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

Development of a Classification Model for Compound Bioactivity Against SARS-CoV-2 Using an Integrated Approach of K-Means Clustering and K-Nearest Neighbors (KNN)

Catherine Firdhasari Maulina Sinaga[1], Revaldo Dafa Fahmindo[2] *, Patricia Gaby Rahmawati Tamba[3], Saiful Haris Muhammad[4], Adisty Syawalda Arianto[5], Deodry Siahaan[6]

[1] Sains Data, Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
**E-mail: revaldo.121450085@student.itera.ac.id*

Abstrak

Upaya pencarian senyawa bioaktif yang berpotensi menjadi kandidat obat untuk SARS-CoV-2 menjadi salah satu fokus utama dalam penelitian di bidang farmasi dan bioteknologi. Penelitian ini mengembangkan model klasifikasi bioaktivitas senyawa terhadap SARS-CoV-2 dengan mengintegrasikan K-Means Clustering pada algoritma K-Nearest Neighbors (KNN). Tujuan utama penelitian ini adalah membandingkan performa model KNN biasa dengan model KNN yang terintegrasi K-Means Clustering. Dataset bioaktivitas senyawa diperoleh dari API ChEMBL dan difilter berdasarkan nilai IC50. Setelah preprocessing dan pemberian label bioaktivitas, deskriptor molekuler dihitung menggunakan RDKit dan PaDEL. Algoritma K-Means digunakan untuk clustering data, yang hasilnya dijadikan fitur tambahan dalam pemodelan KNN. Evaluasi menunjukkan bahwa model KNN dengan integrasi K-Means memiliki akurasi lebih tinggi (99,03%) dibandingkan KNN tanpa integrasi (99,01%). F1-Score juga meningkat dari 0,8787 menjadi 0,8811. Hasil ini menunjukkan bahwa integrasi K-Means Clustering pada KNN dapat meningkatkan performa model dalam klasifikasi bioaktivitas senyawa, serta memberikan landasan untuk pengembangan metode serupa di penyakit lain.

Kata kunci: SARS-CoV-2, bioaktivitas senyawa, klasifikasi, K-Means Clustering, K-Nearest Neighbors (KNN).

Abstract

Efforts to discover bioactive compounds with potential as drug candidates for SARS-CoV-2 have become a major focus in pharmaceutical and biotechnology research. This study developed a bioactivity classification model for compounds against SARS-CoV-2 by integrating K-Means Clustering into the K-Nearest Neighbors (KNN) algorithm. The primary objective of this research is to compare the performance of a standard KNN model with a KNN model integrated with K-Means Clustering. The bioactivity compound dataset was obtained from the ChEMBL API and filtered based on IC50 values. After preprocessing and bioactivity labeling, molecular descriptors were calculated using RDKit and PaDEL. The K-Means algorithm was employed for data clustering, and the results were used as additional features in the KNN modeling. Evaluation results indicate that the KNN model integrated with K-Means achieved higher accuracy (99.03%) compared to the standard KNN model (99.01%). The F1-Score also improved from 0.8787 to 0.8811. These findings demonstrate that integrating K-Means Clustering into KNN can enhance model performance in classifying compound bioactivity and provide a foundation for developing similar methods for other diseases.

Keywords: SARS-CoV-2, compound bioactivity, classification, K-Means Clustering, K-Nearest Neighbors (KNN

PENDAHULUAN

Coronavirus Disease 2019 (COVID-19) merupakan penyakit yang ditemukan pada akhir 2019 di Wuhan, China yang disebabkan oleh virus varian baru bernama *coronavirus 2* (SARS-CoV-2) (1). Secara resmi, WHO menetapkann COVID-19 sebagai status pandemi pada 11 Maret 2020.

Hal ini menjadi tantangan global bagi seluruh lapisan masyarakat dikarenakan dampak dari pandemi yang signifikan terhadap kesehatan dan ekonomi masyarakat. Upaya pencarian senyawa bioaktif yang berpotensi menjadi kandidat obat termasuk salah satu fokus utama dalam penelitian di bidang farmasi dan bioteknologi.

