**Pokok Bahasan II**

**Data Pre-processing with Python**

**Kode Pokok Bahasan**: TIK.RPL03.001.002.01

**Deskripsi Pokok Bahasan**:

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Elemen Kompetensi | Indikator Kinerja | Jml Jam | Hal |
| 1 | Memahami data preprocessing menggunakan Python. | 1.1 Menganalisa data preprocessing dari data yang disediakan | 1 |  |
| 2 | Mengimplementasikan preprocessing data pada console Jupyter Notebooks. | 1.1 Mampu memproses data dengan Missing Values  1.2 Mampu menangani missing Value | 2 |  |

**TUGAS PENDAHULUAN**

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

1. Menginstal Anaconda Python pada PC masing-masing praktikan.
2. Menginstal Jupyter Notebooks pada Anaconda Python untuk masing-masing praktikan.

**DAFTAR PERTANYAAN**

1. Apa yang dimaksud dengan Missing Value?

Missing Value adalah hilangnya beberapa data yang telah diperoleh.

1. Adakah perbedaan penanganan Missing Value antara R dan Python, berikan Alasannya?

R: Mean dan Algoritma K-Means merupakan metode yang dapat digunakan untuk imputasi missing data.

Python: Kita dapat gunakan fungsi dropna() untuk menghapus data yang tidak memiliki nilai dan fungsi fillna() untuk mengisi nilai yang kosong.

**TEORI SINGKAT**

Salah satu Python IDE yang banyak digunakan adalah Jupyter Notebook yang mana mulai digunakan pada tahun 2004. Jupyter Notebook merupakan aplikasi web berdasarkan struktur server-client yang memungkinkan kita melakukan manipulasi pada notebook yang sedang dijalankan.

Jupyter Notebook adalah singkatan dari 3 bahasa pemrograman yaitu Ju (Julia), Py (Python), dan R. Ia merupakan tools yang populer digunakan untuk melakukan pengolahan data bagi seorang data scientist yang memungkinkan untuk mengintegrasikan antara kode dengan output di dalam satu dokumen secara interaktif.

**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 Jupyter Notebooks.

**ELEMEN KOMPETENSI I**

**Deskripsi:**

Memahami data pre-processing menggunakan Python.

**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!**

**ELEMEN KOMPETENSI** **2**

**Deskripsi:**

Mengimplementasikan pre-processing data.

**Kompetensi Dasar:**

1. Mampu memproses data dengan Missing Values
2. Mmapu menangani missing Value

**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 menggunakan python.

**Langkah-Langkah Praktikum:**

1. Letakan dataset dalam sebuah folder, dataset adalah data train.csv yang didownload dari https://www.kaggle.com/c/titanic/data

2. Buka anaconda prompt dan jalankan perintah berikut ini :

(base) C:\Users\Sectio>

(base) C:\Users\Sectio>>cd documents

(base) C:\Users\Sectio\Documents >cd Aslab Data Mining

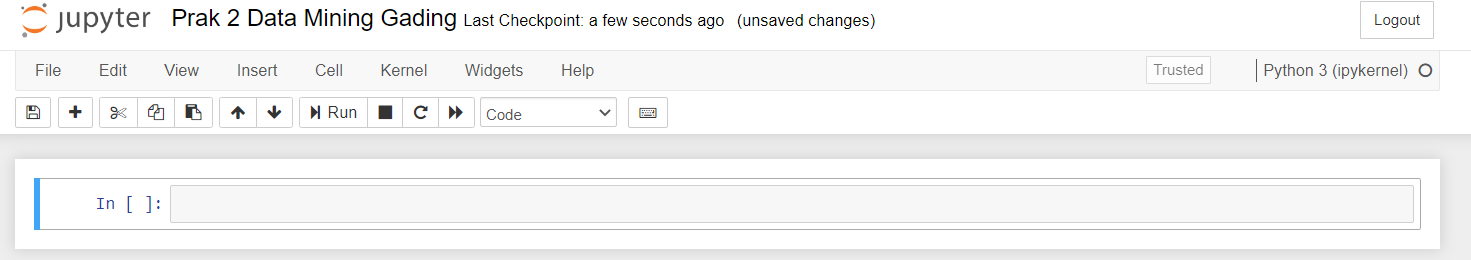
(base) C:\Users\Sectio\Documents\Aslab Data Mining >jupyter notebook

