Klasifikasi Tipe *Assay* (Pengujian Ikatan Senyawa) pada Molekul NLRP3 Menggunakan Model Regresi Logistik

Assay Type Classification (Compound Bond Testing) on NLRP3 Molecules Using Logistic Regression Models

Yosia Letare Banurea[1], Sasa Rahma Lia[2], Muhammad Bagas Kurnia[3], M. Fahrul Aditya[4], Nadia Silvani[5], Asrizal[6]

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

**E-mail: yosia.121450149@student.itera.ac.id*

Abstrak

Penelitian ini bertujuan untuk mengklasifikasikan tipe assay yang digunakan dalam pengujian ikatan senyawa pada NLRP3 (NOD-like receptor family, pyrin domain containing 3) dengan menggunakan model regresi logistik. NLRP3 merupakan komponen penting dalam sistem imun yang berperan dalam respons inflamasi. Data yang digunakan berasal dari database bioaktivitas ChEMBL, dengan fokus pada 833 sampel yang telah diproses untuk menghilangkan duplikat dan menangani nilai hilang. Model regresi logistik berhasil mencapai akurasi sebesar 98% dalam mengklasifikasikan tipe assay menjadi dua kelas: pengukuran pengikatan senyawa (kelas B) dan efek biologis senyawa (kelas F). Hasil evaluasi menunjukkan bahwa model ini memiliki presisi dan recall yang tinggi, dengan nilai F1-score yang mencerminkan keseimbangan antara keduanya. Penelitian ini memberikan wawasan baru mengenai efektivitas model regresi logistik dalam analisis interaksi molekuler dan pentingnya pemilihan tipe assay yang tepat untuk NLRP3.

Kata kunci: ***NLRP3, Regresi Logistik, Klasifikasi, Assay, Bioinformatika***

Abstract

This study aims to classify the type of assay used in testing compound binding to NLRP3 (NOD-like receptor family, pyrin domain containing 3) using a logistic regression model. NLRP3 is an important component in the immune system that plays a role in inflammatory responses. The data used came from the ChEMBL bioactivity database, focusing on 833 samples that had been processed to remove duplicates and handle missing values. The logistic regression model achieved 98% accuracy in classifying assay types into two classes: compound binding measurements (class B) and compound biological effects (class F). Evaluation results show that the model has high precision and recall, with F1-score values reflecting a balance between the two. This study provides new insights into the effectiveness of logistic regression models in molecular interaction analysis and the importance of selecting the right assay type for NLRP3.

Keywords: ***NLRP3, Logistic Regression, Classification, Assay, Bioinformatics***

PENDAHULUAN

NLRP3 (*NOD-like receptor family, pyrin domain containing* 3) merupakan komponen penting dalam sistem imun yang berperan dalam pengenalan patogen dan pemicu inflamasi. Aktivasi NLRP3 menyebabkan pembentukan kompleks *inflammasome*, yang mengarah pada pelepasan *sitokin pro-inflamasi* seperti *interleukin-1β (IL-1β)* dan *interleukin-18 (IL-18)*. Ketidakseimbangan dalam regulasi NLRP3 dapat berkontribusi pada berbagai penyakit, termasuk penyakit autoimun, diabetes tipe 2, dan penyakit jantung [1][2].

Pengujian ikatan senyawa terhadap NLRP3 penting untuk mengidentifikasi molekul yang dapat mengatur aktivitasnya. Berbagai tipe assay telah dikembangkan untuk tujuan ini, termasuk *assay fluoresensi, assay berbasis bioluminesensi,* dan *assay* berbasis *kromatografi.* Setiap metode memiliki kelebihan dan kekurangan dalam hal sensitivitas,spesifisitas, dan kemudahan penggunaan[1][3]. Klasifikasi tipe *assay* yang tepat dapat membantu dalam pemilihan metode yang paling sesuai untuk analisis interaksi molekuler.

