PERTEMUAN 4

TEKNIK-TEKNIK PRAPROSES DATA MENGGUNAKAN R

TUJUAN PRAKTIKUM

Mahasiswa akan dapat menggunakan teknik-teknik dalam praproses data meliputi pembersihan data, reduksi data, transformasi data dan diskretisasi data dengan bahasa pemrograman R.

TEORI PENUNJANG

Pra-proses data dilakukan karena dimungkinkan dataset yang ada 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.

Seperti bahasa pemrograman umumnya, R memiliki nilai-nilai khusus yang merepresentasikan pengecualian-pengecualian untuk tipe data normal lainnya. Nilai tersebut yaitu:

- NA, yang berarti *not available*. NA menggantikan nilai *missing value*. Operasi dasar dari R dapat memproses dataset yang berisikan nilai NA dan terkadang R juga dapat mengembalikan nilai NA sebagai hasil dari suatu operasi walaupun *input* dari *argument* tersebut tidak terdapat NA. Cobalah kode berikut untuk memahami arti nilai NA di R:

```
> NA + 1
> sum(c(NA, 1, 2))
> median(c(NA, 1, 2, 3), na.rm = TRUE)
> length(c(NA, 2, 3, 4))
> 3 == NA
> NA == NA
> TRUE | NA
```

- NULL, yang berarti nilai yang kosong dan memiliki panjang 0. Cobalah kode berikut untuk memahami arti nilai NULL di R:

```
> length(c(1, 2, NULL, 4))
> sum(c(1, 2, NULL, 4))
> x <- NULL
> c(x, 2)
```

- Inf, yang berarti *infinity* dan hanya digunakan pada vektor dengan kelas numerik. Contoh kasus yang menghasilkan nilai Inf adalah operasi pembagian dengan 0. Cobalah kode berikut untuk memahami arti nilai Inf di R:

```
> pi/0
> 2 * Inf
> Inf - 1e+10
```

```
> Inf + Inf
> 3 < -Inf
>Inf == Inf
```

- NaN, yang berarti *not a number*. Biasanya hal ini menunjukkan hasil perhitungan yang tidak diketahui. Contohnya pada operasi 0/0. Inf-Inf, Inf/Inf menghasilkan nilai NaN. Cobalah kode berikut untuk memahami arti nilai NaN di R:

```
> NaN + 1
> exp(NaN)
```

1. Eksplorasi data
   Dalam R terdapat beberapa cara untuk melakukan eksplorasi data, yaitu dengan mengetahui tipe data dari setiap atribut dan mengetahui persebaran data setiap atribut.

```
> data<-airquality
> str(data)
> summary(data)
```

   Untuk mengetahui jumlah data yang missing, dapat dilakukan dengan cara berikut:
   - Install paket **`mice`**
   - Gunakan fungsi `md.pattern (dataset)`

```
> library(mice)
> data<-airquality
> md.pattern(dataset)
     Wind Temp Month Day Solar.R Ozone
111    1    1     1   1       1     1  0
 35    1    1     1   1       1     0  1
  5    1    1     1   1       0     1  1
  2    1    1     1   1       0     0  2
       0    0     0   0       7    37 44
```

   Dari hasil program R tersebut kita dapat melihat distribusi dari nilai missing value untuk setiap atribut.

2. Pembersihan data
   - Dari kode sebelumnya kita akan mencoba mengisikan nilai *missing value* untuk setiap atribut

```
> data<-airquality
> #mengisikan nilai mean untuk missing value di atribut
Solar.R
> data$Solar.R[is.na(data$Solar.R)] <- mean (data$Solar.R,
na.rm= TRUE)
> md.pattern(data)
     Solar.R Wind Temp Month Day Ozone
116       1    1    1     1   1     1  0
 37       1    1    1     1   1     0  1
          0    0    0     0   0    37 37
```

- Untuk atribut dengan tipe data kategorikal dapat menggunakan fungsi modus, yaitu: `names(sort(-table(x))[1]`

3. Transformasi data
   - Membuat variabel baru bernama variabel bulan untuk dataset kode sebelumnya, yaitu:

```
> data$bulan<-NULL
> data$bulan[data$Month == 5] <- “Mei”
> data$bulan[data$Month == 6] <- “Juni”
> data$bulan[data$Month == 7] <- “Juli”
> data$bulan[data$Month == 8] <- “Agustus”
> data$bulan[data$Month == 9] <- “September”
```

4. Reduksi data
   - Menghapus variabel tertentu:

```
> data$bulan<-NULL
```

   - Sampling data
     - Memilih data berdasarkan kriteria tertentu dapat menggunakan fungsi `which()`. Misal: ingin memilih data yang memiliki nilai atribut month lebih besar dari 7 dan nilai atribut wind lebih besar dari 10.

```
> dataBaru<-data [which(data$Month>7 & data$Wind>=10),]
```

     - Memilih data secara random

```
>dataRandom<-data[sample(1:nrow(data),50,replace=FALSE),]
```

5. Integrasi data
   - Pada R, dataset biasanya dalam format data frame. Untuk menggabungkan data dari lebih satu sumber, prasyarat yang harus dipenuhi adalah jumlah atribut dan tipe atribut dari dataset yang akan digabungkan sama.
   - Untuk menggabungkan dua dataset dapat menggunakan kode berikut:

```
> #menggabungkan dataset dibagian kolom, syarat: jumlah baris harus sama
> total <- merge(data frameA,data frameB,by="ID")
> # menggabungkan 2 dataset secara baris, syarat: jumlah kolom harus sama.
> # contoh penggabungan data dari tahap reduksi sampling data
> dataGabung<- rbind(dataBaru, dataRandom)
```

6. Diskretitasi data
   - Data binning dapat menggunakan fungsi *cut*.

```
> Contoh membagi atribut wind menjadi 3 kelompok
> data$Wind<-cut (data$Wind, 3, include.lowest = TRUE)
```

- Terdapat 2 pendekatan dalam melakukan diskretisasi, yaitu unsupervised dan supervised.
- Diskretisasi tidak terbimbing (*unsupervised*) dapat dilakukan dengan paket infotheo pada fungsi discretize.

```
> install.packages("infotheo")
> library(infotheo)
> #contoh melakukan diskretisasi pada atribut sepal dengan membagi menjadi 3 kategori dengan metode equal width
> ew.Sepal <- discretize(data$Sepal.Length, "equalwidth", 3)

> #contoh melakukan diskretisasi pada atribut petal dengan membagi menjadi 3 kategori dengan metode equal frequency
> ef.Petal <- discretize(data$Petal.Length, "equalfreq", 3)
```

- Diskretisasi terbimbing (*supervised*) dapat dilakukan dengan paket discretization,

```
> install.packages("discretization")
> library(discretization)
```

LAPORAN PENDAHULUAN

1. Apakah setiap akan melakukan tahapan data mining harus dilakukan praproses data?
2. Apakah semua teknik praproses data harus dilakukan ketika melakukan praproses data?
3. Bagaimana menetukan jenis praproses data yang tepat untuk menyiapkan suatu dataset?

MATERI PRAKTIKUM

1 Praproses Data
2 Pembersihan Data
3 Integrasi Data
4 Transformasi Data
5 Reduksi Data
6 Diskretisasi Data