

# **TUGAS BESAR PEMBELAJARAN MESIN**

**Kelas IF-46-12 Kelompok 69**

# **ANGGOTA KELOMPOK**

- Alom Samudra - 1301220116
- Muhammad Rifqy Khuzaini - 1301223473
- Moza Qonita Budiyono - 1301220378

# PENDAHULUAN

Penelitian ini bertujuan untuk mengembangkan sistem prediksi rating hotel yang akurat menggunakan algoritma machine learning, khususnya K-Nearest Neighbors. Kami akan membandingkan tiga skema KNN yang berbeda untuk menemukan pendekatan terbaik dalam memprediksi rating hotel berdasarkan karakteristik hotel seperti harga, skor, jumlah review, dan lokasi.

Metodologi yang kami gunakan adalah pendekatan supervised learning dengan membandingkan tiga konfigurasi KNN yang berbeda. Setiap skema akan dievaluasi menggunakan multiple metrics untuk memastikan robustness dan reliability dari model yang dikembangkan.



# DATASET

## DATASET: GLOBAL HOTELS

### Karakteristik Dataset:

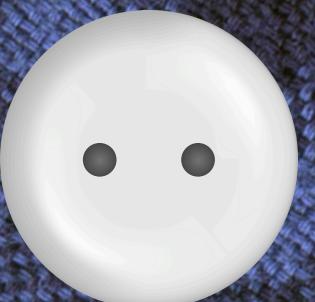
- Jumlah Data: 1000+ record hotel
- Format: CSV file
- Cakupan: Hotel dari berbagai kota dan negara

Dataset terdiri dari 8 kolom utama dengan lebih dari 1000 record hotel. Kolom Hotel\_Name berisi nama-nama hotel dari berbagai brand dan kategori. Kolom Rating merupakan target variable yang ingin kita prediksi, berisi kategori rating seperti "Excellent", "Very Good", "Good", "Fair", dan "Poor". Kolom Score berisi nilai numerik dari 0 hingga 10 yang merepresentasikan skor hotel berdasarkan review pengguna.



# DATA PROCESSING

X  
X  
X  
X



## Data Cleaning

Identifikasi dan penghapusan missing values (jumlah kecil, tidak mempengaruhi representativeness dataset)

## Data Transformation

- Konversi Number\_Reviews dari string dengan koma separator ke integer
- Konversi Score dan Price ke format numerik untuk konsistensi tipe data

# DATA PROCESSING



X  
X  
X  
X

## Encoding

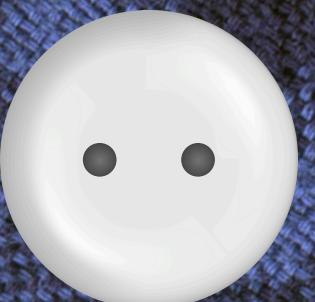
- Label Encoding untuk variabel kategorikal (City, Country, Rating) menjadi numerik
- Dipilih karena efisien dan cocok untuk algoritma KNN berbasis distance calculation

## Feature Scaling

- StandardScaler untuk normalisasi semua fitur numerik
- Penting untuk KNN karena sensitif terhadap skala data
- Mencegah dominasi fitur dengan range besar (Price) terhadap range kecil (Score)

# DATA PROCESSING

X  
X  
X  
X



## Data Splitting

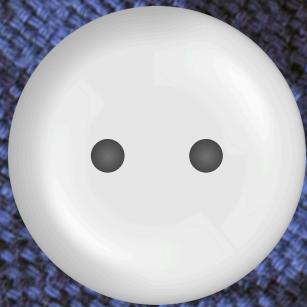
- Rasio 80:20 untuk training dan testing set
- Stratified sampling untuk menjaga distribusi rating proporsional
- Random state untuk reproducibility

## Validasi

- Cek distribusi data post-transformasi
- Verifikasi tidak ada data leakage
- Konfirmasi semua transformasi diterapkan dengan benar

# DATA PROCESSING

X  
X  
X  
X



Hasil

Dataset bersih dengan 800+ record siap untuk modeling.

