# BAB 3 DATA PREPARATION

*Data preparation* merupakan tahap setelah dilakukan pengumpulan data awal yang telah dilakukan pada fase crisp-dm sebelumnya, yaitu *bussiness understanding*. Pada tahap *data preparation* ini, dilakukan proses menyiapkan data awal, memilih variabel yang akan dianalisis dan membersihkan data. Dalam pengerjaan proyek, bahasa pemrograman yang digunakan adalah pemrograman *python* dengan *software* pengolah data **Jupyter Notebook.**

## ***Package***

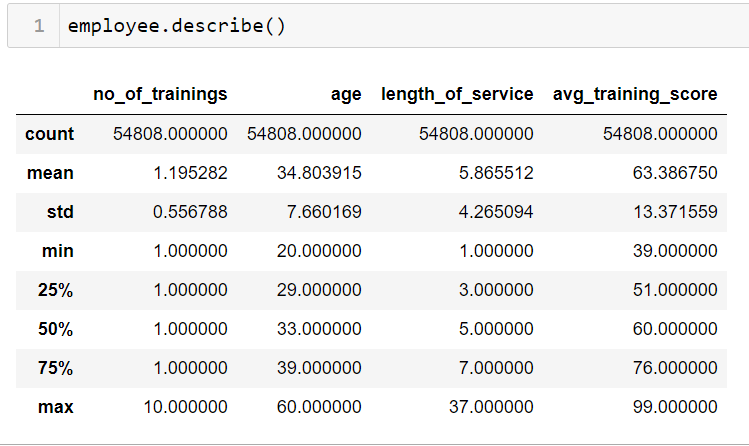
Untuk dapat menjalankan beberapa kode program yang akan dijalankan, dibutuhkan beberapa *package* yang harus diinstal, yaitu:

1. **Pandas**, untuk memuat sebuah file ke dalam tabel virtual seperti *spreadsheet*, mengumpulkan data, dan mengolahnya.
2. **Numpy**, untuk operasi verktor dan matriks serta analisis data.
3. **Mlxtend,** *library* untuk *machine learning* dalam menganalisis data.
4. **Matplotlib**, untuk menyajikan data ke dalam visual yang lebih menarik dan rapi.
5. **Seaborn**, untuk menyajikan data ke dalam visualisasi data secara statistik (dibangun di atas matpotlib.
6. **Apriori**, untuk memuat algoritma apriori.
7. **Association\_rules,** untuk membangun sebuah variabel yang memiliki *association rules* (aturan asosiasi).
8. **Image**, untuk menampilkan gambar.

## ***Dataset Description***

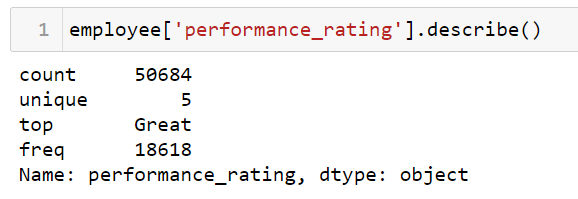
Pada fase ini, *dataset* akan dideskripsikan dengan memanfaatkan bahasa pemrograman python. Berikut beberapa fungsi yang dijalankan dalam mendeskripsikan *dataset* tersebut:

1. **.describe()**, untuk menampilkan berbagai ringkasan atau deskripsi statistik data, seperti jumlah data di setiap kolom (*count*), rata-rata nilai per kolom (*mean*), standard deviasi (*std*), nilai minimun (*min*), nilai maksimum (*max*), serta batas nilai dari masing-masing kuartil (25%, 50%, 75%). Berikut beberapa ringkasan atau deskripsi statistik data pada atribut yang bertipe data numerik.



