

Pengembangan Model Klasifikasi Bioaktivitas Senyawa terhadap SARS-CoV-2 melalui Integrasi K-Means Clustering dan K-Nearest Neighbors (KNN)

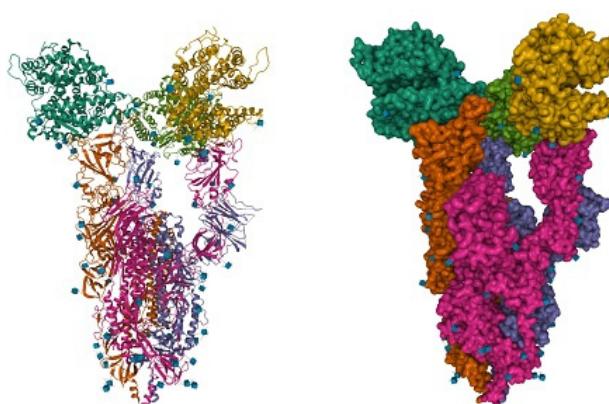
Catherine Firdhasari Maulina Sinaga, Revaldo Dafa Fahmindo, Patricia Gaby Rahmawati Tamba, Saiful Haris Muhammad, Adisty Syawalda Arianto, Deodry Siahaan



Pendahuluan

Pandemi COVID-19 yang disebabkan SARS-CoV-2 mendorong penelitian bioaktif senyawa untuk kandidat obat. Representasi senyawa menggunakan SMILES dievaluasi dengan metode klasifikasi seperti Modified KNN, K-Means heuristic, dan KMNB. Penelitian ini menerapkan Hybrid KNN dengan K-Means Clustering untuk meningkatkan akurasi dibandingkan metode KNN konvensional pada klasifikasi bioaktif.

Dataset



Sumber Data :
API ChEMBLa

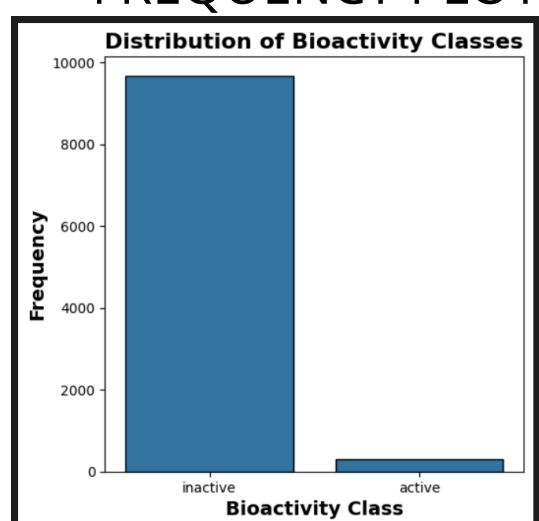
Jumlah Molekul
Yang didapatkan :
9.968 Molekul

Tujuan

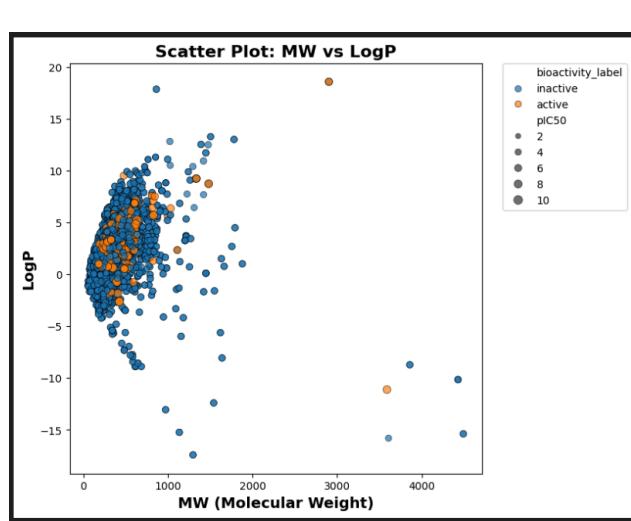
- Membandingkan performa model KNN biasa dengan model KNN yang terintegrasi dengan K-Means Clustering.
- Mengoptimalkan klasifikasi bioaktivitas senyawa SARS-CoV-2 berdasarkan data SMILES.

