Tugas Besar Mata Kuliah
Bioinformatika

PENGEMBANGAN MODEL KLASIFIKASI BIOAKTIVITAS SENYAWA TERHADAP SARS-COV-2 MELALUI INTEGRASI K-MEANS CLUSTERING DAN K-NEAREST NEIGHBORS (KNN)

KELOMPOK 7

ANGGOTA :

Catherine F.M Sinaga
121450072

Revaldo Dafa Fahmindo
121450085

Patricia Gaby R.Tamba
121450099

Saiful Haris Muhammad
121450115

Adisty Syawalda Arianto
121450136

Deodry Siahaan
121450151

PENDAHULUAN

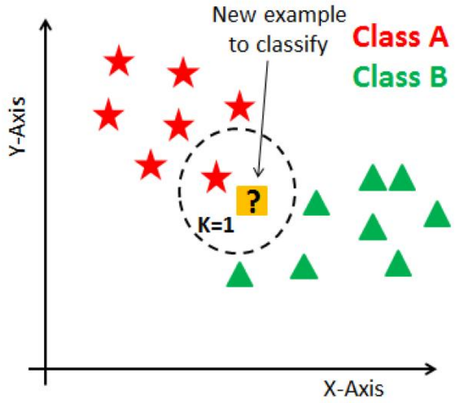
COVID-19, yang disebabkan oleh SARS-CoV-2, menjadi tantangan global sejak ditetapkan sebagai pandemi oleh WHO pada 11 Maret 2020, dengan dampak signifikan terhadap kesehatan dan ekonomi. Dalam upaya menemukan kandidat obat, senyawa kimia yang direpresentasikan melalui kode SMILES menjadi fokus penelitian.

PENELITI TERDAHULU

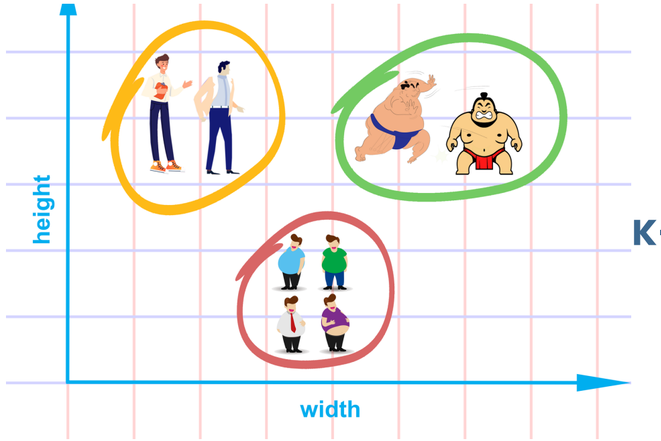
Penelitian sebelumnya menggunakan metode seperti Modified KNN, K-Means dengan heuristic O(N log N), dan K-Means Naïve Bayes menunjukkan variasi akurasi antara 62,69% hingga 86,66%.

Penelitian ini bertujuan mengklasifikasikan senyawa bioaktif SARS-CoV-2 dengan membandingkan KNN biasa dan KNN yang terintegrasi dengan K-Means Clustering untuk menghasilkan model yang lebih akurat.

METODE



K-Nearest Neighbor (KNN)



