Penerapan K-Means dan LGBM Classifier untuk Identifikasi Senyawa BACE1 sebagai Inhibitor Alzheimer

The application of K-Means and LGBM Classifier for the identification of BACE1 compounds as Alzheimer inhibitors.

Kharisma Gumilang[1] , Dede Masita[2], Divania Rahmadani[3], Eunike Bunga S[4] , Ramadhita Atifa Hendri[5], Ericson Chandra Sihombing[6]

[1]Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia

**E-mail: dede.121450007@student.itera.ac.id*

Abstrak

Penelitian ini bertujuan untuk mengidentifikasi senyawa yang berpotensi sebagai inhibitor BACE1, enzim kunci dalam pembentukan beta-amyloid yang berperan dalam penyakit Alzheimer. Metodologi penelitian ini melibatkan penerapan algoritma K-Means untuk klasterisasi senyawa berdasarkan fingerprint molekul serta penggunaan Light Gradient Boosting Machine (LGBM) Classifier untuk klasifikasi senyawa aktif dan tidak aktif. Dataset yang digunakan mencakup informasi molekular dalam format SMILES dan nilai pIC50 yang dihitung dari IC50. Proses preprocessing mencakup transformasi SMILES, perhitungan fingerprint molekuler, dan penghapusan data yang tidak relevan. Hasil penelitian menunjukkan bahwa LGBM Classifier memberikan akurasi 93%, F1-Score 93%, dan ROC AUC 95% dalam mengklasifikasikan senyawa aktif dan tidak aktif, sementara klasterisasi menggunakan K-Means mampu mengelompokkan senyawa berdasarkan kesamaan strukturnya. Berdasarkan hasil ini, algoritma yang diterapkan dapat membantu proses virtual screening senyawa potensial sebagai inhibitor BACE1. Dengan demikian, penelitian ini memberikan kontribusi dalam pengembangan metode komputasi untuk mendukung desain obat yang lebih efisien dalam pengobatan penyakit Alzheimer.

Kata kunci: Alzheimer; BACE1; Fingerprint Molekuler; K-Means; LGBM Classifier

Abstract

This study aims to identify compounds with potential as BACE1 inhibitors, a key enzyme in beta-amyloid production associated with Alzheimer's disease. The methodology involves the application of the K-Means algorithm for molecular clustering based on molecular fingerprints and the Light Gradient Boosting Machine (LGBM) Classifier for classifying active and inactive compounds. The dataset includes molecular information in SMILES format and pIC50 values derived from IC50. Preprocessing steps include SMILES transformation, molecular fingerprint calculation, and the removal of irrelevant data. The findings indicate that the LGBM Classifier achieved an accuracy of 93%, F1-Score of 93%, and ROC AUC of 95% in classifying active and inactive compounds, while K-Means clustering effectively grouped compounds based on structural similarities. These results demonstrate the potential of the implemented algorithms to support virtual screening of candidate compounds as BACE1 inhibitors. This study contributes to the development of computational methods for efficient drug design in Alzheimer's treatment.

Keywords: Alzheimer; BACE1; K-Means; LGBM Classifier; Molecular Fingerprints

PENDAHULUAN

Penyakit Alzheimer (AD) merupakan penyakit degeneratif otak yang paling umum dari demensia. Hal ini biasanya terjadi pada penuaan yang memberikan beban pada kesehatan dengan penurunan memori, bahasa, pemecahan masalah dan keterampilan kognitif lainnya yang mempengaruhi kemampuan seseorang dalam melakukan kegiatan sehari-hari [1].

Dalam kondisi Alzheimer, adanya pembentukan plak neuritik yang mengandung peptida Aβ, serta neurofibrillary tangles (NFTs). Penumpukan Aβ di otak yang menjadi salah satu faktor penting dalam patogenesis penyakit Alzheimer. Akumulasi Aβ, terutama peptida Aβ42 di otak memicu terjadinya disfungsi neuron, neurodegenerasi, dan demensia [2]. Protein BACE1 (Beta-Secretase 1) memainkan peran penting dalam patogenesis alzheimer yang menjadikannya target utama untuk pengembangan inhibitor.

Pembentukan peptida beta-amiloid melibatkan dua enzim utama yang berasal dari protein transmembran yang dikenal sebagai Protein Prekursor Amiloid (APP) [3]. Pada proses pemecahan APP, jalur awal yang dilalui melibatkan aktivitas enzimatik pembelahan Beta-sekretase (BACE1), yang juga disebut sebagai enzim pemecah protein prekursor amiloid di β-site 1, dimana memiliki peran krusial dalam pembentukan peptida Aβ [4].