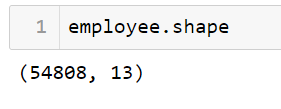
**Gambar 2. Fungsi .describe() pada Atribut bertipe data Numerik**

Jika atribut yang ingin diketahui informasi deskriptif data statistiknya tidak bertipe data numerik, maka informasi yang diberikan adalah *count* (jumlah data setiap kolom), *unique* (jumlah data yang unik), *top* (data dengan nilai paling banyak atau paling besar), dan *freq* (jumlah data yang paling banyak atau yang paling besar). Berikut penerapan fungsi .*describe()* pada atribut *performance\_rating*.



**Gambar 3. Fungsi .describe() pada Atribut performance\_rating**

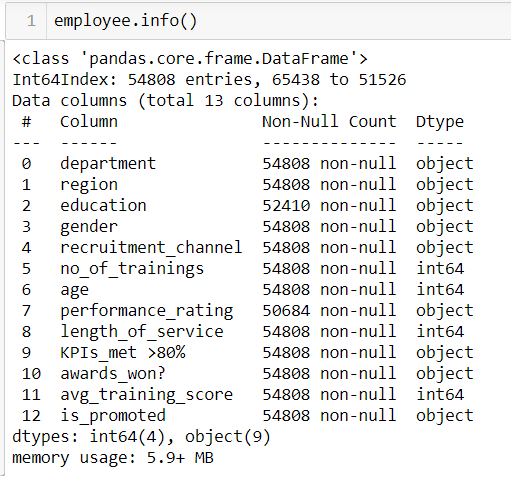
1. **.shape()**, untuk menampilkan dimensi data (jumlah baris dan kolom). Berikut keluaran data yang diperoleh dari fungsi .shape tersebut.



**Gambar 4. Fungsi .shape()**

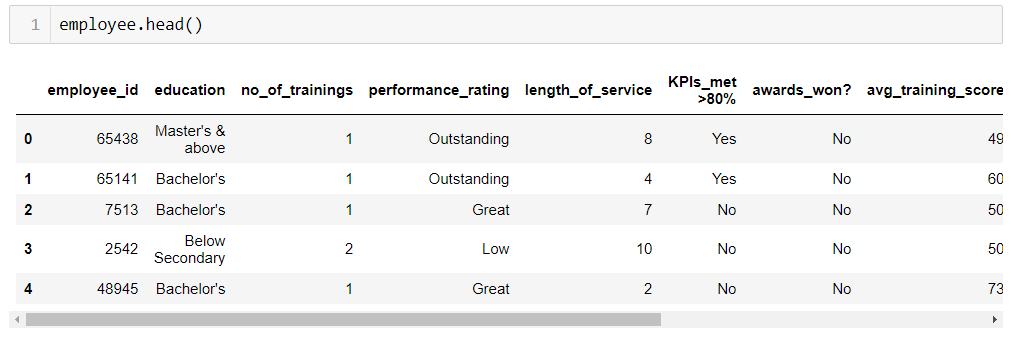
Pada gambar 4 diatas menunjukkan bahwa *dataset* yang digunakan memiliki 54808 baris dan 13 kolom.

1. **.info()**, untuk menampilkan gambaran mengenai *dataset*.



**Gambar 5. Fungsi .info()**

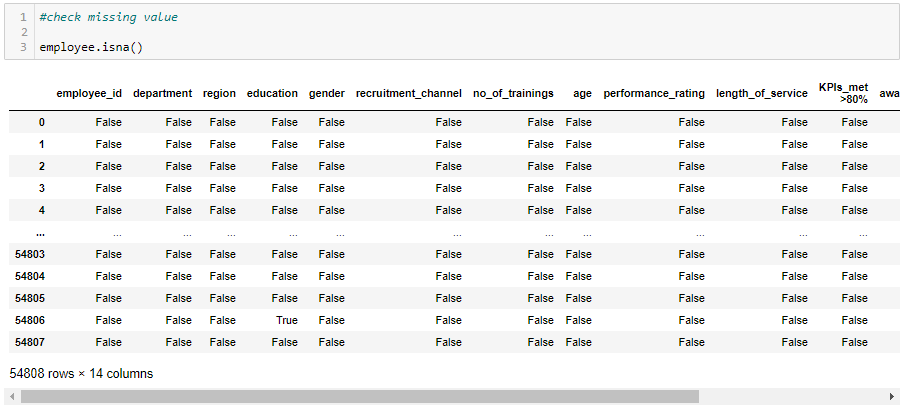
1. **.head()**, untuk melihat 5 sampel data teratas.



