Pengolahan Data Kategorikal dan Numerikal (Encoding dan Scaling)

Setelah memahami penting nya membersihkan data, sekarang kita akan membahas bagaimana mengolah data kategorikal dan numerikal agar siap digunakan dalam analisis dan pemodelan.

# Pengolahan Data Kategorikal (Encoding)

Data kategorikal adalah data yang merepresentasikan kategori atau kelompok. Contohnya termasuk jenis kelamin (pria/wanita), warna(merah/hijau/biru), atau tingkat pendidikan(SD/SMP/SMA).

## Mengapa Perlu Encoding?

Kebanyakan algoritma machine learning hanya dapat bekerja dengan data numerik. Oleh karena itu, kita perlu mengubah data kategorikal menjadi bentuk numerik agar bisa digunakan dalam analisis.

## Teknik Encoding:

* Label Encoding: Setiap kategori unik diberi label numerik (misalnya, 0, 1, 2).
* One-Hot Encoding: Setiap kategori diubah menjadi kolom baru dengan biner (0 atau 1).
* Ordinal Encoding: Kategori yang memiliki urutan alami diberi label numerik yang sesuai dengan urutannya(misalnya, kecil=1, sedang=2, besar=3).

# Pengolahan Data Numerikal (Scaling)

Data numerikal adalah data yang merepresentasikan kuantitas atau angka. Contohnya termasuk usia, suhu, atau pendapatan.

## Mengapa Perlu Scaling?

Beberapa algoritma machine learning sensitif terhadap skala data. Jika variabel memiliki rentang nilai yang sangat berbeda, variabel dengan rentang nilai yang lebih besar dapat mendominasi analisis. Scaling bertujuan untuk membawa semua variabel ke skala yang sama.

## Teknik Scaling:

* Standardisasi (Standardization): Mengubah data sehingga memiliki rata-rata 0 dan standar deviasi 1.
* Normalisasi (Normalization): Menskalakan data ke rentang tertentu (biasanya 0 hingga 1).

### **Tugas Hari 6:**

1. Label Encoding: Pada dataset Titanic, terapkan Label Encoding pada kolom sex. Tampilkan 5 baris pertama dari dataset encoding.
2. One-Hot Encoding: Pada dataset Titanic, terapkan One-Hot Encoding pada kolom Embarked. Tampilkan 5 baris pertama dari dataset setelah encoding.
3. Standardisasi: Pada dataset Titanic, standardisasikan kolom Age. Tampilkan 5 baris pertama dari dataset setelah scaling.

**Contoh Kode (Pandas dan Scikit-Learn):**

import pandas as pd

from sklearn.preprocessing import LabelEncoder, OneHotEncoder, StandardScaler

# ... (kode untuk membaca dataset Titanic)

# Label Encoding

le = LabelEncoder()

titanic\_df['Sex'] = le.fit\_transform(titanic\_df['Sex'])

# One-Hot Encoding

ohe = OneHotEncoder()

embarked\_encoded = ohe.fit\_transform(titanic\_df[['Embarked']]).toarray()

embarked\_df = pd.DataFrame(embarked\_encoded, columns=ohe.get\_feature\_names\_out(['Embarked']))

titanic\_df = pd.concat([titanic\_df, embarked\_df], axis=1).drop('Embarked', axis=1)

# Standardisasi

scaler = StandardScaler()

titanic\_df['Age'] = scaler.fit\_transform(titanic\_df[['Age']])

print(titanic\_df.head().to\_markdown(index=False, numalign="left", stralign="left"))

#### **Catatan:** Pastikan Anda telah menginstall library scikit-learn.

### **Tugas:**

1. Label Encoding:

import pandas as pd

from sklearn.preprocessing import LabelEncoder, OneHotEncoder, StandardScaler

# PERLU INSTALL 'tabulate'. contoh instalasi:

# pip install tabulate

le = LabelEncoder()

titanic\_df['Sex'] = le.fit\_transform(titanic\_df['Sex'])

print(titanic\_df['Sex'].head().to\_markdown(numalign="left", stralign="left"))

1. One-Hot Encoding:

# Mengatasi nilai yang hilang di kolom 'Embarked'

titanic\_df['Embarked'].fillna('Unknown', inplace=True)

# Pastikan semua nilai dalam kolom 'Embarked' adalah string

titanic\_df['Embarked'] = titanic\_df['Embarked'].astype(str)

# One-Hot Encoding

ohe = OneHotEncoder(handle\_unknown='ignore', sparse\_output=False)

embarked\_encoded = ohe.fit\_transform(titanic\_df[['Embarked']])

embarked\_df = pd.DataFrame(embarked\_encoded, columns=ohe.get\_feature\_names\_out(['Embarked']))

# Menggabungkan DataFrame hasil encoding dengan DataFrame asli

titanic\_df = pd.concat([titanic\_df, embarked\_df], axis=1).drop('Embarked', axis=1)

print(titanic\_df.head().to\_markdown(index=False, numalign="left", stralign="left"))

1. Standardisasi:

from sklearn.preprocessing import StandardScaler

# Mengatasi nilai yang hilang di kolom 'Age'

titanic\_df['Age'].fillna(titanic\_df['Age'].mean(), inplace=True)

# Standardisasi

scaler = StandardScaler()

titanic\_df['Age'] = scaler.fit\_transform(titanic\_df[['Age']])

print(titanic\_df.head().to\_markdown(index=False, numalign="left", stralign="left"))