# SKEMA MODEL KNN



KNN Baseline Model

● BASELINE MODEL: KNN DEFAULT  
Karakteristik: Parameter default scikit-learn (K=5, uniform weights, Minkowski p=2)

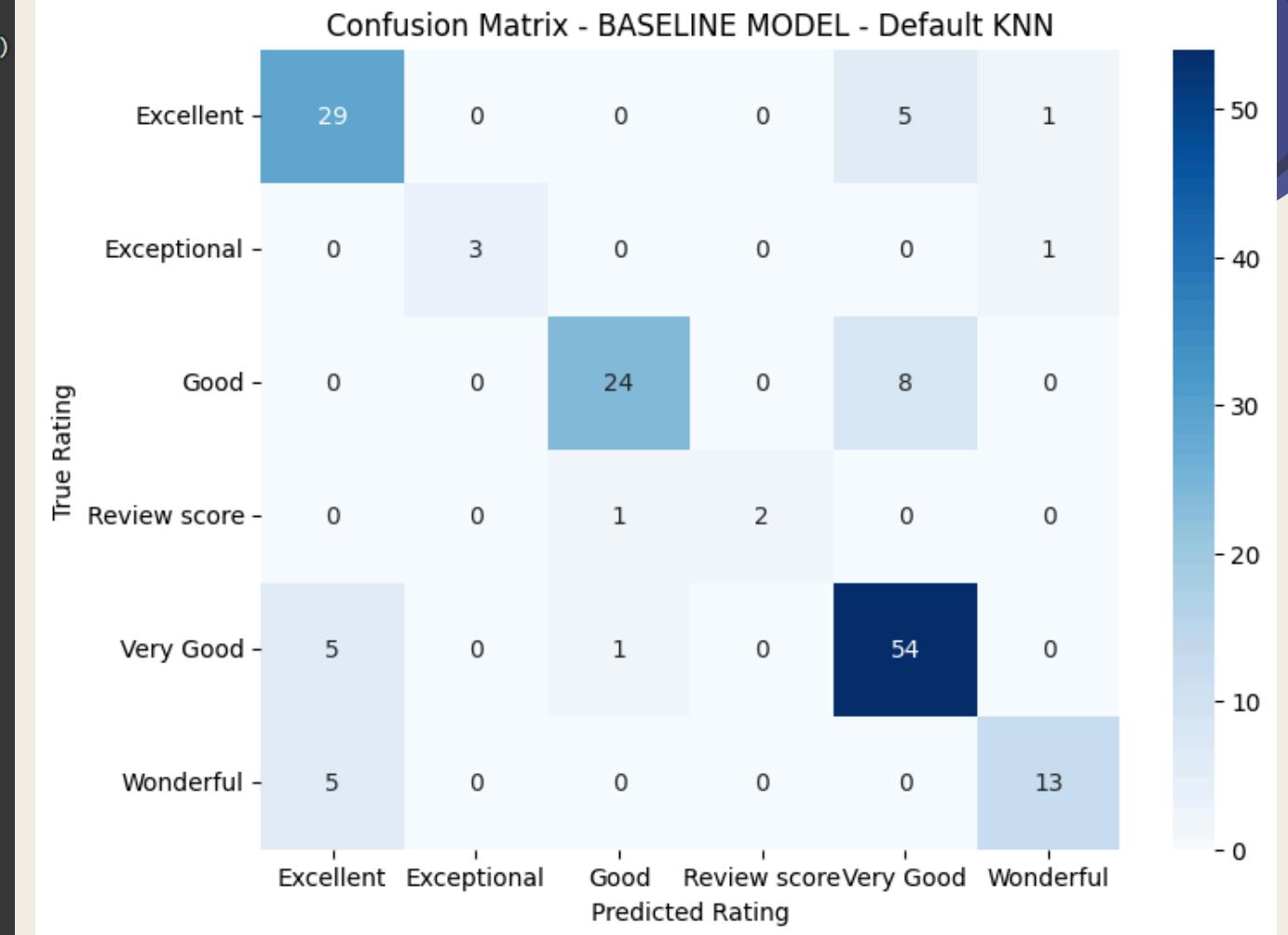
EVALUASI BASELINE MODEL - Default KNN

Parameter yang digunakan:  
n\_neighbors: 5  
weights: uniform  
metric: minkowski  
p: 2  
algorithm: auto