Senyawa kimia adalah zat murni yang terdiri dari dua atau lebih unsur yang dapat membentuk ikatan kimia (2). Senyawa dapat dikategorikan menjadi dua, yaitu senyawa aktif dan senyawa tidak aktif. Data senyawa-senyawa tersebut dalam bentuk kode SMILES (*Simplified Molecular Input Line Entry System*).

Kode SMILES merupakan notasi yang merepresentasikan struktur kimia dalam format baris yang diperkenalkan pada tahun 1980. SMILES termasuk bentuk data yang berisifar *case sensitive* dikarenakan memperhatikan penulisan huruf kapital dan non-kapital (3).

Berdasarkan penelitian terdahulu, penelitian Yunita Dwi Alfiyanti dkk. (2019) menggunakan metode Modified *K-Nearest Neighbor* (k = 3) pada data SMILES, menghasilkan akurasi pengujian 73% dengan 90% data latih (234 data) dan 10% data uji (26 data), serta akurasi rata-rata 62,69% melalui *k-fold cross-validation* (4). Penelitian Sherly Witanto dkk. (2019) *menggunakan K-Means* dengan inisialisasi pusat klaster berbasis *heuristic O(N log N),* menghasilkan akurasi pengujian 63% dengan 512 data latih dan 128 data uji, serta akurasi rata-rata 52,58% dari 10 kali *K-Fold Cross Validation*. Studi ini menunjukkan bahwa *K-Means* dengan *heuristic O(N log N)* lebih akurat dibandingkan *K-Means* konvensional (5). Penelitian Revi Anistia Masykuroh, dkk (2019) menggunakan metode *K-Means Naïve Bayes* (KMNB) dengan rasio data latih 80% dan data uji 20% menghasilkan akurasi rata-rata sebesar 85,45% dengan sistem diuji menggunakan *K-Fold Cross Validation* sebanyak 10 menghasilkan akurasi tertinggi sebesar 86,66% (6).

Berdasarkan penelitian terdahulu, peneliti mencoba menyelesaikan klasifikasi bioaktif senyawa SARS-CoV-2 dengan menerapkan metode Hybrid KNN (*K Nearest Neighbor*) dengan *K-Means Clustering* dimana peneliti akan membandingkan model KNN biasa dengan model KNN yang terintegrasi *K-Means Clustering*.

METODE

Penelitian ini menggunakan pendekatan kualitatid dan pembelajaran mesin dengan metode Hybrid KNN dengan *K-Means Clustering* untuk menklasifikasikan bioaktif senyawa SARS-CoV-2. Algoritma *K-Means Clustering* dipilih untuk mendapatkan efisiensi dan akurasi yang lebih baik (7).

K-Means Clustering

K-Means adalah algoritma pembelajaran mesin tanpa pengawasan (unsupervised learning) yang digunakan untuk membagi dataset menjadi beberapa kelompok (cluster) berdasarkan kesamaan fitur (8). Algoritma ini bekerja dengan menentukan sejumlah pusat cluster (centroid) secara acak, kemudian mengiterasikan proses penugasan data ke centroid terdekat dan memperbarui posisi centroid hingga posisi tersebut stabil. Tujuannya adalah meminimalkan variasi dalam setiap cluster, sehingga data dalam satu cluster memiliki kemiripan yang tinggi.

Rumus Update Centroid:

$$\boldsymbol{\mu j} = \frac{\mathbf{1}}{\boldsymbol{nj}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{nj}} \boldsymbol{x}_{\boldsymbol{i}}^{(j)} \qquad (1)$$

Keterangan :
μ_j : pusat cluster (centroid) ke-*j*
n_j : jumlah data dalam cluster *j*
$x_i^{(j)}$: data ke-i dalam cluster j .