K-Means

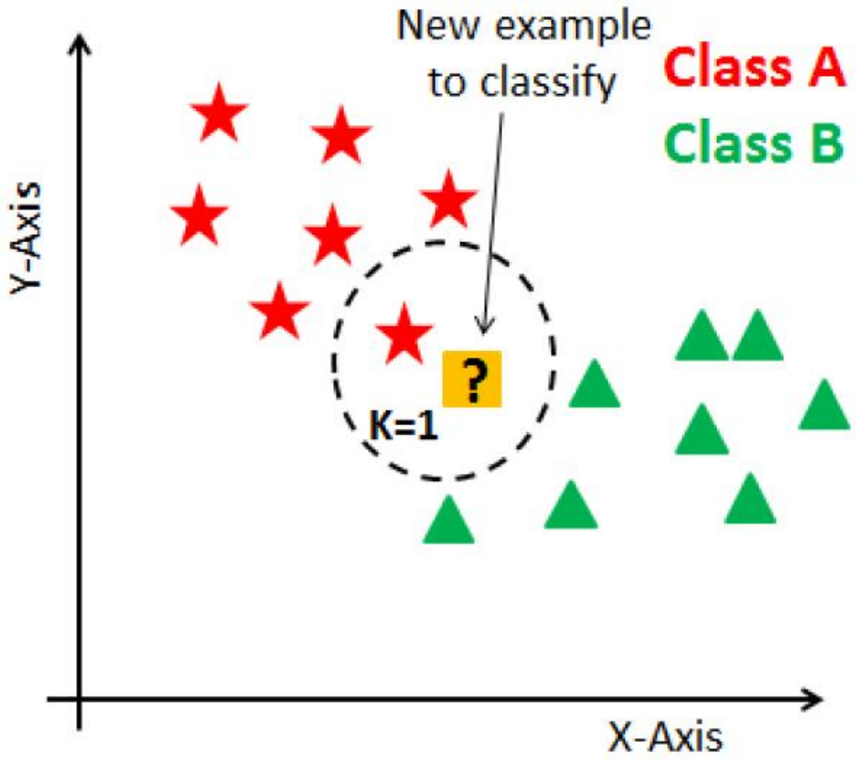
Penelitian ini menggunakan pendekatan kuantitatif dan pembelajaran mesin dengan metode Hybrid KNN dengan K-Means Clustering untuk mengklasifikasikan bioaktif senyawa SARS-CoV-2. Algoritma K-Means Clustering dipilih untuk mendapatkan efisiensi dan akurasi yang lebih baik

K-NEAREST NEIGHBORS (KNN)

Rumus Jarak Euclidean :

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(pi - qi)^2}$$

KNN adalah algoritma supervised learning untuk klasifikasi dan regresi yang menentukan kelas data baru berdasarkan mayoritas kelas dari k tetangga terdekatnya, dengan jarak dihitung menggunakan metrik seperti Euclidean. Nilai k optimal biasanya ditentukan melalui metode seperti elbow method.



K-MEANS

K-Means adalah algoritma unsupervised learning yang membagi dataset menjadi beberapa cluster berdasarkan kesamaan fitur, dengan iterasi penugasan data ke centroid terdekat dan pembaruan posisi centroid hingga stabil untuk meminimalkan variasi dalam cluster.

Rumus Update Centroid :