Penelitian ini berfokus pada penerapan pembelajaran mesin, khususnya **K-Means Clustering** dan **Light Gradient Boosting Machine (LGBM) Classifier**, untuk mengidentifikasi senyawa yang memiliki potensi sebagai inhibitor BACE1. Dengan memanfaatkan dataset dari database ChEMBL, penelitian ini bertujuan untuk mengevaluasi karakteristik struktural senyawa menggunakan pendekatan deskriptif dan prediktif untuk mempercepat proses penemuan obat.

METODE

a. Dataset Penelitian

Penelitian ini diawali dengan pengumpulan data dari database ChEMBL yang difokuskan pada target BACE1. Data ini memiliki peran penting dalam memahami mekanisme penyakit Alzheimer, dikarenakan data BACE1 (*Beta-Site Amyloid Precursor Protein Cleaving Enzyme 1*) berperan dalam pembentukan *beta-amyloid* yang merupakan salah satu karakteristik utama dalam perkembangan penyakit Alzheimer.

Proses pengolahan data ini melibatkan konversi data SMILES (*Simplified Molecular Input Line Entry System*) ke dalam format yang lebih mudah dianalisis, serta kalkulasi descriptor Lipinski untuk menilai apakah senyawa tersebut memenuhi kriteria obat yang ideal. Setelah itu, nilai IC50 dari senyawa aktif dikonversi menjadi pIC50, yang merupakan ukuran yang lebih tepat dalam menilai potensi bioaktivitas senyawa.

Tabel 1. Dataset Senyawa dan Aktivitas pIC50 pada Target Biologis

NO	canonical_smiles	pIC50
1	CC(C)C[C@H](NC(=O)[C@@H](NC(=O)[C@@H](N)CCC(=O)O)C(C)C)C(=O)N[C@@H](Cc1ccccc1)[C@@H](O)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](Cc1ccccc1)C(=O)O	6.38404994834359
2	CC(C)C[C@H](NC(=O)[C@H](CC(N)=O)NC(=O)[C@@H](NC(=O)[C@@H](N)CCC(=O)O)C(C)C)[C@@H](O)C[C@@H](C)C(=O)N[C@@H](C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](Cc1ccccc1)C(=O)O	8.69897000433601
….	…….	...
8618	CCCN1CCOCCc2cccc(c2)C[C@@H]([C@H](O)CNC(C)(C)c2cccc(OC)c2)NC(=O)c2cccc(c2)C1=O	4.64397414280687

b. Preprocessing

Dataset akan difilter dan hanya menggunakan data dari spesies Homo sapiens. kemudian, memilih tiga kolom utama yaitu *molecule_chembl_id* (ID unik senyawa dari ChEMBL), *canonical_smiles* yaitu Representasi molekul dalam format SMILES (*Simplified Molecular Input Line Entry System)* dan *standard_value* (nilai IC50 senyawa terhadap target biologis). Data yang mengandung nilai missing values dan duplikat akan dihapus untuk memastikan kualitas dan konsistensi dataset yang digunakan. Data kemudian diberi label berdasarkan nilai IC50 yang tercatat pada kolom standard_value. Tiga kategori bioaktivitas yang ditetapkan adalah:

Tabel 2. Kategori Bioaktivitas Berdasarkan Nilai IC50

Kelas Bioaktivitas	Rentang IC50
Active	IC50 ≤ 1.000 nM
Inactive	IC50 ≥ 10.000 nM
Intermediate	1.000nM<IC50< 10.000 nM

Kolom canonical_smiles yang berisi representasi molekul dalam format SMILES ditransformasi menjadi daftar untuk mendukung perhitungan deskriptor Lipinski. Menggunakan toolkit kimia RDKit, beberapa deskriptor Lipinski dihitung yaitu; Berat Molekul (Molecular Weight, MW), Koefisien Partisi Octanol-Air (logP), Jumlah Donor Hidrogen (NumHDonors), Jumlah Akseptor Hidrogen (NumHAcceptors). Dataset kemudian digabungkan dengan nilai deskriptor Lipinski yang telah dihitung. Nilai IC50 yang tercatat pada kolom standard_value dikonversi menjadi pIC50 dengan rumus:

$$pIC50 = -log_{10}(IC50)$$

Untuk memastikan konsistensi data, nilai IC50 dinormalisasi. Selanjutnya, kelas intermediate dihapus agar fokus analisis hanya pada dua kelas utama, yaitu active dan inactive.