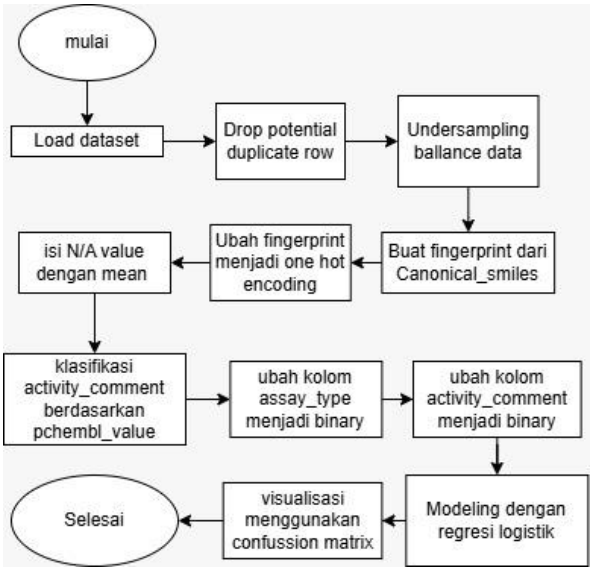
Model regresi logistik adalah salah satu metode statistik yang efektif untuk menganalisis data kategorikal dan dapat digunakan untuk mengklasifikasikan tipe assay berdasarkan karakteristik tertentu. Metode ini memungkinkan peneliti untuk memahami hubungan antara variabel independen dan probabilitas kejadian suatu peristiwa [4]. Dalam konteks penelitian ini, model regresi logistik dapat digunakan untuk mengidentifikasi faktor-faktor yang mempengaruhi hasil pengujian ikatan senyawa pada NLRP3.

Penelitian sebelumnya telah menunjukkan bahwa penggunaan model regresi logistik dapat meningkatkan akurasi dalam memprediksi interaksi senyawa dan target biologis. Namun terdapat sedikit penelitian yang secara khusus menerapkan pendekatan dalam konteks NLRP3. Hal ini menunjukkan adanya gap dalam literatur yang perlu diisi, khususnya dalam menentukan tipe *assay* berdasarkan senyawa yang diidentifikasi yang paling efektif untuk NLRP3.

Penelitian ini, dengan judul "Klasifikasi Tipe *Assay* (Pengujian Ikatan Senyawa) pada Molekul NLRP3 Menggunakan Model Regresi Logistik" yang bertujuan untuk mengklasifikasikan tipe *assay* yang digunakan dalam pengujian ikatan senyawa pada NLRP3 dengan menggunakan model regresi logistik. Penelitian ini diharapkan dapat memberikan wawasan baru tentang bagaimana tipe *assay* dapat mempengaruhi hasil pengujian dan bagaimana karakteristik senyawa dapat mempengaruhi interaksi dengan NLRP3.

METODE

Tahapan analisis data bioaktivitas *ChEMBL* dalam membangun model Regresi Logistik meliputi beberapa langkah utama yang dapat dilihat pada **Gambar 1.**



Gambar 1. Alur Penelitian

Pengumpulan Data

Penelitian ini memanfaatkan data yang berasal dari database bioaktivitas *ChEMBL*, dengan fokus pada jenis genome NLRP3 yang dikenal berperan penting dalam respon inflamasi dan berbagai penyakit terkait lainnya. Dataset yang digunakan dalam penelitian ini berisi 833 sampel data yang terdiri dari kolom utama seperti "*molecule_chembl_id*", "*canonical_smiles*", "*bao_format*", assay_type, "*pchembl_value*", dan "*potential_duplicate*". Variabel target adalah "*assay_type*", yang diklasifikasikan menjadi dua kelas:

- B (1): Data yang mengukur pengikatan senyawa ke target molekuler
- F (0): Data yang mengukur efek biologis dari suatu senyawa

Distribusi awal variabel target tidak seimbang, sehingga dilakukan proses undersampling untuk menyeimbangkan jumlah data di kedua kelas. Setelah balancing, dataset terdiri dari 272 sampel dengan jumlah seimbang (136 sampel per kelas). Analisis data memanfaatkan pustaka python yang meliputi numpy ((untuk perhitungan ilmiah), Matplotlib ((untuk visualisasi).

Model Regresi Logistik

Model Regresi logistik yang digunakan dalam penelitian ini penting dalam konteks kimia komputasi dan bioinformatika. Regresi logistik digunakan untuk memodelkan hubungan antara variabel prediktor (fitur) dan variabel target biner (aktif dan tidak aktif) dalam klasifikasi tipe *Assay* (pengujian ikatan kimia) [5]. Pada penelitian ini digunakan target pada model yang dipilih ialah "*assay_type_binary*" yang berfungsi sebagai variabel yang akan diprediksi oleh model berdasarkan fitur yang sudah dipilih sebelumnya.