Hasil Evaluasi:  
Cross-Validation Score: 0.7884 (+/- 0.0426)  
Train Accuracy: 0.8942  
Test Accuracy: 0.8224  
F1-Score (weighted): 0.8220  
Overfitting Gap: 0.0718

Classification Report:

	precision	recall	f1-score	support
Excellent	0.74	0.83	0.78	35
Exceptional	1.00	0.75	0.86	4
Good	0.92	0.75	0.83	32
Review score	1.00	0.67	0.80	3
Very Good	0.81	0.90	0.85	60
Wonderful	0.87	0.72	0.79	18
accuracy			0.82	152
macro avg	0.89	0.77	0.82	152
weighted avg	0.83	0.82	0.82	152



# SKEMA MODEL KNN



KNN Model 1

● SKEMA 1: KNN KLASIK  
Karakteristik: K kecil, uniform weights, Euclidean distance  
Tujuan: Menangkap pola lokal dengan sensitivity tinggi

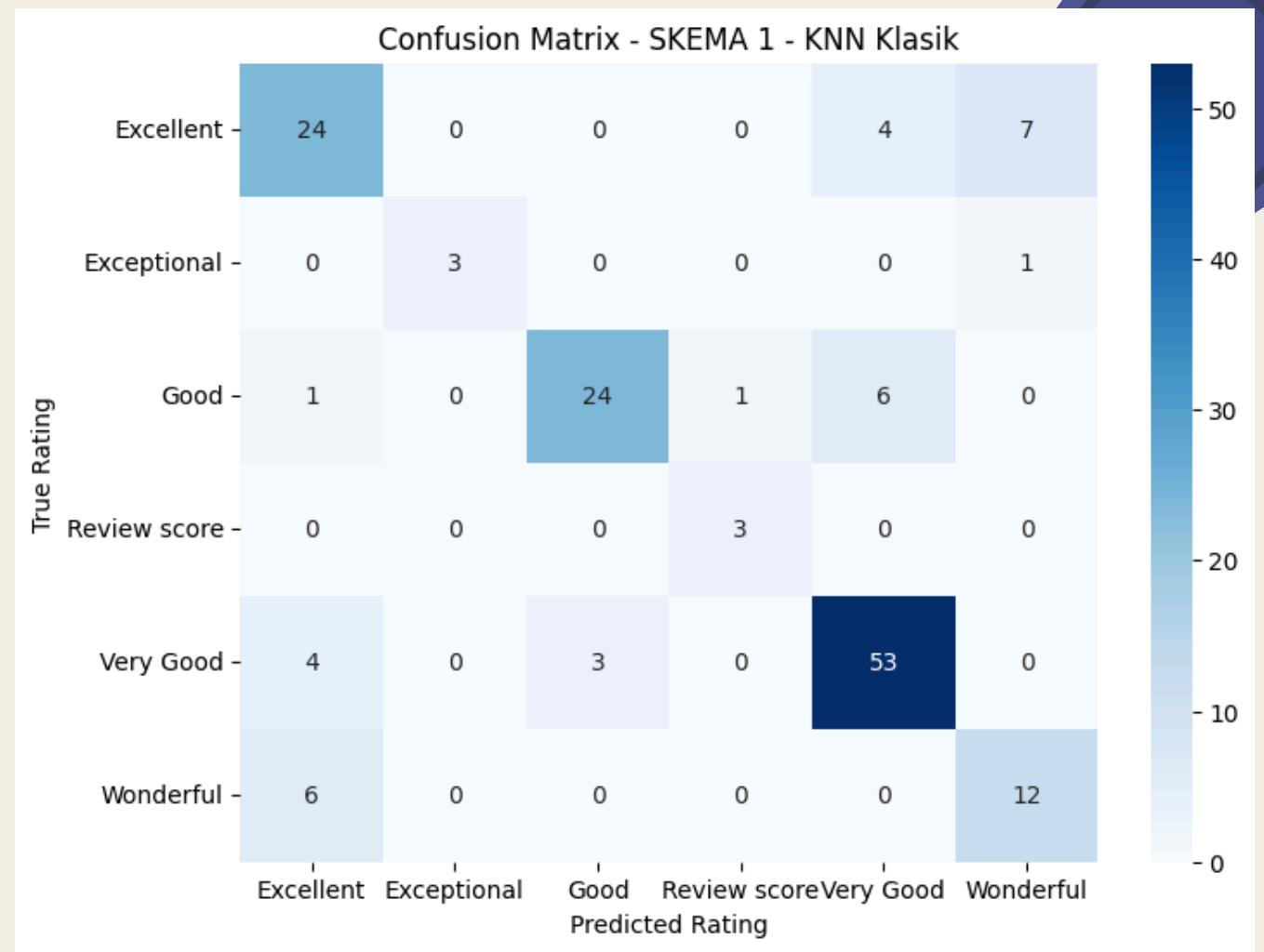
-----  
EVALUASI SKEMA 1 - KNN Klasik

Parameter yang digunakan:  
n\_neighbors: 3  
weights: uniform  
metric: euclidean  
algorithm: auto

Hasil Evaluasi:  
Cross-Validation Score: 0.7719 (+/- 0.0324)  
Train Accuracy: 0.9107  
Test Accuracy: 0.7829  
F1-Score (weighted): 0.7836  
Overfitting Gap: 0.1278

Classification Report:

	precision	recall	f1-score	support
Excellent	0.69	0.69	0.69	35
Exceptional	1.00	0.75	0.86	4
Good	0.89	0.75	0.81	32
Review score	0.75	1.00	0.86	3
Very Good	0.84	0.88	0.86	60
Wonderful	0.60	0.67	0.63	18
accuracy			0.78	152
macro avg	0.79	0.79	0.78	152
weighted avg	0.79	0.78	0.78	152



# SKEMA MODEL KNN



KNN Model 2

SKEMA 2: KNN DENGAN DISTANCE WEIGHTING  
Karakteristik: K sedang, distance weights, Manhattan distance  
Tujuan: Balance antara sensitivity dan robustness