c. K-Means

Analisis klastering dilakukan menggunakan algoritma K-Means. Metode ini digunakan untuk mengelompokkan senyawa berdasarkan kemiripan struktur mereka, yang dapat membantu dalam memahami hubungan antara struktur molekul dan aktivitas biologisnya. Klastering dengan K-Means dilakukan dengan menentukan jumlah klaster (K) yang optimal, yang diukur menggunakan teknik *elbow method.* Setiap klaster mewakili kelompok senyawa dengan kesamaan struktur yang lebih tinggi. Berikut merupakan tahapan-tahapan dalam melakukan clustering menggunakan algoritma K-means :

1. Tentukan jumlah cluster k
2. Tentukan posisi awal pusat cluster secara acak.
3. Kelompokkan setiap data ke cluster yang paling dekat, berdasarkan jarak ke pusat cluster. Jarak antar data dihitung menggunakan rumus jarak Euclidean. biasanya digunakan rumus jarak Euclidean seperti yang dijelaskan di bawah ini:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Dimana :
d = Jarak Euclidean
x_i = Koordinat data ke-i (nilai fingerprint)
y_i = Koordinat Centroid pada dimensi ke-i
n = Jumlah Dimensi (fitur fingerprint molekul)

4. Hitung kembali posisi pusat cluster berdasarkan keanggotaan cluster yang baru. Pusat cluster dihitung sebagai rata-rata dari semua data dalam cluster tersebut. Bisa juga menggunakan median jika diperlukan

$$R_k = \frac{1}{N_k}(\mathrm{X_{1k}} + \mathrm{X_{2k}} + \cdots + \mathrm{X_{nk}}) \quad (2)$$

dimana:
R_k = Rata rata baru
N_k = Jumlah training pattern pada kluster (k)
X_{nk} = pola ke (n) yang menjadi bagian dari kluster (k)

5. Tentukan kembali data ke cluster yang baru. Proses selesai jika pusat cluster tidak berubah, atau kembali ke langkah 3 jika masih ada perubahan[5].

Pada penelitian ini, K-Means menggunakan fingerprint Morgan sebagai representasi fitur untuk setiap molekul. Fingerprint Morgan memiliki keunggulan dalam menangkap pola-pola lokal dalam struktur molekul dan efektif dalam klastering berdasarkan kemiripan struktur.

c. Cross Validation

Cross-validation adalah teknik yang digunakan untuk menilai kinerja model dengan membagi dataset menjadi beberapa subset (folds). Pada penelitian ini menggunakan *ShuffleSplit* dengan membagi data menjadi 10 bagian secara acak (shuffle) ke dalam sejumlah subset (folds) untuk pelatihan dan pengujian. Setiap data point digunakan beberapa kali dalam pelatihan dan pengujian, tetapi dengan pembagian yang berbeda di setiap iterasi. Data menggunakan 10 iterasi dan 30% data sebagai data uji pada setiap iterasi. Hal ini dilakukan untuk mendapatkan evaluasi model yang lebih baik dan stabil.

d. LGBM Classifier

Light Gradient Boosting Machine (LightGBM) adalah metode *gradient boosting* yang cepat, terdistribusi, dan memiliki kinerja tinggi yang berbasis pada *decision tree*. *LightGBM* adalah implementasi dari *Gradient Boosting Decision Tree* (GBDT). Selama proses pelatihan, setiap *decision tree* individual akan melakukan pemisahan data. LightGBM menggunakan dua strategi yaitu *gradient-based one-side sampling* (GOSS) dan pertumbuhan berbasis daun (*leaf-wise growth*) [6].

Pada penelitian ini model dilatih dengan dataset yang telah dibagi menjadi data pelatihan (70%) dan data pengujian (30%). Proses pelatihan dilakukan dengan menggunakan *fingerprint* molekul sebagai fitur (X) dan pIC50 sebagai target (y). Fingerprint yang digunakan untuk klasifikasi adalah *fingerprint Avalon*, yang lebih sederhana dan efisien untuk pengklasifikasian aktivitas biologis molekul. Avalon memiliki performa tinggi untuk menangkap fitur struktural yang relevan dengan pIC50.

e. Evaluasi Model

1. Akurasi (Accuracy)

Akurasi mengukur proporsi prediksi yang benar dari total prediksi yang dilakukan oleh model[7].