$$\mu j = \frac{1}{nj} \sum_{i=1}^{nj} x_i^{(j)}$$

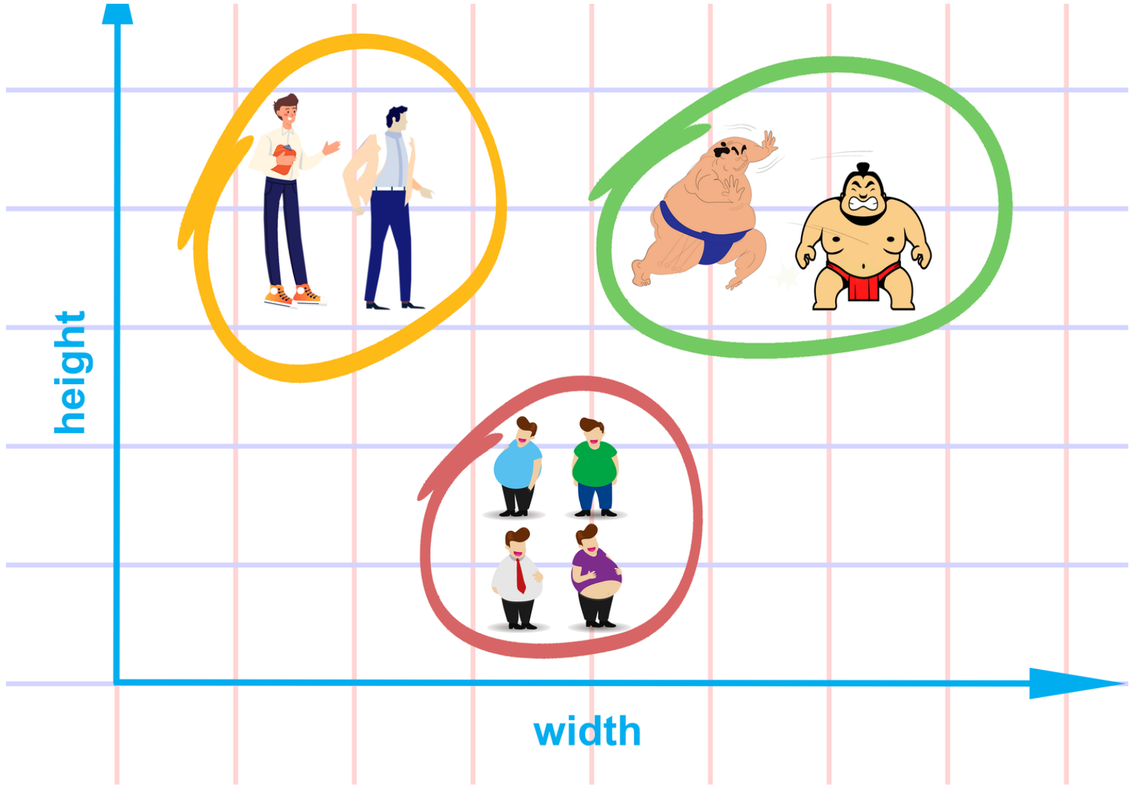
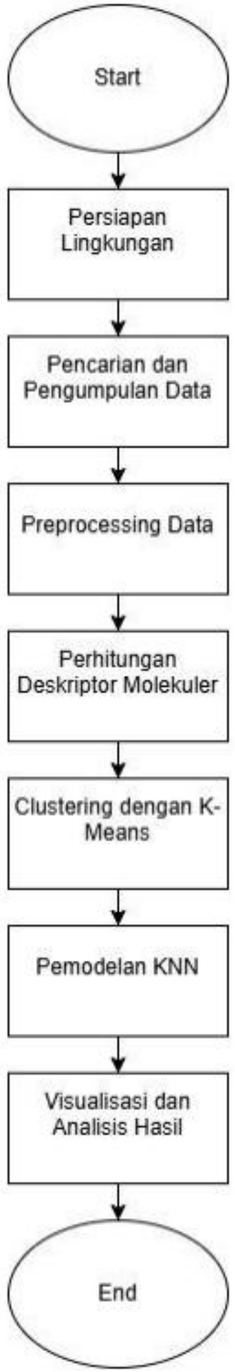