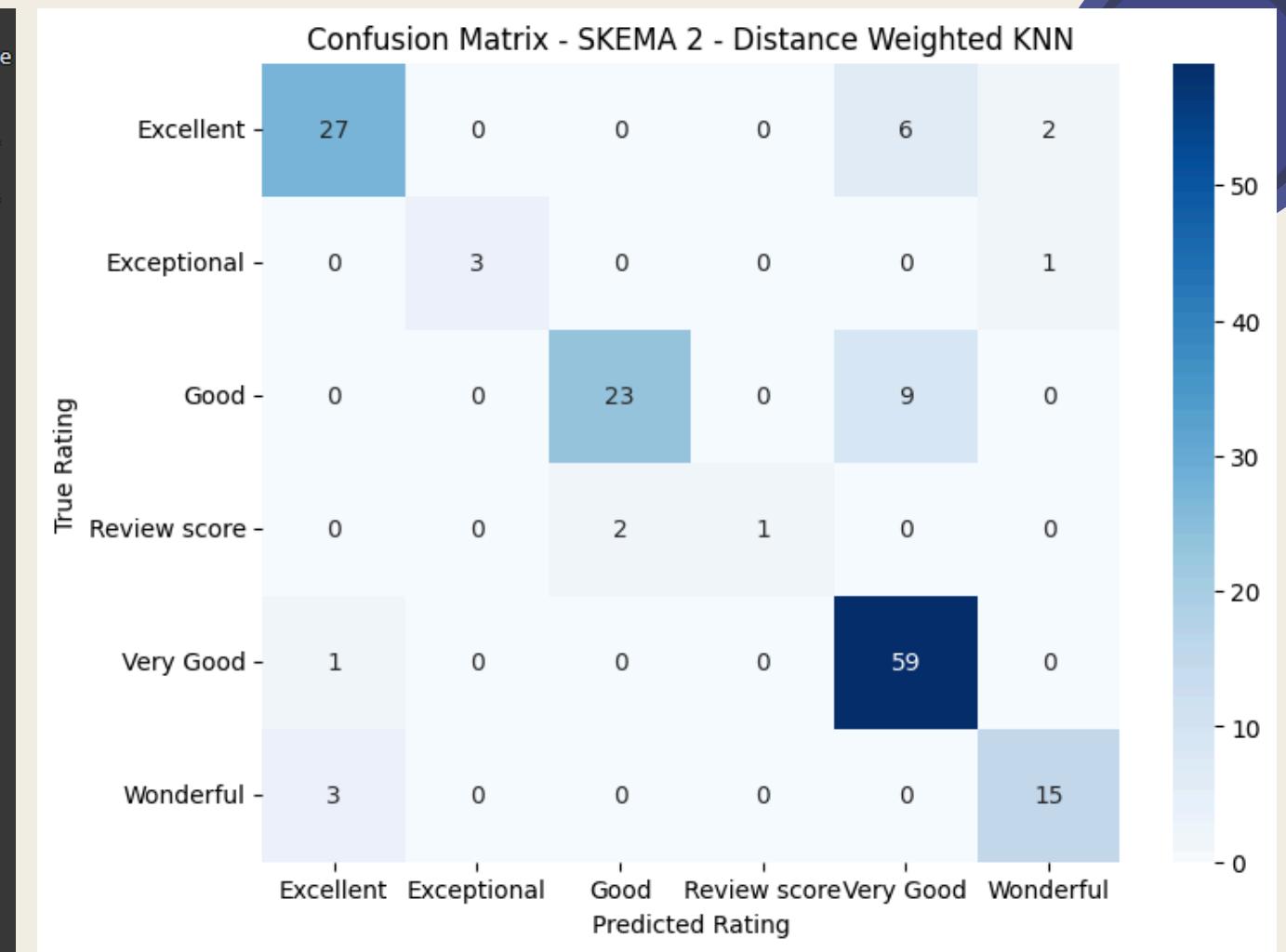
EVALUASI SKEMA 2 - Distance Weighted KNN

Parameter yang digunakan:  
n\_neighbors: 7  
weights: distance  
metric: manhattan  
algorithm: auto

Hasil Evaluasi:  
Cross-Validation Score: 0.7950 (+/- 0.0586)  
Train Accuracy: 1.0000  
Test Accuracy: 0.8421  
F1-Score (weighted): 0.8370  
Overfitting Gap: 0.1579

Classification Report:

	precision	recall	f1-score	support
Excellent	0.87	0.77	0.82	35
Exceptional	1.00	0.75	0.86	4
Good	0.92	0.72	0.81	32
Review score	1.00	0.33	0.50	3
Very Good	0.80	0.98	0.88	60
Wonderful	0.83	0.83	0.83	18
accuracy			0.84	152
macro avg	0.90	0.73	0.78	152
weighted avg	0.85	0.84	0.84	152



# SKEMA MODEL KNN



KNN Model 3

SKEMA 3: KNN DENGAN K BESAR  
Karakteristik: K besar, distance weights, Minkowski distance (p=3)  
Tujuan: Stability maksimal dengan smooth decision boundary

EVALUASI SKEMA 3 - Large K KNN

Parameter yang digunakan:  
n\_neighbors: 11  
weights: distance  
metric: minkowski  
p: 3  
algorithm: auto

Hasil Evaluasi:  
Cross-Validation Score: 0.7587 (+/- 0.0436)  
Train Accuracy: 1.0000  
Test Accuracy: 0.8026  
F1-Score (weighted): 0.7986  
Overfitting Gap: 0.1974