2. F1-Score

F1-Score adalah rata-rata dari precision dan recall, untuk memberikan keseimbangan antara precision dan recall, terutama pada dataset yang tidak seimbang.

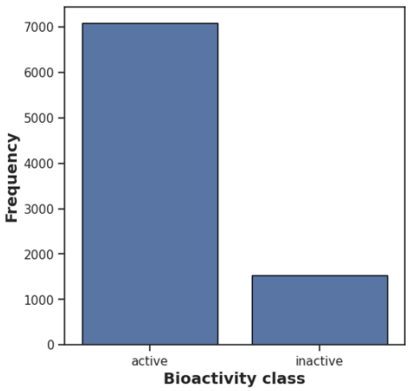
3. Area Under the Curve (AUC) - Receiver Operating Characteristic (ROC)

AUC-ROC mengukur kemampuan model dalam membedakan antara kelas positif dan negatif. Nilai AUC berkisar antara 0 hingga 1, nilai yang tinggi menunjukkan kinerja model yang lebih baik[8].

HASIL DAN PEMBAHASAN

Distribusi Data Senyawa

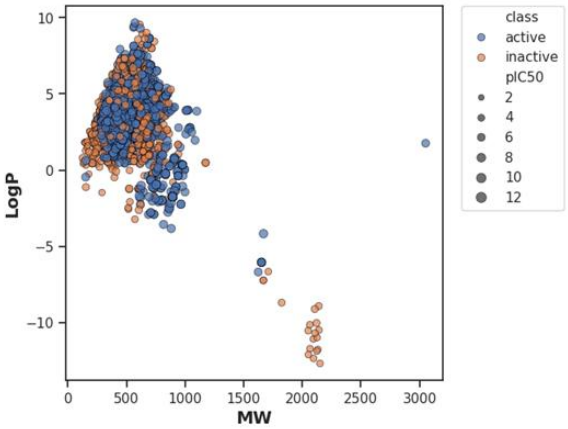
a. Distribusi Kelas Senyawa



Gambar 1. Distribusi Kelas Senyawa

Berdasarkan Gambar 1, dalam dataset sebelum dilakukannya klasifikasi dan *processing data*, menunjukan bahwa jumlah senyawa aktif sebanyak lebih dari 7.000 senyawa, dibandingkan dengan senyawa tidak aktif berjumlah sekitar 1.500 senyawa.

b. Distribusi Massa Molekul dan Potensi Bioaktivitas (plC50)



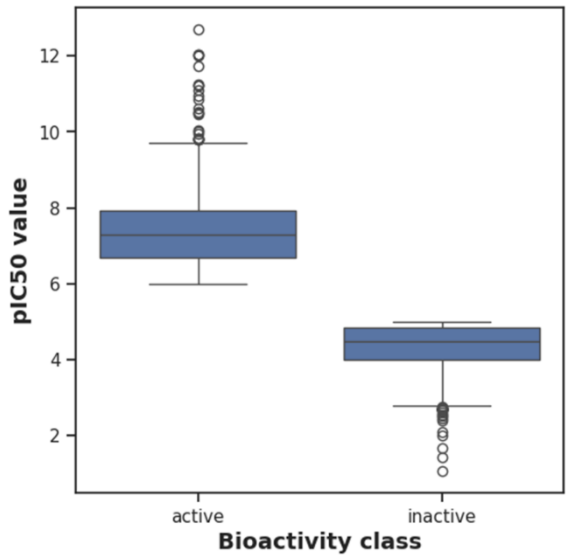
Gambar 2. Scatter Plot MW vs LogP

Berdasarkan Gambar 2 diatas, menunjukan adanya hubungan antara massamolekul (MW) dan koefisien distribusi dari senyawa yang dianalisis, dimana variabel MW mempresentasikan berat molekul dari masing masing senyawa, Sedangkan LogP pada sumbu y mengukur tingkat lipofilisitas atau kecenderungan senyawa untuk larut dalam lemak dibandingkan dengan air. Pada kategori kelas terdapat kelas senyawa aktif dan tidak aktif serta nilai plC50 sebagai indikator sebagai potensi biologis senyawa.