EXPLORATORY DATA ANALYSIS (EDA)

• FREQUENCY PLOT



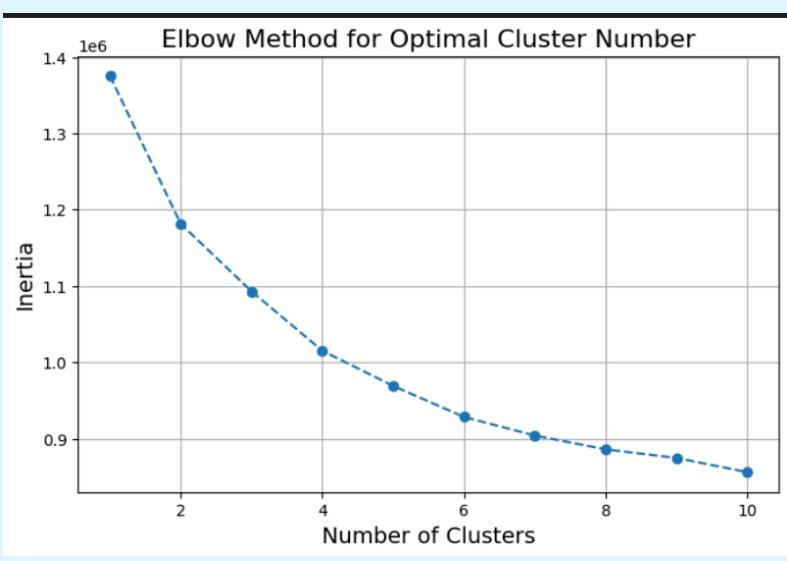
• SCATTER PLOT



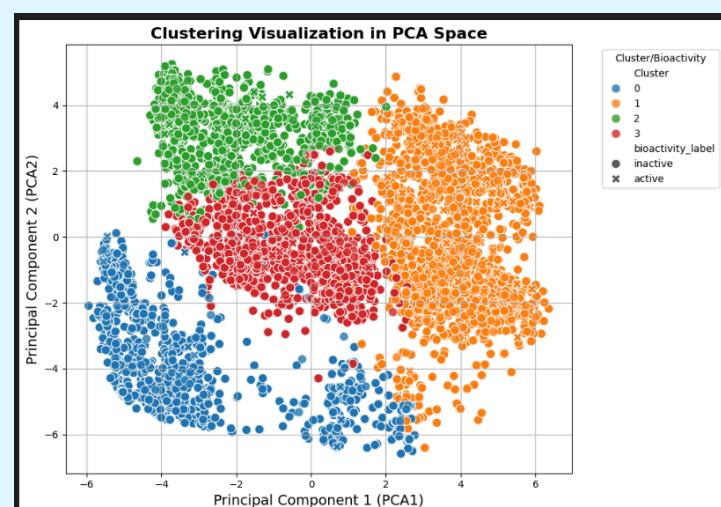
Interpretasi:

- MW: Molekul aktif (<1000 Da) lebih kecil dan efisien, molekul besar (>3000 Da) cenderung tidak aktif.
- LogP: Aktivitas terbaik pada logP 0-5 (seimbang lipofilisitas dan kelarutan).
- pic50: Molekul aktif dengan pic50 tinggi optimal pada MW moderat dan logP ideal.