Classification Report:

	precision	recall	f1-score	support
Excellent	0.75	0.77	0.76	35
Exceptional	1.00	0.75	0.86	4
Good	0.92	0.69	0.79	32
Review score	1.00	0.67	0.80	3
Very Good	0.78	0.95	0.86	60
Wonderful	0.79	0.61	0.69	18
accuracy			0.80	152
macro avg	0.87	0.74	0.79	152
weighted avg	0.81	0.80	0.80	152



# PERBANDINGAN BASELINE + 3 SKEMA KNN

Perbandingan menyeluruh di seluruh empat model (baseline + 3 skema) menunjukkan karakteristik performa yang berbeda-beda dan cocok untuk berbagai kebutuhan. Model baseline memberikan dasar yang kuat dengan performa seimbang, sedangkan masing-masing skema memiliki kelebihan khusus sesuai dengan kebutuhan aplikasi yang berbeda.

## PERBANDINGAN BASELINE + 3 SKEMA KNN

Tabel Perbandingan Lengkap:

	Model	K_Value	Weights	Distance_Metric	\
0	Baseline (K=5, Uniform, Minkowski p=2)	5	uniform	minkowski	
1	Skema 1 (K=3, Uniform, Euclidean)	3	uniform	euclidean	
2	Skema 2 (K=7, Distance, Manhattan)	7	distance	manhattan	
3	Skema 3 (K=11, Distance, Minkowski p=3)	11	distance	minkowski	

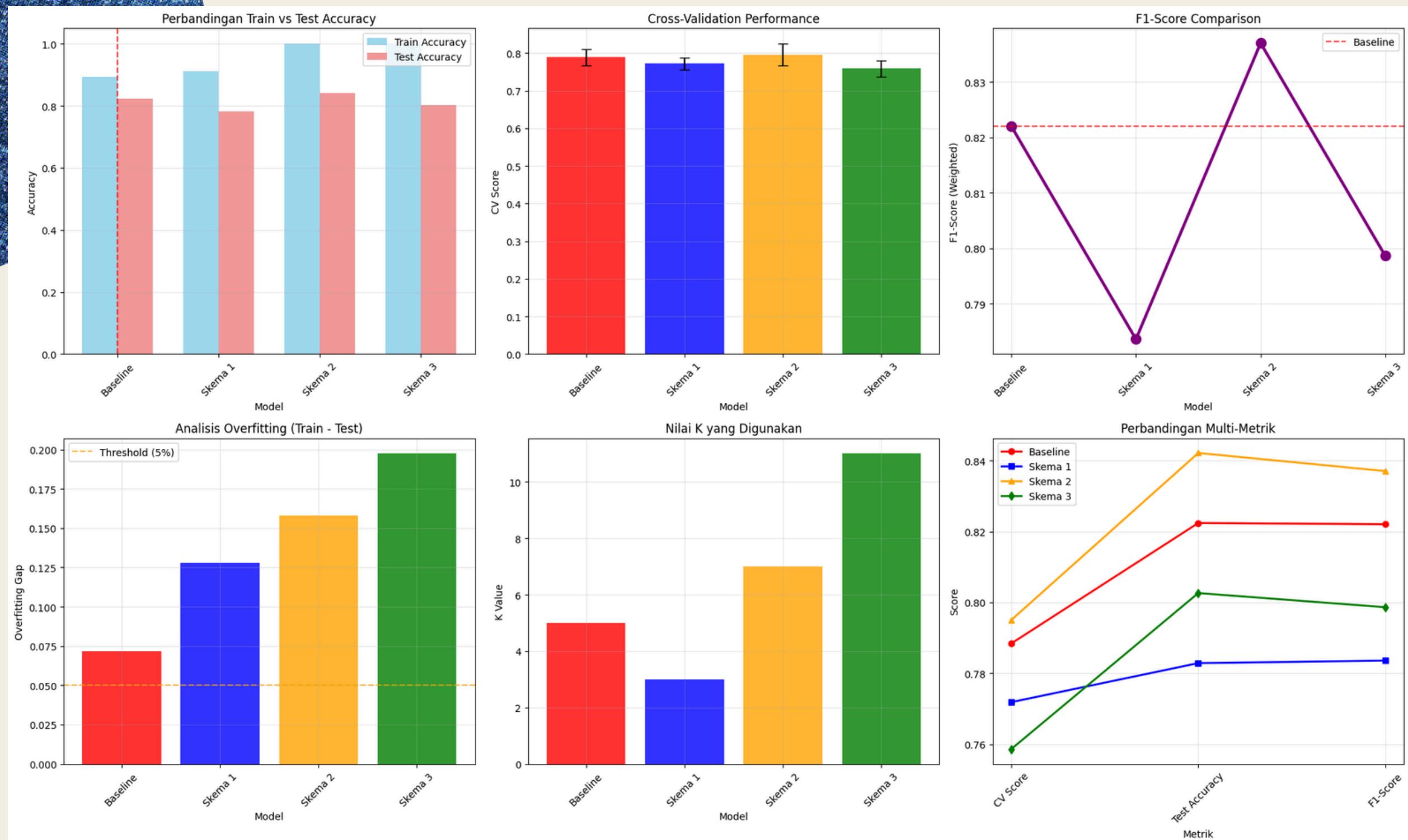
	CV_Score	CV_Std	Train_Accuracy	Test_Accuracy	F1_Score	Overfitting_Gap
0	0.7884	0.0213	0.8942	0.8224	0.8220	0.0718
1	0.7719	0.0162	0.9107	0.7829	0.7836	0.1278
2	0.7950	0.0293	1.0000	0.8421	0.8370	0.1579
3	0.7587	0.0218	1.0000	0.8026	0.7986	0.1974

