Pembersihan Data (Data Cleaning) dan Transformasi Data (Data Transformation)

Data yang Anda kumpulkan dari dunia nyata seringkali tidak sempurna. Data tersebut mungkin mengandung kesalahan, nilai yang hilang, format yang tidak konsisten, atau outlier (nilai ekstrim). Oleh karena itu, sebelum data dapat digunakan untuk analisis atau pemodelan, data tersebut perlu dibersihkan dan ditransformasi.

# Pembersihan Data (Data Cleaning)

Pembersihan data adalah proses mengidentifikasi dan memperbaiki kesalahan, inkonsistensi, dan ketidaklengkapan dalam data. Tujuannya adalah untuk menghasilkan data yang bersih, akurat, dan konsisten. Beberapa langkah umum dalam pembersihan data meliputi:

1. Penanganan Nilai yang Hilang (Missing Values):

* Menghapus baris atau kolom yang memiliki banyak nilai yang hilang.
* Mengisi nilai yang hilang dengan nilai yang masuk akal (misalnya, mean, median, atau modus).
* Menggunakan teknik imputasi yang lebih canggih (misalnya, regresi atau machine learning).

1. Penanganan Duplikat:

* Mengidentifikasi dan menghapus data yang terduplikasi.

1. Penanganan Outlier:

* Mendeteksi outlier menggunakan metode statistik atau visualisasi.
* Menghapus atau mengganti outlier dengan nilai yang lebih representatif.

1. Standarisasi Format Data:

* Memastikan semua data memiliki format yang konsisten (misalnya, tanggal, waktu, mata uang).

# Transformasi Data (Data Transformation)

Transformasi data adalah proses mengubah data dari satu bentuk ke bentuk lain. Tujuannya adalah untuk membuat data lebih sesuai untuk analisis atau pemodelan. Beberapa teknik transformasi data yang umum meliputi:

1. Normalisasi: Menskalakan data ke rentang tertentu (misalnya, 0 hingga 1) agar variabel dengan skala berbeda dapat dibandingkan.
2. Encoding: Mengubah data kategorikal menjadi bentuk numerik yang dapat digunakan oleh model machine learning.
3. Agregasi: Menggabungkan data dari beberapa sumber atau kolom menjadi satu.
4. Pembuatan Fitur (Feature Engineering): Membuat fitur baru dari data yang ada untuk meningkatkan kinerja model.

# Pentingnya Pembersihan dan Transformasi Data

Pembersihan dan Transformasi data adalah langkah-langkah yang sangat penting dalam proses Data Science. Data yang bersih dan terstruktur dengan baik akan menghasilkan analisis yang lebih akurat dan model yang lebih andal.

### Tugas:

Anda akan bekerja dengan dataset “Titanic” yang dapat diunduh dari kaggle:

[Titanic](https://www.kaggle.com/datasets/heptapod/titanic)

Dataset ini berisi informasi tentang penumpang kapal Titanic, termasuk apakah mereka selamat atau tidak.

### Tugas:

1. Eksplorasi Data: Gunakan Visual Studio Code untuk membuka dataset Titanic (dalam format CSV). Amati kolom-kolom yang ada dan jenis data dari masing-masing kolom.
2. Identifikasi Masalah Kualitas Data: Identifikasi setidaknya 3 masalah kualitas data (misalnya, nilai yang hilang, duplikat, outlier, format yang tidak konsisten) yang ada dalam dataset. Jelaskan bagaimana Anda mengidentifikasi masalah-masalah tersebut.

#### Catatan:

* Jangan melakukan pembersihan atau transformasi data pada tahap ini. Tugas ini hanya untuk mengidentifikasi masalah yang ada.
* Anda dapat menggunakan pustaka Pandas di python untuk membantu Anda dalam eksplorasi data.

**Selamat mengerjakan tugas!**

### Tugas:

1. Terdapat 28 kolom dan tipe data yaitu: Passengerid: int64, Age: float64, Fare: float64, Sex: int64, sibsp: int64, zero: int64, zero.1: int64, zero.2: int64, zero.3: int64, zero.4: int64, zero.5: int64, zero.6: int64, Parch: int64, zero.7: int64, zero.8: int64, zero.9: int64, zero.10: int64, zero.11: int64, zero.12: int64, zero.13: int64, zero.14: int64, Pclass: int64, zero.15: int64, zero.16: int64, Embarked: float64, zero.17: int64, zero.18: int64, 2urvived: int64. dtype: object
2. Mengidentifikasi Nilai yang Hilang:

missing\_values\_df = titanic\_df

missing\_values\_df.isnull().sum()

Mengidentifikasi Duplikat:

duplicated\_df = titanic\_df

duplicated\_df.duplicated().sum()

Mengidentifikasi Outliers:

Q1 = titanic\_df.quantile(0.25)

Q3 = titanic\_df.quantile(0.75)

IQR = Q3 - Q1

outliers = ((titanic\_df < (Q1 - 1.5 \* IQR)) | (titanic\_df > (Q3 + 1.5 \* 1.5))).sum()

### Tugas Tambahan (Opsional):

Jika Anda ingin memperdalam pemahaman, coba lakukan hal berikut:

1. Visualisasikan distribusi data dari beberapa kolom numerik (misalnya, Age, Fare) menggunakan histogram atau boxplot. Apakah Anda melihat adanya outlier?
2. Coba identifikasi outlier menggunakan metode lain, seperti Z-score. Bandingkan hasilnya dengan metode IQR.