Rumus untuk Within-Cluster Sum of Squares (WCSS):

$$\boldsymbol{WCSS} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{nj}} \boldsymbol{x_i^{(j)}} \qquad (3)$$

Keterangan :

k : jumlah cluster.
n_j: jumlah data dalam cluster j .
$x_i^{(j)}$: data ke-i dalam cluster j .
μ_j : centroid cluster j .
$\| x_i^{(j)} - \mu_j \|^2$: jarak kuadrat antara titik data dan centroid.

***K-Nearest Neighbors* (KNN)**

KNN adalah algoritma pembelajaran mesin dengan pengawasan (supervised learning) yang digunakan untuk klasifikasi dan regresi. Dalam konteks klasifikasi, KNN menentukan kelas suatu data baru berdasarkan mayoritas kelas dari k tetangga terdekatnya dalam ruang fitur. Jarak antara data dihitung menggunakan metrik seperti jarak Euclidean, dan nilai k yang optimal sering ditentukan melalui metode seperti elbow method (9).

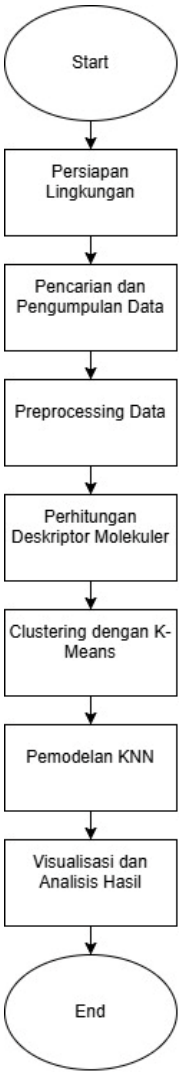
Rumus Jarak Euclidean (untuk menentukan tetangga terdekat):

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

Keterangan:
$d(p,\ q)$: Jarak Euclidean antara titik p dan q
$pi,\ qi$: nilai fitur pada dimensi ke- i
n : jumlah dimensi fitur.

Diagram Alir

Diagram alir penelitian ini dapat dilihat pada **Gambar 1.**



Gambar 1. Alur Penelitian

Proses dimulai dengan Persiapan Lingkungan untuk instalasi pustaka dan alat. Data bioaktivitas senyawa diunduh melalui API ChEMBL dan difilter berdasarkan IC50 pada tahap Pencarian dan Pengumpulan Data. Kemudian, pada Preprocessing Data, data dibersihkan dan diberi label bioaktivitas. Selanjutnya, Perhitungan Deskriptor Molekuler dilakukan menggunakan RDKit dan PaDEL. Setelah itu, Clustering dengan K-Means dilakukan untuk mengelompokkan data, diikuti dengan Pemodelan KNN untuk melatih model dan mengevaluasinya. Terakhir, Visualisasi dan Analisis Hasil dilakukan untuk memvisualisasikan dan menganalisis karakteristik cluster dan hasil klasifikasi.

Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini diperoleh melalui API ChEMBL. Dataset mencakup atribut utama seperti SMILES yang berisi data bioaktivitas senyawa terhadap target protein SARS-CoV-2 yang diukur menggunakan nilai IC50 (*Inhibitory Concentration* 50%). Data yang tidak relevan atau memiliki nilai IC50 yang tidak tersedia dihapus.

a. Target: SARS-CoV-2 Main Protease (Mpro) sebagai target utama.
b. Jumlah Data Awal: 40.521 entri aktivitas senyawa.
c. Filter Data: Dataset difilter berdasarkan parameter IC50 untuk menyaring senyawa dengan aktivitas yang relevan dengan 10.530 entri tersisa.

Proses Preprocessing

Proses preprocessing bertujuan untuk membersihkan data dan menjadikannya siap untuk analisis lebih lanjut.

a. Pembersihan Data:
   - Menghapus nilai kosong pada kolom penting seperti canonical_smiles,standard_value, dan standard_type.
   - Normalisasi nilai IC50 untuk menangani skala besar (maksimal 100 juta nM).
b. Pemberian Label Bioaktivitas:
   Kriteria label:
   - Active: IC50 ≤ 1.000 nM.
   - Inactive: IC50 ≥ 10.000 nM.
   - Intermediate: Di antara nilai tersebut.