Distribusi data menunjukan bahwa

Sebagian besar senyawa aktif dan tidak aktif terpusat pada kisaran MW antara 300-800 dan LogP antara 0-7, dimana pada pola ini mengindikasikan bahwa senyawa dengan karakteristik tersebut memiliki potensi biologis yang lebih relevan. Senyawa dengan MW > 1000 dan LogP < -5 mengindikasikan bahwa senyawa dengan karakteristik ini memiliki kemungkinan lebih rendah dalam aktivitas biologis yang signifikan. Dari hasil plot diatas, pola persebaran titik nilai plC50 cenderung terkonsentrasi pada kelas aktif, yang mana hal ini menunjukan bahwa senyawa aktif memiliki potensi biologis yang lebih kuat dan lebih sering diklasifikasikan sebagai senyawa aktif.

c. Perbandingan Potensi Bioaktivitas Senyawa

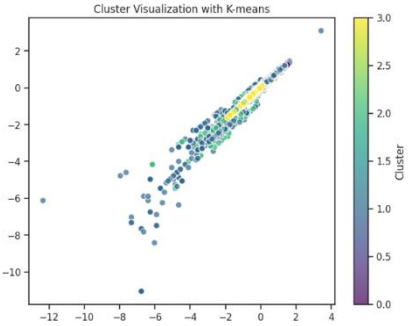


Gambar 3. Kelas Bioaktivitas

Nilai plC50 merupakan logaritma negatif dari konsentrasi inhibitor setengah maksimal (IC50), yang mengindikasikan kekuatan suatu senyawa dalam menghambat aktivitas biologis target tertentu. Semakin tinggi nilai plC50, semakin kuat afinitas senyawa tersebut terhadap target.

Berdasarkan Gambar 3 diatas, dapat dilihat bahwa median nilai plC50 pada kelompok aktif secara signifikan lebih tinggi dibandingkan dengan kelompok tidak aktif. Hal ini menunjukan bahwa secara umum, senyawa aktif memiliki afinitas yang lebih kuat terhadap target biologis dibandingkan dengan senyawa tidak aktif.

Hasil Klastering Senyawa



Gambar 4. Plot Klastering

Pada model clustering yang ditunjukan pada Gambar 4 menggunakan algoritma K-Means untuk mengelompokan senyawa berdasarkan karakteristik strukturalnya, dimana proses ini menghasilkan 4 kluster yang berbeda, dengan masing-masing senyawa menunjukan pola karakter dan struktur kimia yang unik. Hasil dari klustering dapat dilihat pada Tabel 1 berikut:

Tabel 3. Klastering Senyawa

Cluster	Jumlah Senyawa	Karakteristik Struktur
0	3.558	Mengandung heteroatom seperti nitrogen (N) dan klorin (Cl).
1	2.494	Struktur kompleks dengan gugus karboksil (COOH) dan amida (NC=O).
2	1.323	Memiliki gugus hidroksil (-OH) dan fenil (C6H5).
3	1.242	Struktur fenol dan alkohol, yang berpotensi menghambat aktivitas biologis.

Berdasarkan Tabel 1 diatas, dapat dilihat bahwa cluster 0 terdiri dari 3558 senyawa yang secara dominan mengandung heteroatom seperti nitrogen (N) dan klorin (Cl). Kehadiran heteroatom ini sering diasosiasikan dengan potensi bioaktivitas yang tinggi, terutama dalam aplikasi farmasi dan kimia medis. Senyawa dalam cluster ini berpotensi memiliki sifat antibakteri, antijamur, dan aktivitas biologis lainnya yang relevan.

Cluster 1 mencakup 2494 senyawa dengan struktur kompleks yang memiliki gugus karboksil (COOH) dan amida (NC=O). Gugus fungsi ini sering ditemukan pada senyawa dengan sifat asam lemah dan kelarutan yang baik dalam air, yang membuatnya cocok untuk formulasi farmasi dan agen terapeutik. Keberadaan amida juga memberikan kontribusi pada stabilitas struktural dan potensi reaktivitas biologis.

Cluster 2 berisi 1323 senyawa yang menunjukkan keberagaman struktur dengan gugus hidroksil (-OH) dan fenil (C6H5). Gugus hidroksi sering terlibat dalam pembentukan ikatan hidrogen, meningkatkan interaksi dengan target biologis. Senyawa dengan gugus fenil, di sisi lain, memiliki sifat hidrofobik yang dapat memfasilitasi penetrasi membran sel, menjadikannya kandidat potensial untuk pengembangan obat.

Cluster 3 terdiri dari 1242 senyawa yang mengandung struktur fenol dan alkohol. Struktur ini sering dihubungkan dengan aktivitas antioksidan dan penghambatan aktivitas enzim tertentu, yang menjadikannya relevan untuk pengembangan agen antiinflamasi dan antiproliferatif.

