Feature Engineering dan Feature Selection

Setelah data Anda bersih dan terstruktur dengan baik, langkah selanjutnya adalah mempersiapkan fitur-fitur yang akan digunakan dalam model machine learning. Ini adalah langkah penting yang disebut feature engineering dan feature selection.

# Feature Engineering

Feature engineering adalah proses mengubah data mentah menjadi fitur yang lebih informatif dan relevan untuk model machine learning. Tujuannya adalah untuk meningkatkan kinerja model dengan mengekstrak informasi yang paling berguna dari data.

## Teknik Feature Engineering:

* Pembuatan Fitur Baru: Menggabungkan atau mengubah fitur yang ada untuk membuat fitur baru yang lebih relevan.
* Encoding Kategorikal: Mengubah fitur kategorikal menjadi bentuk numerik(seperti yang telah kita bahas sebelumnya)
* Scaling Numerikal: Menskalakan fitur numerik ke rentang yang sama (seperti yang telah kita bahas sebelumnya)
* Ekstraksi Fitur dari Teks: Mengubah teks menjadi representasi numerik (misalnya, menggunakan bag-of-words atau word embeddings)
* Ekstraksi Fitur dari Gambar: Mengubah gambar menjadi representasi numerik (misalnya, menggunakan histogram warna atau fitur HOG)

# Feature Selection

Feature selection adalah proses memilih subset fitur yang paling relevan untuk digunakan dalam model machine learning. Tujuannya adalah untuk mengurangi dimensi data dan meningkatkan kinerja model dengan menghilangkan fitur yang tidak relevan atau redundan.

## Teknik Feature Selection:

* Filter Methods: Memilih fitur berdasarkan statistik univariat (misalnya, chi-squared test, ANOVA F-value).
* Wrapper Methods: Memilih fitur berdasarkan kinerja model (misalnya, recursive feature elimination).
* Embedded Methods: Memilih fitur selama proses pelatihan model (misalnya, LASSO regression).

### Tugas Hari 7:

Lanjutkan dengan dataset Titanic.

### Tugas:

1. Feature Engineering:

* Buat fitur baru bernama FamilySize yang merupakan jumlah dari SibSp (jumlah saudara kandung/pasangan di kapal) dan Parch (jumlah orang tua/anak di kapal) ditambah 1 (untuk penumpang itu sendiri).
* Buat fitur baru bernama IsAlone yang bernilai 1 jika FamilySize sama dengan 1, dan 0 jika tidak.

1. Feature Selection (Opsional):

* Jika Anda ingin feature selection, Anda bisa menggunakan metode filter seperti chi-squared test (untuk fitur kategorikal) atau ANOVA F-value (untuk fitur numerik) untuk memilih yang paling relevan dengan variabel target (Survived).

### Tugas:

1. Feature Engineering:

titanic\_df['FamilySize'] = titanic\_df['sibsp'] + titanic\_df['Parch'] + 1

titanic\_df['IsAlone'] = 1

titanic\_df.loc[titanic\_df['FamilySize'] > 1, 'IsAlone'] = 0

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

1. Feature Selection(***Opsional***):