   Hanya senyawa dengan label Active dan Inactive yang digunakan.
c. Perhitungan Deskriptor Molekuler:
   Menggunakan pustaka RDKit untuk menghitung deskriptor seperti:
   - Molecular Weight (MW)
   - LogP (hidrofobisitas)
   - Jumlah donor dan akseptor hidrogen.

   Menggunakan alat PaDEL untuk menghasilkan fingerprint molekuler tambahan.

Proses Clustering dan Pemodelan

a. Clustering dengan K-Means:
   - Normalisasi fitur numerik menggunakan MinMaxScaler.
   - Menentukan jumlah cluster optimal dengan metode Elbow.
   - Melakukan clustering untuk mengelompokkan senyawa berdasarkan karakteristik deskriptornya.
   - Dataset akhir untuk clustering memiliki 22.740 entri unik senyawa setelah penghapusan nilai hilang.
b. Pemodelan dengan KNN:
   Melatih model KNN menggunakan dataset :
   - Tanpa fitur clustering: Hanya berdasarkan fitur deskriptor.
   - Dengan fitur clustering: Termasuk hasil clustering sebagai fitur tambahan.

   Mengevaluasi performa model menggunakan:
   - Akurasi
   - F1-Score
   - MCC (Matthews Correlation Coefficient)
   - ROC AUC (untuk analisis probabilitas)

Alat dan bahan

a. Software:
   - Google Colab untuk pengolahan data dan eksekusi kode.
b. Pustaka Python:
   - scikit-learn: Untuk preprocessing, clustering, dan pemodelan.
   - RDKit: Untuk penghitungan deskriptor molekuler.
   - PaDEL-Descriptor: Untuk menghasilkan fingerprint molekuler.
   - Matplotlib dan Seaborn: Untuk visualisasi data dan hasil clustering.
c. Alat Tambahan:
   - API ChEMBL untuk mendapatkan data bioaktivitas senyawa.

- PaDEL untuk menghitung fingerprint molekuler.

Prosedur kerja

Prosedur kerja dalam penelitian ini dimulai dengan pengumpulan data bioaktivitas senyawa melalui API ChEMBL, diikuti dengan preprocessing data untuk membersihkan, menyaring, dan memberikan label bioaktivitas. Data yang telah diproses kemudian digunakan untuk menghitung deskriptor molekuler menggunakan RDKit dan PaDEL.

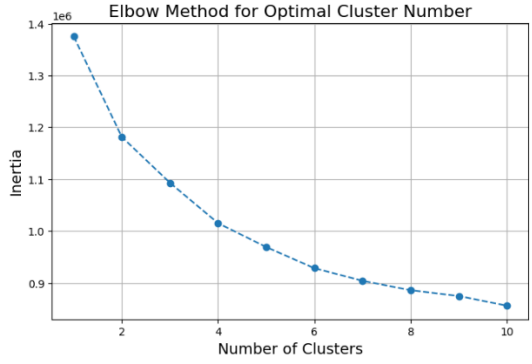
Tahap berikutnya adalah clustering data dengan algoritma K-Means untuk mengidentifikasi pola dalam dataset. Hasil clustering digunakan bersama dengan fitur deskriptor dalam pelatihan model KNN untuk klasifikasi bioaktivitas. Evaluasi dilakukan menggunakan metrik akurasi, F1-Score, MCC, dan ROC AUC. Seluruh proses dikelola menggunakan Google Colab dan pustaka Python seperti scikit-learn, RDKit, dan matplotlib.

HASIL DAN PEMBAHASAN

Pengklasifikasian tersebut menggunakan dataset dengan dimensi (22740, 889) yang merupakan hasil akhir dari proses *Fingerprint Molecular*. Peneliti melakukan percobaan menggunakan KNN (*K-Nearest Neighbor*) dan KNN (*K Nearest Neighbor*) yang diintegrasikan dengan *K-Means Clustering*.