Hasil Klasifikasi Senyawa

Proses klasifikasi menggunakan algoritma LightGBM (LGBM) untuk mengelompokkan senyawa aktif dan tidak aktif. Hasil klasifikasi pada Tabel 2 menunjukan bahwa dari seluruh dataset yang ada, jumlah senyawa aktif berkurang menjadi 2.219 senyawa, sedangkan senyawa tidak aktif meningkat menjadi 6.398 senyawa. Hal tersebut mengindikasikan bahwa model yang dimiliki mempunyai tingkat sensitivitas yang lebih tinggi dalam mengklasifikasikan senyawa sebagai tidak aktif dibandingkan sebelumnya.

Tabel 4. Klasifikasi Senyawa

Kategori	Jumlah Senyawa
Senyawa Aktif	2.219
Senyawa Tidak Aktif	6.398

Dalam membedakan senyawa berdasarkan karakteristik struktural yang lebih mendalam, model mengklasifikasikan senyawa dengan nilai plC50 > 5 menunjukan senyawa aktif sedangkan jika nilai plC50 < 5 menunjukan senyawa tidak aktif. Model LGBM *Classifier* ini memiliki kemampuan yang baik dalam membedakan senyawa, namun ditemukan beberapa kasus dimana senyawa dengan nilai plC50 > 5 yang seharusnya diidentifikasi sebagai senyawa aktif, namun pada model diidentifikasi sebagai senyawa tidak aktif. Hal ini mengindikasikan bahwa model LGBM ini terlalu sensitif terhadap pola tertentu dalam struktur kimia sehingga mengabaikan senyawa aktif potensial yang memiliki struktur berbeda dari pola yang dikenalinya.

Metric Evaluasi

Pada model ini dilakukan evaluasi model menggunakan metrik-metrik untuk mengukur kinerja model dalam mengklasifikasikan senyawa aktif dan tidak aktif, yang ditunjukan pada tabel 3 dibawah ini.

Tabel 5. Evaluasi Model

Metric	Value
Train Accuracy	97%
Test Accuracy	93%
Accuracy(C-V)	93%
F1-Score (C-V)	93%
ROC-AUC(C-V)	95%

Berdasarkan Tabel 3 evaluasi diatas, dapat dilihat bahwa model LGBM sangat

efektif dalam mengklasifikasikan senyawa sebagai aktif atau tidak aktif. Akurasi yang tinggi pada data pelatihan (97%) dan pengujian (93%) menunjukkan bahwa model dapat memprediksi aktivitas biologis senyawa dengan baik, bahkan pada data yang belum dilihat sebelumnya. F1-Score dan ROC AUC yang tinggi lebih mengkonfirmasi bahwa model tidak hanya akurat, tetapi juga seimbang dalam menangani kedua kelas.

KESIMPULAN

Penelitian ini berhasil memanfaatkan data dari ChEMBL untuk mengidentifikasi senyawa potensial sebagai inhibitor BACE1, yang merupakan target utama dalam pengembangan terapi penyakit Alzheimer. Proses transformasi data dan perhitungan descriptor, seperti berat molekul (MW), logP, jumlah donor hidrogen, dan akseptor hidrogen, telah memberikan pemahaman yang mendalam terkait karakteristik senyawa bioaktif. Konversi nilai IC50 ke pIC50 juga memungkinkan evaluasi potensi senyawa secara lebih relevan.

Analisis klastering menggunakan algoritma K-Means berdasarkan fingerprint molekul berhasil mengelompokkan senyawa berdasarkan kemiripan struktur, sehingga memberikan wawasan tambahan terkait hubungan antara struktur dan aktivitas biologis. Selain itu, model klasifikasi menggunakan LGBM Classifier dengan fingerprint Avalon menunjukkan performa yang baik dalam memprediksi aktivitas biologis senyawa, dengan akurasi mencapai 93%, F1-Score sebesar 93%, dan AUC-ROC sebesar 95%.