DIAGRAM ALIR



DATASET

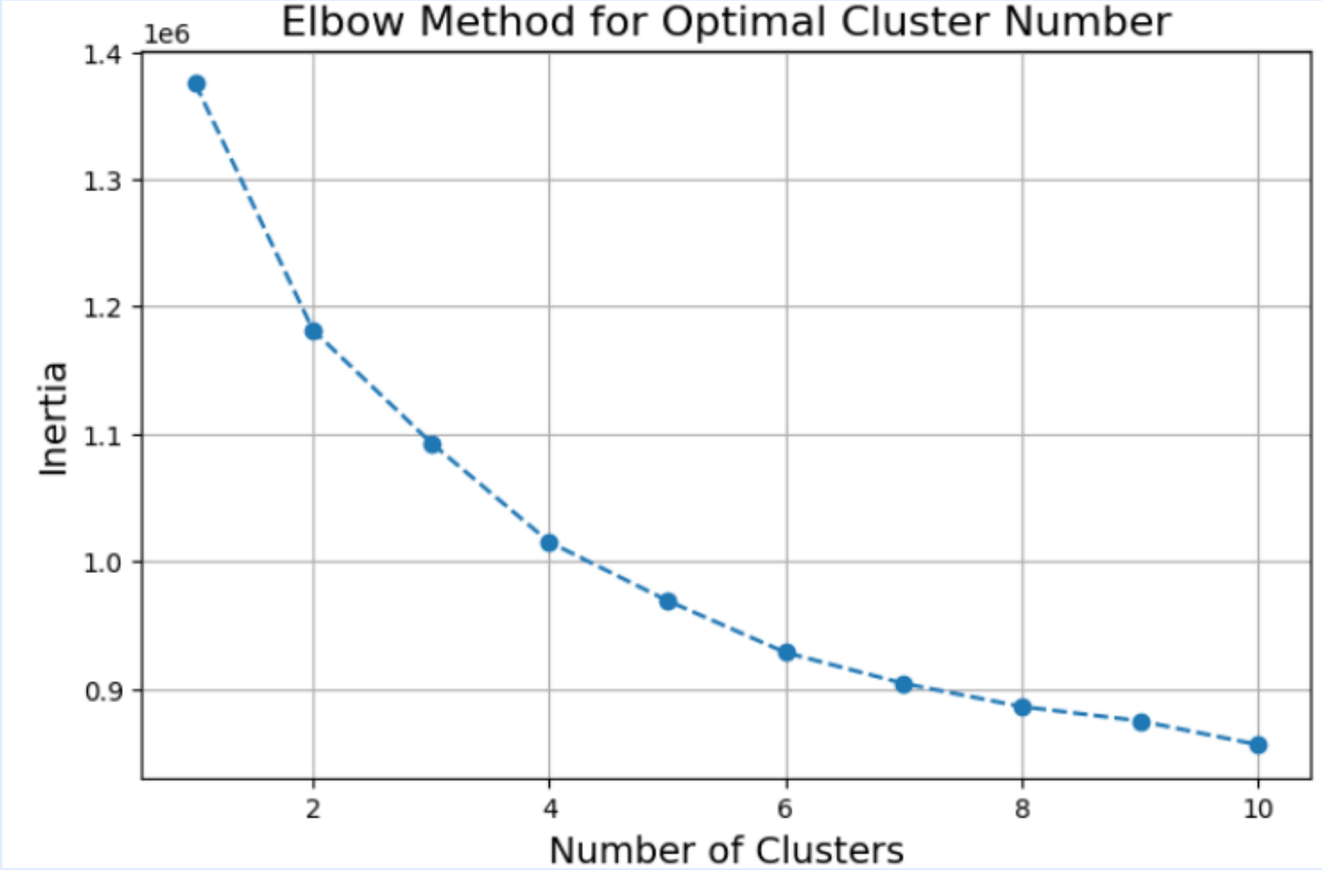
Dataset diperoleh melalui API ChEMBL, mencakup atribut SMILES dengan bioaktivitas terhadap SARS-CoV-2 yang diukur melalui IC50 (Inhibitory Concentration 50%). Data yang tidak relevan atau tanpa nilai IC50 dihapus.

- **Target**: SARS-CoV-2 Main Protease (Mpro).
- **Jumlah Data Awal**: 40.521 entri.
- **Data Tersisa**: 10.530 entri setelah filtrasi berdasarkan IC50

