Validasi Silang (Cross-Validation) dan pemilihan Model dalam Regresi

Evaluasi model yang telah kita lakukan sebelumnya, seperti menghitung MAE, MSE, RMSE, dan R-squared, dilakukan pada data yang sama yang digunakan untuk melatih model. Hal ini dapat menyebabkan masalah overfitting, dimana model berkinerja baik pada data latih tetapi buruk pada data baru.

Untuk mengatasi masalah ini, kita dapat menggunakan teknik validasi silang (cross-validation) dan pemilihan model.

# Validasi Silang (Cross-Validation)

Validasi silang adalah teknik untuk mengevaluasi kinerja model dengan cara membagi dataset menjadi beberapa bagian (fold), kemudian melatih model pada sebagian fold dan mengujinya pada fold yang lain. Proses ini diulang beberapa kali, dan hasil evaluasi pada setiap fold dirata-ratakan untuk mendapatkan estimasi kinerja model yang lebih baik.

## Jenis-jenis Cross-Validation:

* **K-Fold Cross-Validation:** Membagi dataset menjadi K fold.
* **Leave-One-Out Cross-Validation (LOOCV):** Setiap sampel data menjadi satu fold.
* **Stratified K-Fold Cross-Validation:** Memastikan proporsi kelas yang sama di setia[ fold (untuk masalah klasifikasi).

## keuntungan Cross-Validation:

* **Mengurangi Overfitting:** Memberikan estimasi kinerja model yang lebih realistis karena model diuji pada data yang belum pernah dilihatnya.
* **Memanfaatkan Data dengan Lebih Baik:** Memungkinkan kita untuk menggunakan semua data yang tersedia untuk pelatihan dan pengujian model.

# Pemilihan Model

Pemilihan model adalah proses memilih model terbaik dari beberapa model yang berbeda. Model terbaik adalah model yang memberikan kinerja terbaik pada matrik evaluasi yang relevan.

## Teknik Pemilihan Model:

* **Grid Search:** mencoba semua kombinasi hyperparameter yang mungkin.
* **Randomized Search:** Mencoba kombinasi hyperparameter secara acak.
* **Bayesian Optimization:** Mencari hyperparameter optimal secara lebih efisien dengan menggunakan model probabilistik.

**Tugas Hari 19**

Melanjutkan dengan dataset “Advertising” dan model regresi linear yang telah di unduh.

**Tugas:**

1. **K-Fold Cross-Validation:** Lakukan 5-fold cross-validation pada model regresi linear Anda. Hitung rata-rata dan standar deviasi dari MAE, MSE, dan RMSE pada setiap fold.
2. **Pemilihan Model:**

* Buat beberapa model regresi linear dengan kombinasi fitur yang berbeda (misalnya, hanya menggunakan **TV,** hanya menggunakan **Radio,** dan seterusnya).
* Lakukan 5-fold cross-validation pada setiap model.
* Bandingkan kinerja model-model tersebut berdasarkan rata-rata MAE. Pilih model terbaik berdasarkan rata-rata MAE terendah.

**Contoh Kode (Scikit-Learn)**

from sklearn.model\_selection import cross\_val\_score

# ... (kode untuk membaca, mengolah data, dan membuat model regresi linear)

# Cross-validation

scores = cross\_val\_score(model, X, y, scoring='neg\_mean\_absolute\_error', cv=5)

print(f'MAE (5-fold cross-validation): {-scores.mean():.2f} (+/- {scores.std():.2f})')

**Catatan:**

* Fungsi cross\_val\_score mengembalikan skor negatif karena scikit-learn menganggap skor yang lebih tinggi lebih baik. Oleh karena itu, kita perlu menambahkan tanda negatif (-scores.mean() ) untuk mendapatkan nilai MAE yang positif.

**Selamat Mengerjakan Tugas! 😀**

**Tugas:**

1. **K-Fold Cross-Validation:**

import pandas as pd

import numpy as np

# Misalkan X adalah dataframe pandas

# Langkah 1: Identifikasi kolom dengan nilai string

print(X.dtypes)

# Langkah 2: Transformasi atau hapus nilai string

# Contoh transformasi: mengubah nilai 'Mega' menjadi 1e6

X = X.replace('Mega', 1e6)

# Alternatif: jika kolom berisi nilai string dan tidak dapat ditransformasi, bisa menghapus kolom tersebut

# X = X.drop(columns=['nama\_kolom\_yang\_mengandung\_string'])

# Jika masih ada kolom dengan tipe objek, ubah atau drop

for col in X.select\_dtypes(include=['object']).columns:

# Jika memungkinkan untuk transformasi

# X[col] = X[col].astype(float) # ini jika bisa diubah ke float

# Jika tidak bisa diubah ke float, drop kolom

X = X.drop(columns=[col])

# Langkah 3: Cek nilai null atau kosong

print(X.isnull().sum())

X = X.dropna()

# Melanjutkan ke cross-validation

from sklearn.model\_selection import cross\_val\_score

scores = cross\_val\_score(model, X, y, scoring='neg\_mean\_absolute\_error', cv=5)

print(f'MAE (5-fold cross-validation): {-scores.mean():.2f} (+/-{scores.std():.2f})')

1. **Pemilihan Model:**