategori metode partitioning dapat digunakan jika ukuran *database* tidak terlalu besar. Algoritma ini didasarkan pada nilai tengah dari objek yang ada dalam cluster. Algoritma *k*-Means meminta inputan parameter *k*, dan mempartisi satu set *n* objek ke dalam *k* cluster sehingga menghasilkan tingkat kemiripan yang tinggi antar objek dalam kelas yang sama (*intra-class similarity*) dan tingkat kemiripan yang paling rendah antar objek dalam kelas yang berbeda (*inter-class similarity*). Kemiripan cluster diukur dengan menghitung nilai tengah dari objek yang ada di dalam cluster.

*c) Regression*

*Linear regression* memodelkan sebuah random variable, *Y* (disebut *response variable*) sebagai sebuah fungsi linier dari random variable yang lain, *X* (disebut sebagai *predictor variable*), dengan persamaan empiris: $Y = \alpha + \beta X$ , dimana α

dan β adalah koefisien regresi. Koefisien ini dapat dihitung menggunakan metode *least squares* dengan persamaan sebagai berikut:

$$\beta = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

dan

$$\alpha = \bar{y} - \beta\bar{x}$$

dimana *x* adalah nilai ratarata dari *x1, x2, …, xi* dan *y* adalah nilai rata-rata dari *y1, y2, …, yi*.

b. *Aggregation* : Ringkasan, kontruksi data cube

Adalah operasi *summary* (peringkasan) diaplikasikan pada data numerik. Misalnya pada data penjualan harian digabungkan untuk menghitung pendapatan perbulan dan pertahun dengan dirata-rata atau ditotal. Langkah ini dilakukan dengan memanfaatkan operator *data cube* (operasi *roll up*/meringkas).

c. Normalisasi : Min-max, Z-score, Decimal Scaling

Normalization atau normalisasi adalah proses transformasi dimana sebuah atribut numerik diskalakan dalam range yang lebih kecil seperti -1.0 sampai 1.0, atau 0.0 sampai 1.0. Ada beberapa metode/teknik yang diterapkan untuk normalisasi data, diantaranya:

5.1 Normalisasi min-max
Menghasilkan [new_min, new_max]

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

Contoh soal:

Penghasilan berkisar dari $10,000 sampai $98,000 dinormalisasikan dari [0,1]. Sehingga untuk penghasilan sebesar $73,000 dipetakan ke $\frac{73{,}000-10{,}000}{98{,}000-10{,}000}$(1-0)+0=0.716

5.2 Normalisasi z-score

*Normalization: μ: mean, σ: standard deviation*

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Contoh soal:

Misal $\mu$ = 55,000, $\sigma$ = 20,000. Maka, $\frac{73{,}000-55000}{20000}$= 0.9

5.3 Normalisasi dengan penskalaan desimal

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

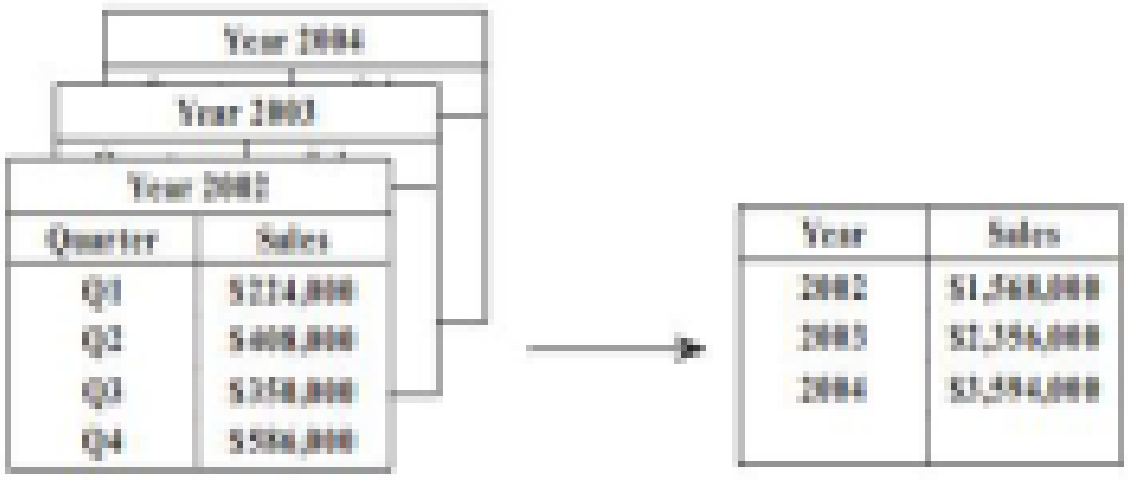
Dimana *j* adalah integer terkecil sehingga Max(| $v'$ |)<1

## 6 REDUKSI DATA (*DATA REDUCTION)*

Memperkecil volume tapi menghasilkan analasis data yang sama. Strategi-strategi data reduksi: Data cube aggregation, reduksi dimensi (menghapus atribut yang tidak penting), kompresi data, dsb. Suatu data warehouse bisa menyimpan terabytes data. Analisis/menambang data kompleks bisa membutuhkan waktu sangat lama untuk dijalankan pada data set komplit (tak efisien). Reduksi data merupakan proses untuk mengurangi ukuran data set tetapi menghasilkan hasil analitis yang sama (hampir sama)

Strategi dalam mereduksi data, antara lain;

b. Agregasi kubus data




b. Reduksi dimensionalitas—menghilangkan atribut tak penting

c. Kompresi data
d. Reduksi Numerosity reduction—mencocokkan data kedalam model
e. Diskritisasi dan pembuatan konsep hierarki

## 7. DISKRITASI DATA *(DATA DISCRETIZATION)*

Terdapat tiga tipe atribut:

a. Nominal = Nilai dari sekumpulan data yang tidak beraturan. Contoh: Warna, Profesi
b. Ordinal = Nilai dari sekumpulan data yang terurut.Contoh: Ip, nomor antrian
c. Kontinu = Nilai real seperti integer atau real number

Metode disktritisasi bisa dilakukan pada data kontinu. Tahap pertama, kita mengelompokkan nilai ke dalam interval. Setelah itu kita menggantikan nilai atribut dengan label atau interval.
Contoh:

Dataset (age, salary): (26;56,000),(28;70,000),(89;99,000)

Metode disktritisasi bisa dilakukan pada data kontinu.Tahap pertama, kita mengelompokkan nilai ke dalam interval. Setelah itu kita menggantikan nilai atribut dengan label atau interval.

Contoh:

Dataset (age, salary): (26;56,000),(28;70,000),(89;99,000)

## DAFTAR PUSTAKA


Gorunescu, F. (2011). *Data Mining : Concepts, Models and Techniques.* New York: Springer-Verlag.

Han, Jiawei dan Micheline Kamber. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann. 2001.

Junaedi H, Budianto H, Maryati I, Melani Y. Data Transformation Pada Data Mining. Surabaya: IdeaTech2011. 2011.

Kurniawan E. Analisa Data Rekam Medis Menggunakan Teknik Data Mining Association Rules Dengan Algoritma Clustering. InPROSIDING SEMINAR NASIONAL & INTERNASIONAL 2017.