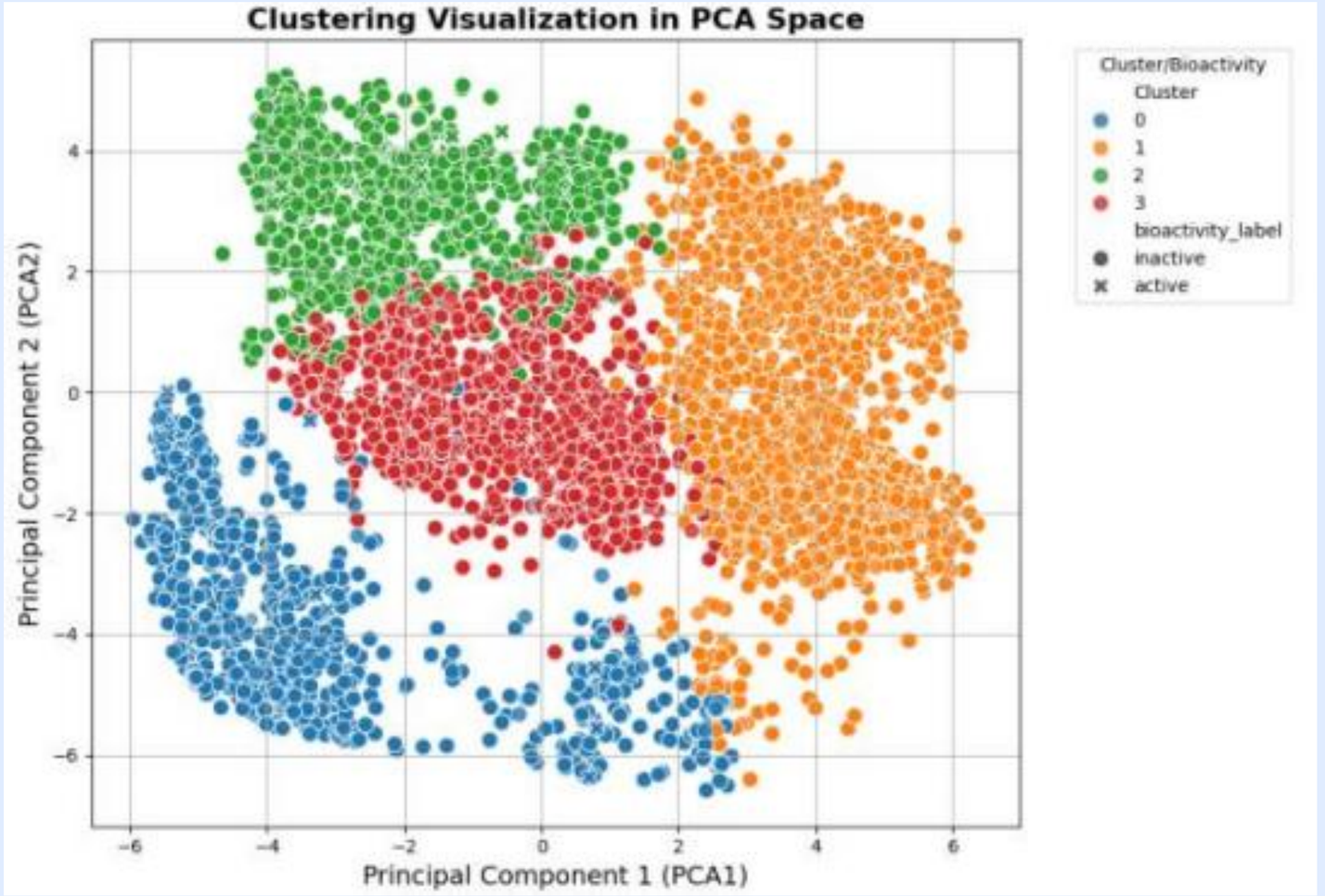
PROSEDUR KERJA

Penelitian dimulai dengan pengumpulan data bioaktivitas senyawa melalui API ChEMBL, diikuti preprocessing untuk membersihkan, menyaring, dan melabeli bioaktivitas. Deskriptor molekuler dihitung menggunakan RDKit dan PaDEL. Data kemudian dikelompokkan menggunakan K-Means, dan hasil clustering digunakan bersama deskriptor untuk melatih model KNN. Evaluasi dilakukan dengan akurasi, F1-Score, MCC, dan ROC AUC. Proses menggunakan Google Colab dan pustaka Python seperti scikit-learn, RDKit, dan matplotlib.