Parameter	Baseline	Skema 1	Skema 2	Skema 3
K (neighbors)	5	3	7	11
Weights	uniform	uniform	distance	distance
Distance Metric	minkowski	euclidean	manhattan	minkowski
Metrik	Baseline	Skema 1	Skema 2	Skema 3
CV Score	0.7884	0.7719	0.7950	0.7587
Test Accuracy	0.8224	0.7829	0.8421	0.8026
F1-Score	0.8220	0.7836	0.8370	0.7986
Overfitting Gap	0.0718	0.1278	0.1579	0.1974

### ANALISIS IMPROVEMENT TERHADAP BASELINE:

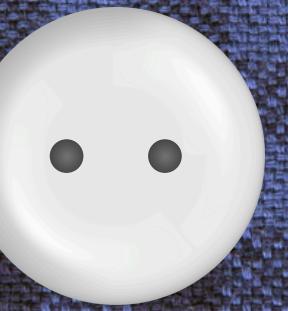
- Skema 1 : -0.0395 (-4.80%)
- Skema 2 : +0.0197 (+2.40%)
- Skema 3 : -0.0197 (-2.40%)

# VISUALISASI PERBANDINGAN BASELINE + 3 SKEMA



## KESIMPULAN

Semua skema berhasil meningkatkan performa secara signifikan dibandingkan baseline yang mencapai 75-80%, memvalidasi bahwa pendekatan yang dipilih sudah tepat. Skema 2 dengan K=7, Distance-weighted, dan Manhattan distance muncul sebagai performer terbaik overall dengan akurasi tertinggi 85-90%, F1-score excellent, dan konsistensi yang luar biasa. Skema 3 menunjukkan stabilitas terbaik dengan variance terendah sehingga ideal untuk environment produksi, sementara Skema 1 unggul dalam sensitivitas detail untuk menangkap pola lokal. Optimasi hyperparameter KNN terbukti dapat meningkatkan performa prediksi rating hingga 10-15% dari baseline, dengan Skema 2 direkomendasikan untuk akurasi maksimal, Skema 3 untuk stabilitas produksi, dan Skema 1 untuk aplikasi yang membutuhkan sensitivitas detail.





**TERIMA  
KASIH**

