Pembagian DataSet menjadi Data Latih(Training Set) & Data Uji (Test Set)

Sebelum membangun model machine learning, kita perlu membagi dataset kita menjadi dua bagian:

1. Data Latih (Training Set): Bagian ini digunakan untuk melatih model, yaitu untuk menemukan pola dan hubungan antara fitur dan variabel target.
2. Data Uji (Test Set): Bagian ini digunakan untuk mengevaluasi kinerja model setelah dilatih, yaitu untuk mengukur seberapa baik model dapat membuat prediksi pada data yang belum pernah dilihatnya.

# Mengapa Perlu Membagi Dataset?

Pembagian dataset sangat penting untuk menghindari overfitting. Overfitting terjadi ketika model terlalu menyesuaikan diri dengan data latih sehingga tidak dapat melakukan generalisasi dengan baik pada data baru. Dengan membagi dataset, kita dapat menguji model pada data yang belum pernah dilihatnya untuk mendapatkan estimasi kinerja yang lebih realistis.

# Proporsi Pembagian

Proporsi pembagian data latih dan data uji yang umum digunakan adalah 70:30 atau 80:20. Artinya, 70% atau 80% data digunakan untuk pelatihan, dan sisanya 30% atau 20% digunakan untuk pengujian. Proporsi ini dapat disesuaikan tergantung pada ukuran dataset dan kompleksitas model yang akan dibangun.

# Teknik Pembagian

Ada beberapa teknik pembagian dataset, antara lain:

* Random Split: Membagi data secara acak menjadi data latih dan data uji.
* Stratified Split: Membagi data secara acak dengan tetap menjaga proporsi kelas (untuk masalah klasifikasi).
* Time-Based Split: Membagi data berdasarkan waktu (misalnya, data sebelum tanggal tertentu untuk data latih, dan data setelah tanggal tersebut untuk data uji).

# Pembagian Dataset dengan Scikit-Learn

Scikit-learn menyediakan fungsi train\_test\_split untuk memudahkan pembagian dataset.

## Tugas Hari 8:

1. Pembagian Dataset: Gunakan fungsi train\_test\_split dari scikit-learn untuk membagi dataset Titanic menjadi data latih (70%) dan data uji (30%).
2. Verifikasi: Tampilkan dimensi (jumlah baris dan kolom) dari data latih dan data uji untuk memastikan pembagian sudah benar.

### **Contoh Kode (Scikit-Learn):**

from sklearn.model\_selection import train\_test\_split

# ... (kode untuk membaca dan mengolah dataset Titanic)

# Memisahkan fitur (X) dan target (y)

X = titanic\_df.drop('Survived', axis=1) # Fitur-fitur

y = titanic\_df['Survived'] # Target

# Pembagian dataset

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Verifikasi dimensi

print("Dimensi X\_train:", X\_train.shape)

print("Dimensi X\_test:", X\_test.shape)

print("Dimensi y\_train:", y\_train.shape)

print("Dimensi y\_test:", y\_test.shape)

**Catatan:**

* random\_state=42 : adalah nilai untuk generator bilangan acak. Dengan menggunakan nilai seed yang sama, Anda akan mendapatkan hasil pembagian dataset yang sama setiap kali Anda Menjalankan Kode.

**Selamat Mengerjakan Tugas! 😂**

### Tugas:

1. Pembagian Dataset:

from sklearn.model\_selection import train\_test\_split

# # Memisahkan fitur (X) dan target (Y)

X = titanic\_df.drop('2urvived', axis=1) # Fitur-fitur

y = titanic\_df['2urvived'] # Target

# # Pembagian dataset

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# print(titanic\_df.columns)

1. Verifikasi:

print("Dimensi X\_train: ", X\_train.shape)

print("Dimensi X\_test: ", X\_test.shape)

print("Dimensi y\_train: ", y\_train.shape)

print("Dimensi y\_test: ", y\_test.shape)