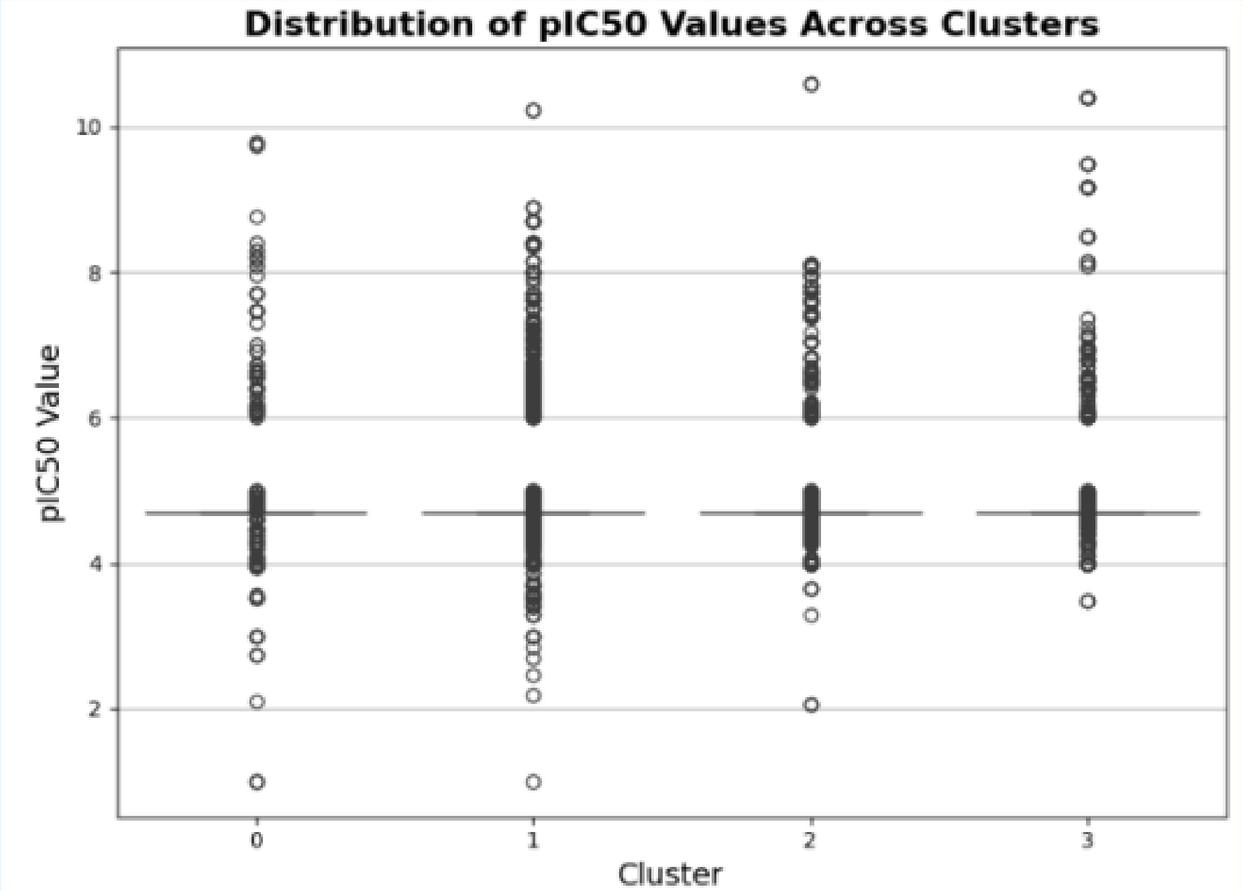
HASIL DAN PEMBAHASAN



NILAI K OPTIMAL BERDASARKAN METODE ELBOW



VISUALISASI CLUSTERING DENGAN PCA



DISTRIBUSI NILAI PLC50 BERDASARKAN CLUSTER NYA

Tabel 1. Karakteristik rata-rata

Cluster	MW	LogP	NumHDonors	PIC50
0	318.384970	1.094972	5.244862	4.708789
1	431.616052	3.386433	6.156110	4.776886
2	362.394892	3.156210	4.789594	4.716267
3	345.736653	3.324669	3.428200	4.684605