Text

Description automatically generated

1. Buat file baru pada jupyter notebook

Graphical user interface, text, application, email

Description automatically generated

1. Import Library

import pandas as pd

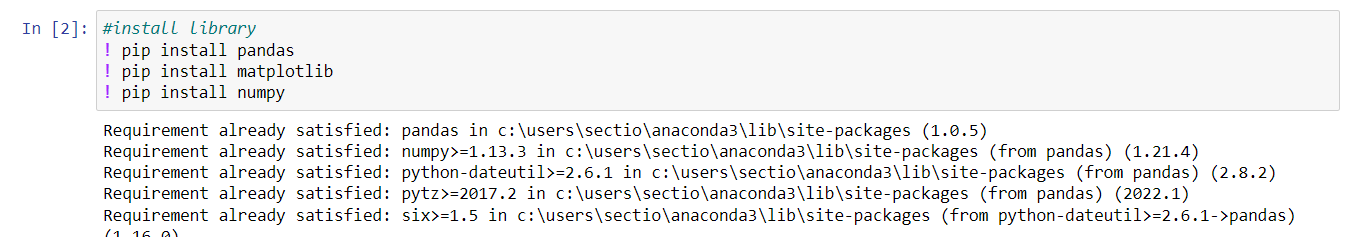
import matplotlib.pyplot as plt

import numpy as np

Graphical user interface, application

Description automatically generated

Jika library belum diinstall maka install terlebih dahulu library tersebut dengan perintah berikut



1. Mengambil data dari file .csv data\_namapraktikan=pd.read\_csv(‘train.csv’) data\_namapraktikan.head(5)

Text

Description automatically generated

1. Adakah missing value pada variabe Age? Y
2. Mengecek data secara umum dengan sintaks data\_namapraktikan.info()

-ada berapa baris dan kolom?

-apa fitur/kolom yang ada pada data?

-apa tipe data dari masing-masing kolom?

-fitur/kolom apa yang mengandung data null?

-kunjungi <https://www.kaggle.com/c/titanic> dan dapatkan informasi lengkap dari setiap kolom yang ada.

Text, table

Description automatically generated with medium confidence

1. Mengambil nama kolom dan tipe data dari masing-masing kolom

Selain dengan data\_frame.info(), pengambilan informasi kolom dapat dilakukan dengan

#data\_frame.columns

#data\_frame.dtypes

Dimana data\_frame adalah nama variabel yang menyimpan data dengan tipe data DataFrame

Cek apa representasi dari data bertipe object pada kolom/fitur Name, Sex, dan Ticket?

Graphical user interface, application, Word

Description automatically generated

1. Mengidentifikasi nilai mean, median, min value, max value quartile dan standar deviasi pada data bertipe numerik.

Table

Description automatically generated

1. Cara lain mengecek missing value selain dengan data\_frame.info() adalah dengan cara mengambil nama kolom dan menghitung jumlah data pada kolom yang bernilai null dengan data[nama\_kolom].isnull().sum()

Kolom mana yang mengandung banyak nilai null?

Graphical user interface, text, application

Description automatically generated

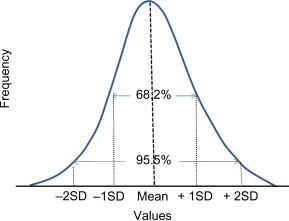
1. Data kolom Cabin mengandung banyak sekali nilai null, kita bisa mengabaikan kolom tersebut dengan cara tidak menyertakan data kolom Cabin dengan cara membuangnya. copy data kedalam variabel baru data\_1 dan panggil fungsi drop kolom

Pada fungsi drop parameter axis di set = 1. Apa maksudnya?

Text

Description automatically generated

Data yang digunakan untuk membangun model seharusnya berdistribusi normal. Cara mudah untuk mengidentifikasi distribusi data adalah dengan membandingkan kurva distribusi dengan kurva gaussian yang berbentuk lonceng seperti gambar dibawah ini. Jika bentuknya tidak mirip dengan kurva berbentuk lonceng maka distribusi datanya tidak normal. (SD : standard deviation).