SARAN

Berdasarkan hasil penelitian ini, terdapat beberapa saran yang dapat diberikan untuk pengembangan penelitian di masa mendatang. Pertama, disarankan untuk memperluas cakupan dataset dengan mengumpulkan data yang lebih beragam dan mencakup berbagai jenis senyawa bioaktif. Hal ini bertujuan untuk meningkatkan generalisasi model sehingga mampu menghasilkan prediksi yang lebih akurat pada senyawa baru. Kedua, penggunaan berbagai jenis fingerprint molekul yang lebih bervariasi, seperti Atom Pair, Topological Torsion, atau Morgan Circular, dapat dieksplorasi lebih lanjut untuk menggali hubungan struktur dan aktivitas yang mungkin tidak teridentifikasi menggunakan fingerprint Avalon saja. Ketiga, pengembangan model dapat ditingkatkan dengan membandingkan performa berbagai algoritma pembelajaran mesin, seperti Random Forest, XGBoost, atau Neural Networks, serta mengoptimalkan hyperparameter untuk meningkatkan akurasi dan reliabilitas model.

UCAPAN TERIMA KASIH

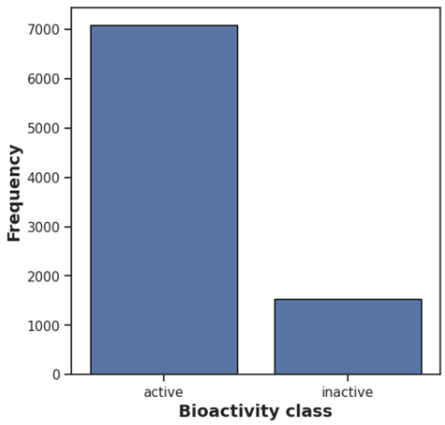
Ucapan terima kasih kepada Bapak Tirta Setiawan, S.Pd., M.Si. selaku dosen pengampu mata kuliah Bioinformatika Sains Data ITERA, yang telah memberikan ilmunya sampai dengan selesainya artikel ini. Penelitian ini diharapkan dapat membantu peningkatan pemahaman dan pengembangan lebih lanjut dalam bidang bioinformatika, khususnya dalam aplikasi pembelajaran mesin untuk identifikasi senyawa sebagai inhibitor Alzheimer.

DAFTAR RUJUKAN

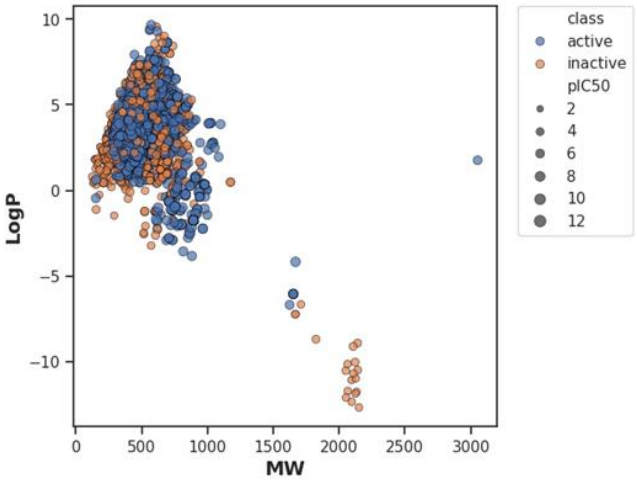
1. A. G. M. Sianturi, “Stadium, Diagnosis, dan Tatalaksana Penyakit Alzheimer,” Maj. Kesehat. Indones., vol. 2, no. 2, pp. 39–44, 2021, doi: 10.47679/makein.202132.
2. K. A. M. Pattni, “Beta-Amyloid As Pathogenesis of Alzheimer Disease,” *e-Jurnal Med. Udayana*, vol. 2, no. 8, pp. 1306–1317, 2013.
3. C. A. B. Rustiawati and J. Pondang, “Penggunaan Lecanemab Pada Alzeimer: Sebuah Harapan Baru,” J. Sehat Indones., vol. 6, no. 02, pp. 479–494, 2024, doi: 10.59141/jsi.v6i02.96.
4. T. R. Noviandy, K. Nisa, G. M. Idroes, I. Hardi, and N. R. Sasmita, “Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer’s Drug Discovery with LightGBM,” *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 358–367, 2024, doi: 10.62411/jcta.10129.
5. M. R. Nugroho, I. E. Hendrawan, and Puwantoro, "Penerapan Algoritma K-Means

Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI," Nuansa Informatika, vol. 16, no. 1, pp. 125, Jan. 2022.
6. F. I. Kurniadi dan P. D. Larasati, "Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke," Jurnal Sistem Komputer dan Kecerdasan Buatan, vol. VI, no. 1, hlm. 67, Sep. 2022.
7. A. M. Carrington et al., "Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation," arXiv preprint arXiv:2103.11357, 2021.
8. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.