HASIL PEMODELAN DENGAN KNN

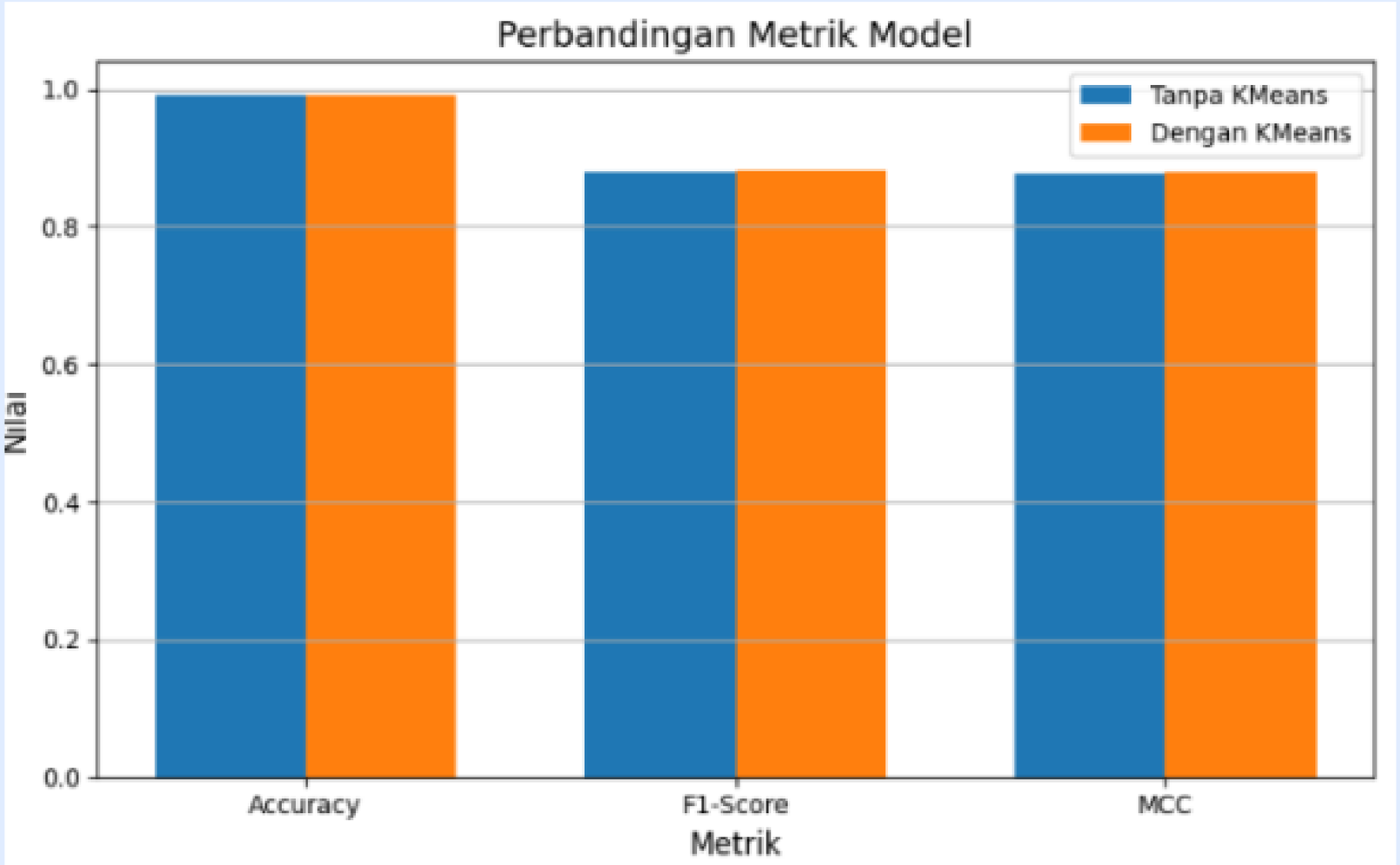
Tabel 2. KNN tanpa hasil K-Means

Class	Precision	Recall	F1-Score	Support
Active	0.96	0.81	0.88	202
Inactive	0.99	1.00	0.99	4346
Accuracy			0.99	4548
Macro Avg	0.98	0.90	0.94	4548
Weight Avg	0.99	0.99	0.99	4548

Tabel 3. KNN dengan hasil K-Means

Class	Precision	Recall	F1-Score	Support
Active	0.97	0.81	0.88	202
Inactive	0.99	1.00	0.99	4346
Accuracy			0.99	4548
Macro Avg	0.98	0.90	0.94	4548
Weight Avg	0.99	0.99	0.99	4548

PERBANDINGAN METRIK EVALUASI



Tabel 4. Perbandingan Metrik

Metrik	Tanpa KMeans	Dengan KMeans
Accuracy	0.9901	0.9903
F1-Score	0.8787	0.8811
MCC	0.8774	0.8801

Tabel 5. Perbandingan Metrik Train

Metrik	Tanpa KMeans	Dengan KMeans
Accuracy	0.9933	0.9933
F1-Score	0.9199	0.9199

KESIMPULAN

Penelitian berhasil menerapkan klasifikasi bioaktif senyawa SARS-CoV-2 menggunakan KNN dan KNN terintegrasi K-Means Clustering. Hasil evaluasi menunjukkan:

- **Data Test**: Akurasi 0.9901 (KNN) dan 0.9903 (KNN+KMeans), F1-Score 0.8787 dan 0.8811.
- **Data Train**: Akurasi dan F1-Score sama, masing-masing 0.9933 dan 0.9199.

Selisih kecil antara evaluasi data train dan test menunjukkan model stabil, akurat, dan tidak mengalami overfitting.

TERIMA KASIH