Clustering* dengan *K-Means

Berdasarkan metode elbow pada K-Means didapatkan nilai *k* optimal sebesar 4, grafik dapat dilihat pada Gambar 2.

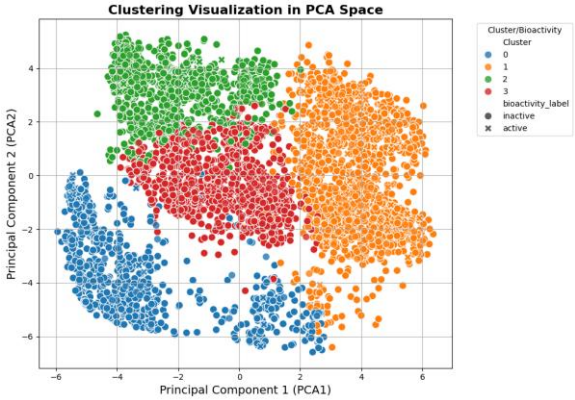


Gambar 2. Nilai *k* optimal berdasarkan metode elbow

Grafik tersebut merupakan visualisasi klasterisasi data dalam ruang dua dimensi yang dihasilkan melalui analisis komponen utama (PCA). Kemudian dibuat scatterplot untuk melihat penyebaran kluster yang dapat dilihat pada Gambar 3.

Pada gambar terdapat empat klaster yang teridentifikasi, diantaranya *cluster* 0 (biru), *cluster* 1 (jingga), *cluster* 2 (hijau), dan *cluster* 3 (merah). Setiap titik data pada grafik mewakili kategori bioaktivitas senyawa, untuk data dengan simbol lingkaran untuk data senyawa yang tidak aktif *(inactive)* dan data dengan simbol "x" untuk data yang aktif *(active)*.

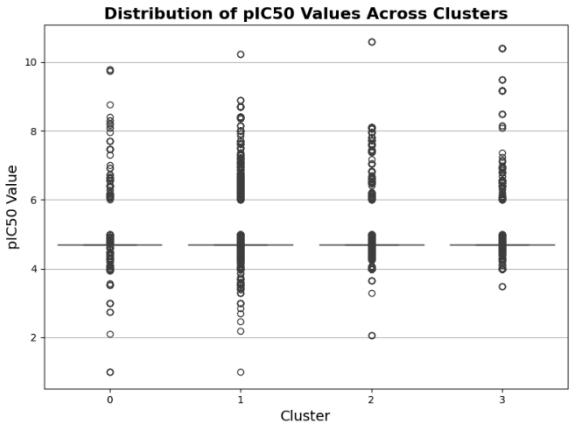
Cluster 0 (biru) berada di area bawah grafik, *cluster* 1 (jingga) mendominasi area kanan grafik, *cluster* 2 (hijau) berada di area atas grafik, dan *cluster* 3 (merah) terkonsentrasi di area tengah grafik dan terlihat tumpang tindih dengan *cluster* lainnya.



Gambar 3. Visualisasi *Clustering* dengan PCA

Grafik tersebut menujukkan distribusi nilai dari plC50 yang berada di antara empat *cluster* (Klaster 0, 1, 2, dan 3). Grafik di bawah menunjukkan distribusi nilai plC50 memiliki pola yang serupa, dengan sebagian besar data terkonsentrasi di sekitar nilai plC50 sebesar 4 hingga 6, namun ada beberapa nilai yang teridentifikasi *outliers* pada beberapa semua klaster yang nilai plC50 nya mendekati 2 dan 10. Hal ini menunjukkan bahwa,

klasterisasi terhadap nilai plC50 tidak menunjukkan perbedaan distribusi yang signifikan.



