**Pokok Bahasan I**

**Data Pre-processing**

**Kode Pokok Bahasan**: TIK.RPL03.001.001.01

**Deskripsi Pokok Bahasan**:

Membahas bagaimana pengolahan data sebelum diproses menggunakan bahasa R. Mempersiapkan sebuah data mentah, agar dapat sesuai dengan spesifikasi untuk tahap processing data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Elemen Kompetensi | Indikator Kinerja | Jml Jam | Hal |
| 1 | Memahami data pre-processing. | 1.1 Menganalisa data pre-processing dari data yang disediakan | 1 | 5 |
| 2 | Mengimplementasikan pre-processing data. | 1.1 Mampu memproses data dengan Missing Values  1.2 Mampu mengidentifikasi nilai NA disetiap column dan menghapus baris yang memiliki nilai NA  1.3 Mampu memvisualisasikan data dengan histogram  1.4 Mampu membagi data  menjadi training dan test  dataset.  1.5 Mampu menerapkan feature scaling pada data. | 2 | 6 |

**TUGAS PENDAHULUAN**

Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :

1. Menginstal R pada PC masing-masing praktikan.
2. Menginstal R Studio pada PC masing-masing praktikan.

**DAFTAR PERTANYAAN**

1. Apa yang dimaksud dengan data pre-processing?

Data preprocessing adalah proses yang mengubah data mentah ke dalam bentuk yang lebih mudah dipahami. Proses ini penting dilakukan karena data mentah sering kali tidak memiliki format yang teratur.

1. Mengapa perlu adanya tahapan pre-processing?

Preprocessing data penting untuk dilakukan karena dapat memberikan fungsi atau manfaat pada data mining. Proses ini utamanya dilakukan untuk memastikan kualitas data baik sebelum digunakan saat analisis data.

1. Sebutkan langkah-langkah dari data pre-processing?
2. Data cleaning.
3. Data integration.
4. Data transformation.
5. Data reduction.

**TEORI SINGKAT**

Data Preparation atau bisa disebut juga dengan data preprocessing adalah suatu proses/langkah yang dilakukan untuk membuat data mentah menjadi data yang berkualitas(input yang baik untuk data mining tools). Tahapan dalam melakukan data mining salah satunya adalah preprosesing data. Pertanyaannya adalah mengapa data perlu di bersihkan sebelum diproses?  
Hal ini terjadi karena biasanya data yang akan digunakan belum baik, penyebabnya antara lain :

– **Incomplete** : kekurangan nilai-nilai atribut atau atribut tertentu lainnya.  
– **Noisy** : berisi kesalahan atau nilai-nilai outlier yang menyimpang dari yang diharapkan.

– **Inconsisten** : ketidakcocokan dalam penggunaan kode atau nama.

Disini kualitas data yang baik didasarkan oleh keputusan yang baik dan data warehouse memerlukan integrasi kualitas data yang konsisten.  
Beberapa hal yang perlu diperhatikan untuk mendapatkan data yang baik adalah :

–**Accuracy**  
–**Completeness**  
–**Consistency**  
–**Timeliness**  
– **Value added**

–**Interpretability**  
–**Accessibility**  
–**Contextual**  
– **Representational**

**LAB SETUP**

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.
2. Menjalankan R Studio.

**ELEMEN KOMPETENSI I**

**Deskripsi:**

Memahami data pre-processing.

**Kompetensi Dasar**:

1. Menganalisa data pre-processing dari data yang telah disediakan.

**Latihan 1.1.1**

**Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk melakukan analisis terhadap data yang telah disediakan.

**Apakah data yang disediakan, termasuk data yang perlu di pre-processing? Jika ya, jelaskan alasannya!**

**-Data yang disediakan adalah data Train,Test,Dataset**

**-Seluruh data harus di processing karena kita sendiri harus mengisi data-data yang NA(Kosong)**

**ELEMEN KOMPETENSI** **2**

**Deskripsi:**

Mengimplementasikan pre-processing data.

**Kompetensi Dasar:**

1. Mampu memproses data dengan Missing Values

2. Mmapu membagi data menjadi training dan test dataset.

3. Mampu menerapkan feature scaling pada data.

**Latihan 1.2.1**

**Penjelasan Singkat:**

Pada latihan ini praktikan diminta untuk memproses missing value dari data yang telah disediakan. Bagaimana cara agar missing value tersebut tidak menjadi kendala dalam tahap processing.

**Langkah-Langkah Praktikum:**

1. Impor data train.csv dan test.csv dari <https://www.kaggle.com/c/titanic/data>**.**
2. Baca file excel dengan perintah :

> dataku\_namapraktikan=read.csv("E:/Kuliah Data Mining gasal 1920/train.csv").