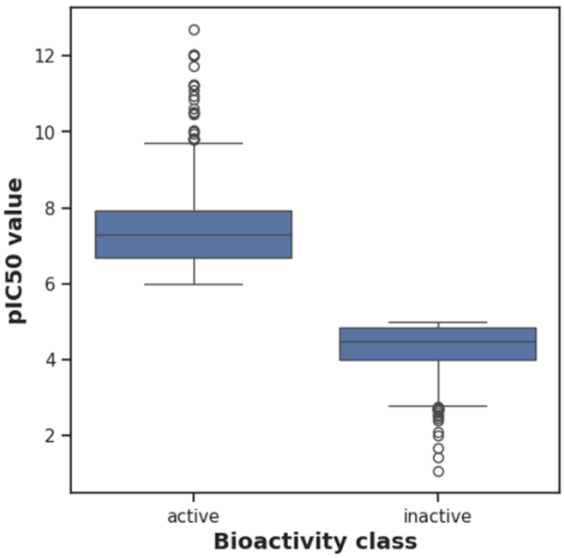
LAMPIRAN GAMBAR



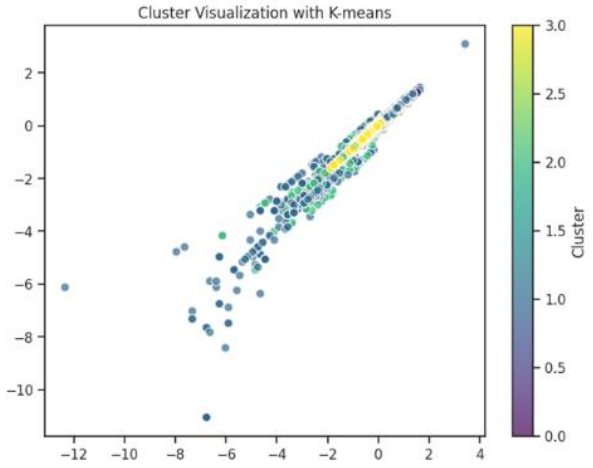
Gambar 1. Distribusi Kelas Senyawa



Gambar 2. Scatter Plot MW vs LogP



Gambar 3 Kelas Bioaktivitas



Gambar 4 Plot Klastering

LAMPIRAN TABEL

Tabel 1. Dataset Senyawa dan Aktivitas pIC50 pada Target Biologis

NO	canonical_smiles	pIC50
1	CC(C)C[C@H](NC(=O)[C@@H](NC(=O)[C@@H](N)CCC(=O)O)C(C)C)C(=O)N[C@@H](Cc1ccccc1)[C@@H](O)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](Cc1ccccc1)C(=O)O	6.3840499483 4359
2	CC(C)C[C@H](NC(=O)[C@H](CC(N)=O)NC(=O)[C@@H](NC(=O)[C@@H](N)CCC(=O)O)C(C)C)[C@@H](O)C[C@@H](C)C(=O)N[C@@H](C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](Cc1ccccc1)C(=O)O	8.6989700043 3601
….	…….	...
8618	CCCN1CCOCCc2cccc(c2)C[C@@H]([C@H](O)CNC(C)(C)c2cccc(OC)c2)NC(=O)c2cccc(c2)C1=O	4.6439741428 0687

Tabel 2. Kategori Bioaktivitas Berdasarkan Nilai IC50

Kelas Bioaktivitas	Rentang IC50
Active	IC50 ≤ 1.000 nM
Inactive	IC50 ≥ 10.000 nM
Intermediate	1.000nM<IC50< 10.000 nM

Tabel 3. Klastering Senyawa

Cluster	Jumlah Senyawa	Karakteristik Struktur
0	3.558	Mengandung heteroatom seperti nitrogen (N) dan klorin (Cl).
1	2.494	Struktur kompleks dengan gugus karboksil (COOH) dan amida (NC=O).
2	1.323	Memiliki gugus hidroksil (-OH) dan

		fenil (C6H5).
3	1.242	Struktur fenol dan alkohol, yang berpotensi menghambat aktivitas biologis.

Tabel 4. Klasifikasi Senyawa

Kategori	Jumlah Senyawa
Senyawa Aktif	2.219
Senyawa Tidak Aktif	6.398

Tabel 5. Evaluasi Model

Metric	Value
Train Accuracy	97%
Test Accuracy	93%
Accuracy(C-V)	93%
F1-Score (C-V)	93%
ROC-AUC(C-V)	95%