Gambar 4. Distribusi nilai plC50 berdasarkan *cluster* nya

Adapun karakteristik rata-rata pada tiap cluster. Parameter yang dianalisis meliputi MW (berat molekul), LogP (logaritma koefisien partisi), NumHDonors (jumlah donor hidrogen),dan plC50 (negatif logaritma dari nilai lC50). Setiap baris mewakili satu kluster dan setiap kolom menunjukkan nilai rata-rata untuk parameter dalam kluster tersebut. Karakteristik setiap kluster dapat dilihat pada Tabel 1.

Tabel 1. Karakteristik rata-rata

Cluster	MW	LogP	NumHDonors	PlC50
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

Model KNN dilatih menggunakan dataset dengan dan tanpa fitur clustering. Hasil evaluasi menunjukkan bahwa penambahan fitur clustering meningkatkan performa model.

a. KNN tanpa Hasil Clustering
Setelah dilakukan filtering pada data, dataset pelatihan (X_train) memiliki 18.192 sampel dan 888 fitur, sedangkan dataset pengujian (X_test) memiliki 4.548 sampel dan 888 fitur dan hasil evaluasi dapat dilihat pada Tabel2.

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 tersebut merupakan hasil evaluasi model klasifikasi KNN (K-Nearest Neighbors) yang dijalankan tanpa menggunakan KMeans. Metrik Evaluasi yang diperoleh menghasilkan Accuracy sebesar 0.9901 yang artinya model memiliki akurasi 99%, F1-Score sebesar 0.8787, yang menunjukkan keseimbangan antara precision dan recall, dan Matthews Correlation Coefficient (MCC) sebesar 0.8774, menunjukkan koefisien korelasi yang mengukur kualitas klasifikasi biner. Nilai ini mendekati 1, yang menunjukkan model memiliki kemampuan klasifikasi yang sangat baik.

b. KNN Dengan Clustering
Setelah dilakukan filtering pada data, dataset pelatihan (X_train) memiliki 18.192 sampel dan 889 fitur, sedangkan dataset pengujian (X_test) memiliki 4.548 sampel dan 889 fitur. Adapun hasil evaluasi dapat dilihat pada Tabel 3.

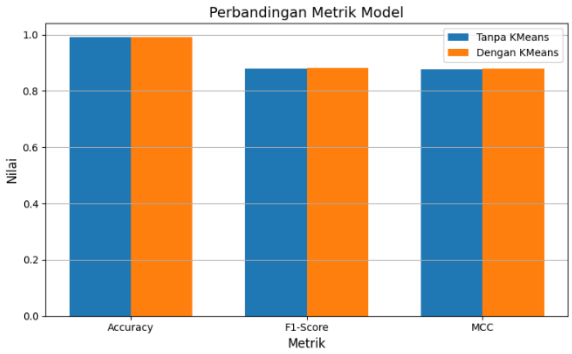
Tabel 3. KNN dengan hasil K-Means

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

Tabel tersebut merupakan hasil evaluasi model klasifikasi KNN (K-Nearest Neighbors) yang dijalankan dengan menggunakan KMeans. Metrik Evaluasi yang diperoleh menghasilkan Accuracy sebesar 0.9903 yang artinya model memiliki akurasi 99%, F1-Score sebesar 0.8811, yang menunjukkan keseimbangan antara precision dan recall, dan Matthews Correlation Coefficient (MCC) sebesar 0.8801, menunjukkan koefisien korelasi yang mengukur kualitas klasifikasi biner. Nilai ini mendekati 1, yang menunjukkan model memiliki kemampuan klasifikasi yang sangat baik.

Perbandingan dengan Publikasi Sebelumnya

Penelitian ini menghasilkan akurasi yang lebih tinggi dibandingkan metode yang hanya menggunakan fitur deskriptor molekuler tanpa clustering. Penambahan clustering sebagai fitur baru terbukti meningkatkan prediksi bioaktivitas senyawa, konsisten dengan temuan yang menunjukkan pentingnya eksplorasi pola di data molekuler yang dapat divisualkan dengan grafik pada Gambar 5.