1. Lihat 5 data teratas dengan perintah :

> head(dataku\_namapraktikan)

Output :

|  |
| --- |
|  |

1. Jalankan perintah berikut :

> library("readxl")

> dataku\_namapraktikan <- read\_excel("E:/Kuliah Data Mining gasal 1920/train.xlsx")

> head(dataku\_namapraktikan)

1. Apakah baris ke-6 mengandung missing value pada variabel Age ? Y/T

> str(dataku\_namapraktikan)

Output :

|  |
| --- |
|  |

6. Membuat data frame dengan logika nilai True and False untuk mengindikasi Missing Value

> is.na(dataku\_namapraktikan)

output :

|  |
| --- |
|  |

> apply(is.na(dataku\_namapraktikan), 2, which)

output :

|  |
| --- |
|  |

7. Selanjutnya Mengatasi missing value dengan cara menghapus baris yang yang terdapat missing value

> data\_drop = na.omit(dataku\_namapraktikan)

output :

|  |
| --- |
|  |

> view(data\_drop)

output :

|  |
| --- |
|  |

Pastika sudah tidak ada missingvalue pada table

8. Buat Histogram berdasarkan Usia

> hist(dataku\_namapraktikan $Age, main = "Histogram Passange",

+ xlab = "Usia",xlim=c(0,100), ylim = c(0,250), col="blue")

output :

|  |
| --- |
|  |

> boxplot(dataku$Age, main = "Boxplot Age")

output :

|  |
| --- |
|  |

9. Mengatasi missing value dengan cara mengganti nilai missing value dengan mean

> dataku\_namapraktikan=read.csv("E:/Kuliah Data Mining gasal 1920/train.csv").

output :

|  |
| --- |
|  |

> mean(dataku\_namapraktikan$Age)

Output :

|  |
| --- |
|  |

> mean(dataku\_namapraktikan$Age, na.rm = 'TRUE')

Output :

|  |
| --- |
|  |

> dataku \_namapraktikan$Age = ifelse(is.na(dataku\_namapraktikan$Age),ave(dataku \_namapraktikan$Age, FUN = function(x) mean(x, na.rm = 'TRUE')), dataku\_namapraktikan$Age)

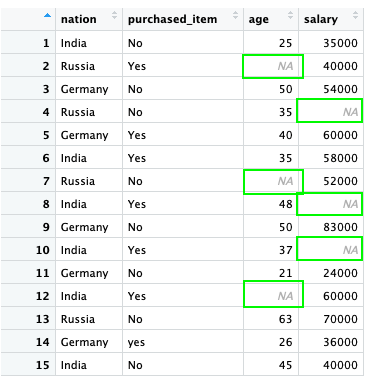
> head(dataku\_namapraktikan)

Output :

|  |
| --- |
|  |

Apakah variabel Age baris ke-6 masih mengandung missing value ? Y/T Y

Latihan :



Lakukan penanganan missing value pada variabel age dan salary sesuai langkah-langkah yang telah dipelajari :

Output :

|  |
| --- |
|  |

**Latihan 1.2.2**

**Penjelasan Singkat:**

Pada latihan ini praktikan diminta untuk mampu membagi data menjadi training dan test dataset.

**Langkah-Langkah Praktikum:**

1. Gunakan dataaset seperti berikut :

**Dataset**

## Country Age Salary Purchased

## 1 France 44 72000 No

## 2 Spain 27 48000 Yes

## 3 Germany 30 54000 No

## 4 Spain 38 61000 No

## 5 Germany 40 **NA**  Yes

## 6 France 35 58000 Yes

## 7 Spain **NA**  52000 No

## 8 France 48 79000 Yes

## 9 Germany 50 83000 No

## 10 France 37 67000 Yes

2. Ganti nilai NA dengan nilai average dari kolom terkait!

> install.packages("catTools")

> library(caTools) #adding caTools to the library

> set.seed(123)

> split = sample.split(dataset$Purchased,SplitRatio = 0.8)

> training\_set = subset(dataset,split == TRUE)

> test\_set = subset(dataset, split == FALSE)

3. Tampilkan 5 data teratas

> head (training\_set)

**output :**

|  |
| --- |
|  |

>head(test\_set)

**output :**

|  |
| --- |
|  |

**Latihan 1.2.3**

**Penjelasan Singkat:**

Pada latihan ini praktikan diminta untuk mampu melakukan featuring scaling pada data uji.

**Langkah-Langkah Praktikum:**

**1.** Lakukan feature scaling pada training\_set dan test\_set

#feature scaling

training\_set[,2:3] = scale(training\_set[,2:3])

test\_set[,2:3] = scale(test\_set[,2:3])

output :

|  |
| --- |
|  |

2. Ambil dataku\_namapraktikan dan lakukan fungsi scale.

> training\_namapraktikan<-dataku\_namapraktikan[,c(2,3)]

> scale\_ namapraktikan <- scale(training\_namapraktikan)

output :

|  |
| --- |
|  |

**CEK LIST**

|  |  |  |  |
| --- | --- | --- | --- |
| Elemen Kompetensi | No Latihan | Penyelesaian | |
| Selesai | Tidak selesai |
| 1 | 1.1.1 |  |  |
| 2 | 1.2.1 |  |  |
| 1.2.2 |  |  |
| 1.2.3 |  |  |

**FORM UMPAN BALIK**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Elemen Kompetensi** | **Tingkat Kesulitan** | | | **Tingkat Ketertarikan** | | | **Waktu Penyelesaian dalam menit** |
| Memahami data pre-processing. |  |  |  |  |  |  |  |
|  |  | Sangat Mudah |  |  | Tidak Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Mudah |  |  | Cukup Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | * Biasa |  |  | Tertarik | 30 |
|  |  |  |  |  |  |  |
|  |  | Sulit |  |  | * Sangat Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sangat Sulit |  |  |  |  |
|  |  |  |  |  |  |  |
| Mengimplementasikan pre-processing data. |  |  |  |  |  |  |  |
|  |  | Sangat Mudah |  |  | Tidak Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Mudah |  |  | Cukup Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | * Biasa |  |  | Tertarik | 30 |
|  |  |  |  |  |  |  |
|  |  | Sulit |  |  | * Sangat Tertarik |  |
|  |  |  |  |  |  |  |
|  |  | Sangat Sulit |  |  |  |  |