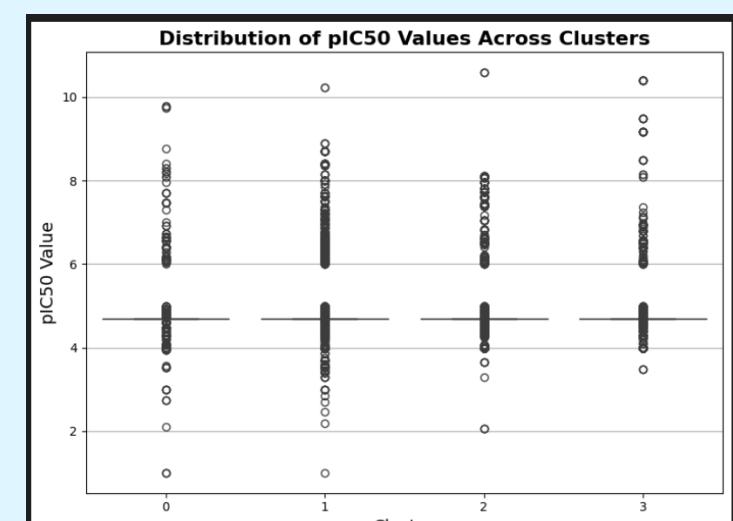
HASIL PEMBAHASAN



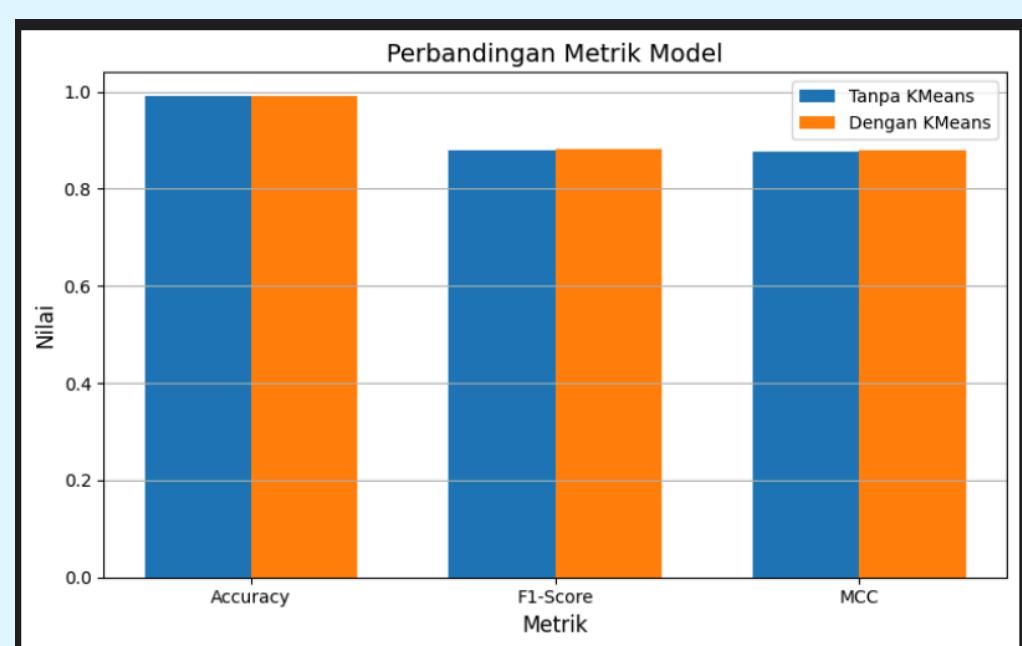
- Nilai k optimal berdasarkan metode elbow



• Visualisasi dengan PCA



- Distribusi nilai pic50 berdasarkan cluster nya



- Perbandingan Metrik Model

Gambar tersebut menunjukkan perbandingan kinerja model klasifikasi bioaktivitas senyawa terhadap SARS-CoV-2 menggunakan tiga metrik: Accuracy, F1-Score, dan MCC, dengan pendekatan Tanpa KMeans dan Dengan KMeans. Hasilnya menunjukkan bahwa integrasi K-Means Clustering dengan K-Nearest Neighbors (KNN) tidak memberikan perubahan signifikan terhadap kinerja model, dengan nilai metrik yang hampir identik pada kedua pendekatan. Hal ini menunjukkan bahwa pengelompokan data dengan KMeans tidak secara substansial memengaruhi hasil klasifikasi bioaktivitas.