Gambar 5. Perbandingan Metrik Model

Grafik menunjukkan perbandingan performa model berdasarkan tiga metrik: Accuracy, F1-Score, dan MCC, dengan dua pendekatan, yaitu tanpa KMeans dan dengan KMeans. Penggunaan KMeans memberikan peningkatan kecil pada semua metrik. Accuracy meningkat dari 0.9901 menjadi 0.9903, F1-Score dari 0.8787 menjadi 0.8811, dan MCC dari 0.8774 menjadi 0.8801. Peningkatan ini menunjukkan bahwa KMeans membantu model memahami pola data dengan lebih baik, meskipun dampaknya tidak signifikan.

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 menunjukkan perbandingan performa model berdasarkan tiga metrik, yaitu Accuracy, F1-Score, dan MCC, dengan dua pendekatan: tanpa KMeans dan dengan KMeans.

Penggunaan KMeans memberikan sedikit peningkatan pada semua metrik. Accuracy meningkat dari 0.9901 menjadi 0.9903, menunjukkan bahwa KMeans memberikan kontribusi kecil pada tingkat keakuratan model. F1-Score meningkat dari 0.8787 menjadi 0.8811, yang menandakan peningkatan keseimbangan antara precision dan recall. Selain itu, MCC juga meningkat dari 0.8774 menjadi 0.8801, yang menunjukkan bahwa penggunaan KMeans memperbaiki korelasi prediksi model dengan data asli.

Secara keseluruhan, meskipun dampaknya kecil, KMeans membantu meningkatkan performa model.

Tabel 5. Perbandingan Metrik Train

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

Hasil evaluasi dari data train dari kedua model memberikan nilai yang tinggi namun sejalan jika disandingkan dengan evaluasi dari data test. Pada model tanpa KMeans ataupun dengan KMeans memiliki hasil evaluasi data train yang sama sebesar 0.9933 untuk nilai akurasi dan 0.9199 untuk nilai F1-Score. Pada model tanpa KMeans selisihnya sebesar 0.0032 sedangkan pada model dengan KMeans sebesar 0.003. Jika disandingkan dengan hasil evaluasi data testnya, memberikan penjelasan bahwa model tidak overfitting karena menghasilkan nilai evaluasi pada data test dan data train memiliki selisih yang tidak jauh berbeda.

KESIMPULAN

Penelitian ini berhasil melakukan klasifikasi bioaktif senyawa SARS-CoV-2 dengan menerapkan metode *hybrid* KNN (*K Nearest Neighbor*) dan *K-Means Clustering* dengan membandingkan model KNN biasa dengan model KNN yang terintegrasi *K-Means Clustering.*

Hasil evaluasi data test model KNN baik tanpa K-Means atau dengan K-Means mendapatkan nilai yang tinggi dengan akurasi masing-masing nilai 0.9901 dan 0.9903 dan F1-Score 0.8787 dan 0.8811.

Adapun hasil evaluasi data train kedua model mendapatkan nilai akurasi dan F1-Score yang sama yaitu sebesar 0.9933 dan 0.9199. Dari hasil evaluasi data train dan test yang memiliki selisih yang sedikit menandakan bahwa model baik KNN tanpa KMeans atau dengan KMeans tidak mengalami overfitting. Ini juga menandakan bahwa model stabil dan sudah mampu menangkap pola data secara akurat.

SARAN

Penelitian selanjutnya disarankan untuk membandingkan performa model KNN (dengan dan tanpa integrasi *K-Means Clustering*) dengan algoritma lain, seperti *Random Forest*, *Support Vector Machine* (SVM), atau model berbasis neural network, guna mengevaluasi potensi peningkatan akurasi, *F1-Score*, dan efisiensi model dalam klasifikasi senyawa bioaktif.

Selain itu, pendekatan serupa dapat diterapkan untuk klasifikasi senyawa bioaktif pada penyakit lain, seperti kanker atau penyakit menular lainnya, untuk memperluas aplikasi metode ini dan menguji keandalannya dalam berbagai domain penelitian biomedis.