Cara memvisualisasikan distribusi data dengan mudah dilakukan dengan fungsi histogram.

Chart, histogram

Description automatically generated

**Latihan 1.2.2**

Penjelasan Singkat:

Pada latihan ini praktikan diminta untuk mampu menangani missing value dengan data baru

Lakukan penanganan missing value pada variabel age dan salary sesuai langkahlangkah yang telah dipelajari dari data berikut :

1. Cara paling mudah untuk membuang nilai missing value adalah dengan memanggil fungsi dropna()

Text

Description automatically generated

Bandingkan dengan data sebelum dilakukan pemanggilan fungsi dropna().

Apa yang terjadi setelah pemanggilan fungsi dropna()?

Data Dihilangkan

1. Melakukan pemberian nilai pada data yang missing value (imputasi)

Ada beberapa cara yang dapat dilakukan untuk melakukan pemberian nilai missing value:

* Memberikan nilai konstan
* Nilai yang diambil secara acak dari kolom yang sama
* Mean, median atau mode dari kolom tersebut
* Diestimasi dari model prediksi lain

Graphical user interface, text

Description automatically generated

Perhatikan pada kolom Age, apakah masih ada data na? Perhatikan apa yang berubah dari data pada kolom Age?

Data Age Ter isi

1. Mengubah representasi data kolom Sex dan Embark.

Lakukan pengisian data missing value pada kolom Embark dengan salah satu nilai selain ‘nan’ terlebih dahulu.

Graphical user interface, text

Description automatically generated with medium confidence

Cek nilai unik pada masing-masing kolom, kemudian petakan data pada kolom tersebut menjadi angka.

Graphical user interface, text, application

Description automatically generated

Ubah representasi dari (‘S’ ⇒ 0, ‘C’ ⇒ 1, ‘Q’ ⇒ 2)

Ubah representasi male ⇒ 0 dan female ⇒ 1

Graphical user interface, text, application, Word

Description automatically generated

1. Setelah dilakukan beberapa preprocessing, tampilkan kembali beberapa baris data untuk mengecek ulang representasi data hasil proses cleaning

Graphical user interface, text, application, chat or text message

Description automatically generated

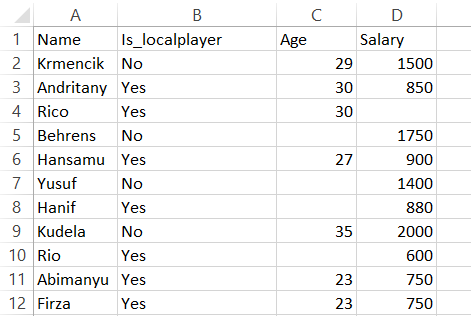
1. Nama dan Tiket bisa diabaikan dengan melakukan drop kolom. Lakukan drop kolom pada kedua kolom tersebut.

Text

Description automatically generated

**Latihan 1.2.3**

Lakukan penanganan missing value pada variabel Age dan Salary sesuai langkahlangkah yang telah dipelajari dari data berikut :



Output

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated



**CEK LIST**

|  |  |  |  |
| --- | --- | --- | --- |
| **Elemen Kompetensi** | **No Latihan** | **Penyelsaian** | |
| **Selsai** | **Tidak Selsai** |
| **1** | **1.1.1** |  |  |
| **2** | **1.2.1** |  |  |
| **1.2.2** |  |  |

**FORM UMPAN BALIK**

|  |  |  |  |
| --- | --- | --- | --- |
| **Elemen Kompetensi** | **Tingkat Kesulitan** | **Tingkat Ketertarikan** | **Waktu**  **Penyelesaian dalam menit** |
| Memahami data preprocessing menggunakan Python. | Sangat Mudah  Mudah   * Biasa   Sulit  Sangat Sulit | * Sangat Tertarik   Cukup Tertarik  Tertarik  Tidak Tertarik |  |
| Mengimplementasikan pre-processing data pada console Jupyter  Notebooks | Sangat Mudah  Mudah   * Biasa   Sulit  Sangat Sulit | * Sangat Tertarik   Cukup Tertarik  Tertarik  Tidak Tertarik |  |