**Kesimpulan Cleaning Data Proyek Kelas.id**

Data Sumber punya 8000, Cleaning data bermain di 5%

1. Data Duplikate ada 7 Data langsung dihapus
2. Missing Value terletak di

* nama\_pengguna 1 0,01% dihapus
* jenis\_kelamin 659 8,24% dilakukan imputasi nilai modus “male” dan mengganti nilai perempuan ke female, laki-laki ke famale
* bidang\_studi 5 0,06% dihapus
* skor\_akhir 1563 19,54% dilakukan imputasi ke nilai median “81.7”
* rating\_kelas 1040 13,006% dilakukan imputasi dengan nilai modus “ 5.0”

Missing value kurang dari 5% langsung dihapus, lebih dari 5% dilakukan Imputasi dengan menggunakan modus dan rata-rata. Setelah dilakukan pengecekan terhadap missing value dan melakukan Imputasi data tersisa sebanyak 7987 data.

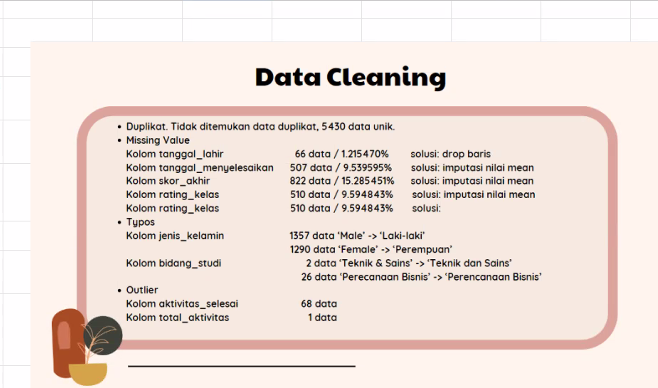
1. Irrelevant value, terdapat “laki-laki”,”perempuan”,”female”,”male” untuk kolom jenis kelamin, Kemudian diperbaiki untuk “female”, dan “male” saja. Lalu pada kolom bidang studi “Manajemen Bisnis”,”Manajemen Business”, dan “Manajemen & bisnis” kemudian diubah menjadi “ Bisnis Manajemen” saja.
2. Pengecekan Outliner, ditemukan 931 baris data dengan Outliner. Yang berarti persentase outliner sebesar 11,63% dari data. Data outliner diimputasi dengan nilai “81.7”
3. Mengecek nilai data dengan nilai tidak masuk akal, pada kolom tanggal lahir ditemukan data dengan tahun lahir

* 2023 6
* 2023 37
* 2021 64
* 2020 73
* 2019 3
* 2018 2
* 2017 1
* 2016 1
* 2014 3
* 2013 3
* 2011 1
* 2010 1
* 2009 1
* 2008 1
* 2007 1
* 2006 1
* Total : 199 data yang dianggap tidak masuk akal(masih diibawah umur) berarti 2,48% data ini bisa dihapus

1. Sisa data sebanyak 7789 baris data ke dalam file baru bernama data\_bersih.csv

perbaikan

* **data > 5% harus diimputasi (outlier,missing value)**
* **data anomali sebaiknya dibiarkan saja, dan kalo di kantor sebaiknya verifikasi data**

****