UCAPAN TERIMA KASIH

Penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan dukungan dan kontribusinya dalam penelitian ini. Terima kasih kepada Tim Peneliti dan Pembimbing **Tirta Setiawan, S.Pd., M.Si** yang telah memberikan arahan, bimbingan, dan wawasan yang sangat berharga sepanjang proses penelitian ini. Terimakasi kepada ChEMBL API atas akses data yang memungkinkan kami untuk mengembangkan model ini dan kepada Pengembang RDKit dan PaDEL, yang menyediakan alat-alat penting dalam perhitungan deskriptor molekuler yang mendukung analisis data. Serta semua pihak yang terlibat langsung maupun tidak langsung, yang telah memberikan dukungan teknis maupun moral yang sangat berarti bagi kelancaran penelitian ini. Semoga hasil dari penelitian ini dapat memberikan kontribusi positif bagi pengembangan ilmu pengetahuan, khususnya dalam upaya penanggulangan SARS-CoV-2 dan penelitian terkait bioaktivitas senyawa..

DAFTAR RUJUKAN

1. Amin S, Aulia L, Salsabila A, Prasetiyo AA. Pencarian Kandidat Obat Baru sebagai Inhibitor Main Protease SARS-

CoV-2 dari Senyawa Aktif Tanaman Andrographis Paniculata: Studi in-Silico [Internet]. Available from: www.penerbitlitnus.co.id

2. Agus Dwinata R, Efendi R, Prima Yudha SS. RANCANG BANGUN APLIKASI TABEL PERIODIK UNSUR DAN PERUMUSAN SENYAWA KIMIA DARI UNSUR KIMIA DASAR BERBASIS ANDROID. Vol. 4, Jurnal Rekursif. 2016.
3. Febry Indarwati D, Ratnawati DE, Anam S. Klasifikasi Fungsi Senyawa Aktif berdasarkan Data Simplified Molecular Input Line Entry System (SMILES) menggunakan Metode Support Vector Machine (SVM) [Internet]. Vol. 3. 2019. Available from: http://j-ptiik.ub.ac.id
4. Dwi Alfiyanti Y, Eka Ratnawati D, Anam S. Klasifikasi Fungsi Senyawa Aktif Data Berdasarkan Kode Simplified Molecular Input Line Entry System (SMILES) menggunakan Metode Modified K-Nearest Neighbor [Internet]. Vol. 3. 2019. Available from: http://j-ptiik.ub.ac.id
5. Witanto S, Ratnawati DE, Anam S. Pengelompokan Fungsi Aktif Senyawa Data SMILES (Simplified Molecular Input Line Entry System) Menggunakan Metode K-Means Dengan Inisialisasi Pusat Klaster Menggunakan Metode Heuristic O(N LogN) [Internet]. Vol. 3. 2019. Available from: http://j-ptiik.ub.ac.id
6. Masykuroh RA, Ratnawati DE, Anam S. Klasifikasi Fungsi Senyawa Aktif Berdasarkan Notasi Simplified Molecular Input Line Entry System (SMILES) Dengan Metode K-Means Naïve Bayes (KMNB) [Internet]. Vol. 3. 2019. Available from: http://j-ptiik.ub.ac.id
7. Habibpour R, Khalilpour K. A new hybrid k-means and K-nearest-neighbor algorithms for text document clustering. Int J Acad Res. 2014 May 30;6(3):79–84.
8. Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies [Internet]. 2013;1(6). Available from: www.ijarcsms.com
9. Sutomo F, Muaafii DA, Naufaldi D, Rasyid A, Kurniawan YI, Afuan L, et al. OPTIMIZATION OF THE K-NEAREST NEIGHBORS ALGORITHM USING THE ELBOW METHOD ON STROKE PREDICTION. Jurnal Teknik Informatika (JUTIF) [Internet]. 2023;4(1). Available from: https://doi.org/10.20884/1.jutif.2023.4.1.839