**Gambar 6. Fungsi .head()**

## ***Clean Data***

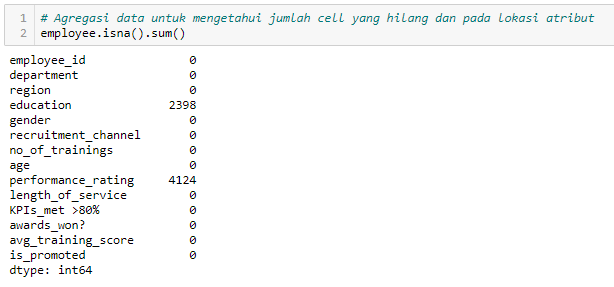
Pada fase ini dilakukan pembersihan data. *Data cleaning* yang dilakukan adalah dengan cara menghapus objek data yang tidak mengandung nilai (*missing value*).



**Gambar 7. Fungsi .isna()**

Fungsi **isna()** berguna untuk mengembalikan nilai *boolean* (True dan False). Jika *cell* berisi *value* “False”, maka artinya *cell* tersebut tidak mengandung *missing value* dan sebaliknya, jika *cell* berisi *value* “True”, maka *cell* tersebut mengandung *missing value*.

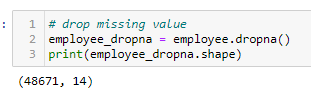
Untuk memudahkan dalam memahami data, maka perlu dilakukan agregasi data dengan fungsi **sum().** Dengan menggunakan fungsi **sum()**, maka akan diketahui berapa jumlah data yang *missing value* dan berasal dari atribut apa.



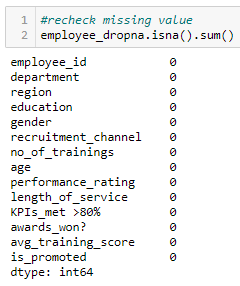
**Gambar 8. Fungsi .isna().sum()**

Dari gambar 8 diatas menunjukkan bahwa terdapat *missing value* pada atribut *education* (sebanyak 2398 data) dan atribut *performance\_rating* (sebanyak 4124 data).

Selanjutnya menggunakan fungsi .**dropna()** untuk menghilangkan data yang hilang. Setelah fungsi .**dropna()** dijalankan, maka data yang mengandung *missing value* terhapus. Pada gambar 9 menunjukkan bahwa baris (*record*) pada *dataset* telah berkurang.



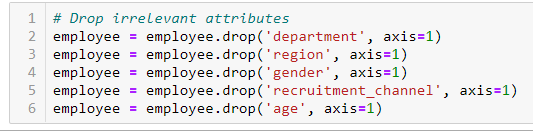
**Gambar 9. Fungsi .dropna()**



**Gambar 10. Pengecekkan Kembali Missing Value setelah Fungsi .dropna() dijalankan**

## ***Select Data (Data Reduction)***

Untuk meningkatkan efisiensi *data mining*, maka perlu memilih beberapa bagian dari *dataset* yang diperlukan sehingga *dataset* yang diproses lebih sedikit. Pada fase *select data* ini, yang dilakukan adalah mengurangi atribut yang tidak relevan. Atribut yang tidak relevan adalah atribut *department*, *region*, *gender*, *recruiment\_channel*, dan *age*.



**Gambar 11. Penghapusan Atribut yang Tidak Relevan**