

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

Kharisma Gumlang, Ericson Chandra S, Dede Masita, Ramadhita Atifa H,  
Divania Rahmadani, Eunike Bunga S

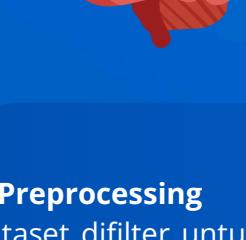
Sains Data, Sains, Institut Teknologi Sumatera, Lampung, Indonesia

## Tujuan

1. Mengelompokkan Senyawa dengan K-Means Clustering
2. Mengklasifikasikan Senyawa Menggunakan LGBM Classifier
3. Mendukung Virtual Screening dengan Model

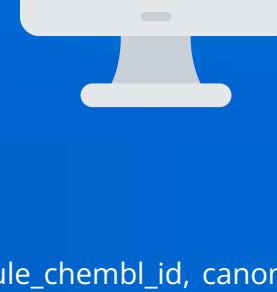
## Pendahuluan

Penyakit Alzheimer (AD) adalah penyakit degeneratif otak yang menyebabkan demensia, ditandai dengan penurunan kognitif akibat akumulasi plak neuritik ( $A\beta$ ) dan neurofibrillary tangles (NFTs), khususnya  $A\beta42$ , yang memicu neurodegenerasi. Protein BACE1, enzim kunci dalam pemecahan Protein Prekursor Amiloid (APP) untuk membentuk  $A\beta$ , menjadi target utama terapi.



## Dataset

- Sumber Data : ChEMBL
- Data Target : BACE1
- Jumlah Molekul : 8618



## Metode

### 1. Preprocessing

Dataset difilter untuk spesies Homo sapiens dengan kolom utama: molecule\_chembl\_id, canonical\_smiles, dan standard\_value. Data duplikat dan missing values dihapus, IC50 dikonversi ke pIC50, dan kelas Intermediate dihapus. Deskriptor Lipinski dihitung menggunakan RDKit.

### 2. K-Means Clustering

Klastering menggunakan fingerprint Morgan dengan elbow method untuk menentukan jumlah klaster (K). Data dikelompokkan berdasarkan jarak Euclidean hingga klaster stabil.

### 3. Cross Validation

ShuffleSplit (10 iterasi, 30% data uji) digunakan untuk membagi data secara acak dan mengevaluasi model.

### 4. LGBM Classifier

Model LightGBM dilatih (70%) dan diuji (30%) dengan fingerprint Avalon sebagai fitur (X) dan pIC50 sebagai target (y).

### 5. Evaluasi

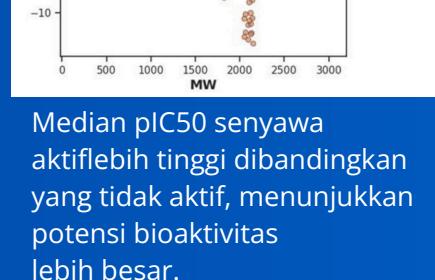
Akurasi: Proporsi prediksi benar.

F1-Score: Rata-rata precision dan recall.

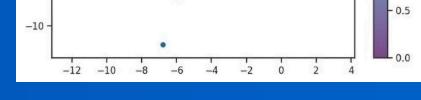
AUC-ROC: Kemampuan model membedakan kelas positif dan negatif.

## Hasil dan Pembahasan

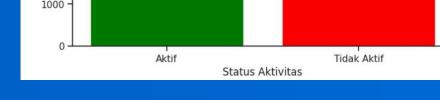
- Sebagian besar senyawa aktif memiliki MW 300-800 dan LogP 0-7.
- Senyawa dengan nilai MW > 1000 cenderung tidak aktif.



Median pIC50 senyawa aktif lebih tinggi dibandingkan yang tidak aktif, menunjukkan potensi bioaktivitas lebih besar.



- Cluster 0: Senyawa dengan heteroatom seperti nitrogen dan klorin (3.558 senyawa).
- Cluster 1: Senyawa dengan gugus karboksil dan amida (2.494 senyawa).
- Cluster 2: Senyawa dengan hidroksi dan fenil (1.323 senyawa).
- Cluster 3: Struktur fenol dan alkohol (1.242 senyawa).



Dataset terdiri dari 2.219 senyawa aktif dan 6.398 senyawa tidak aktif. Model pembelajaran yang digunakan, yaitu LightGBM (LGBM), menunjukkan kinerja yang cukup baik dalam mengklasifikasikan data.

### Metrik Evaluasi

Akurasi (Train): 97%.

Akurasi (Test): 93%.

F1-Score: 93%.

ROC AUC: 95%.

## Kesimpulan

K-Means Clustering berhasil untuk mengelompokkan senyawa berdasarkan karakteristik strukturalnya. Selain itu, model LGBM Classifier menunjukkan kinerja yang baik dalam mengklasifikasi senyawa dengan tingkat akurasi yang tinggi.

Berdasarkan validasi silang, model ini mencapai akurasi 93%, F1-Score 93%, dan ROC-AUC 95%, menunjukkan keandalan dalam membedakan senyawa aktif dan tidak aktif dengan presisi yang tinggi.

## Saran

Pengembangan dataset dapat dilakukan dengan memperluas cakupan data yang lebih beragam. Pendekatan lebih lanjut melibatkan eksplorasi berbagai jenis fingerprint molekul, seperti Atom Pair, Topological Torsion, atau Morgan Circular, untuk menggali hubungan struktur dan aktivitas secara lebih mendalam. Selain itu, performa model dapat ditingkatkan dengan membandingkan algoritma, seperti Random Forest, XGBoost, atau Neural Networks, serta mengoptimalkan hyperparameter untuk meningkatkan akurasi dan keandalan model.