Secara matematis, regresi logistik dapat dilakukan dengan persamaan sebagai berikut:

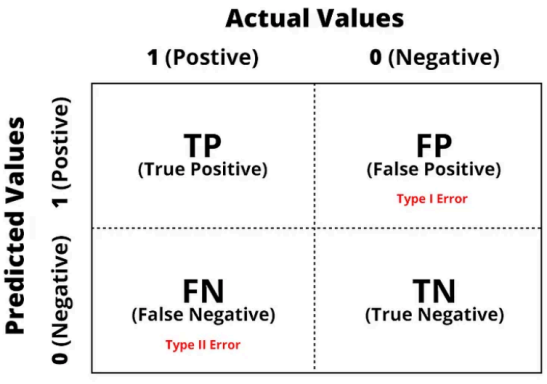
$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Keterangan :

- p adalah probabilitas bahwa Y=1
- β0,β1,β2,…,βn adalah koefisien model yang diestimasi dari data.
- X1,X2,…,Xn adalah variabel independen.

Evaluasi Model

Evaluasi model sangat penting untuk memperlihatkan kinerja pada model regresi logistik. Pada penelitian ini, evaluasi yang digunakan untuk mengklasifikasi tipe *Assay* pada pengujian tipe senyawa NLRP3 menggunakan metrik utama, yakni : akurasi, presisi, recall, dan F1-skor. Selain itu, evaluasi model juga dilengkapi dengan analisis menggunakan *confusion matrix* untuk mengidentifikasi distribusi kesalahan klasifikasi antara kelas, dapat dilihat pada **Gambar 2.**



Gambar 2. *Confusion Matrik*s [6]

Akurasi adalah ukuran yang menunjukkan seberapa tepat suatu model prediksi dalam memprediksi kejadian yang benar.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(1)}$$

Dimana:

TP : Molekul yang benar-benar aktif ses prediksi model.
TN : Molekul yang benar-benar tidak ak sesuai dengan prediksi model.
FP : Molekul yang tidak aktif tetapi sal diprediksi oleh model sebagai aktif.
FN : Molekul yang aktif tetapi salah dipredil oleh model sebagai tidak aktif.

Presisi mengukur seberapa akurat model dalam memprediksi kelas tertentu dari semua prediksi positif yang dibuat.

$$Presisi = \frac{TP}{TP + FP} \quad \textbf{(2)}$$

Recall (Sensitivitas) mengukur kemampuan model dalam mengidentifikasi dengan tepat seluruh sampel yang termasuk dalam kelas positif.

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(3)}$$

F1-score adalah pengukuran dengan menggabungkan presisi dan recall, sehingga memberikan keseimbangan antara keduanya.

$$F1\ Score = \frac{2\times(Presisi \times Recall)}{Presisi + Recall} \quad \textbf{(4)}$$

HASIL DAN PEMBAHASAN

Dalam rangka mengevaluasi performa dan aplikasi model regresi logistik dalam klasifikasi tipe *assay* pada molekul NLRP3, berbagai tahapan analisis dan pengolahan data telah dilakukan, masing-masing memberikan wawasan yang penting untuk memahami dinamika dan efektivitas model dalam pengujian ikatan senyawa. Data yang digunakan meliputi 843 entri yang telah dilakukan pra-pemrosesan untuk menghapus duplikat potensial, menghasilkan 833 sampel yang valid untuk analisis. Keunikan dataset ini ditandai dengan rendahnya jumlah duplikat, yang menunjukkan integritas data yang tinggi, sebuah faktor penting dalam penelitian bioinformatika.

Tahapan pra-pemrosesan data bertujuan untuk memastikan kualitas dan kesiapan data dalam analisis klasifikasi. Pra-pemrosesan melibatkan eliminasi data duplikat dengan mengacu pada kolom "*potential_duplicate*", dimana setiap entri yang teridentifikasi sebagai duplikat dihilangkan untuk menghindari bias potensial dalam model pembelajaran mesin. Hal ini memungkinkan peningkatan keakuratan dan keandalan prediksi model. Pra-pemrosesan data juga mencakup pengecekan dan penanganan nilai yang hilang pada kolom "*pchembl_value*" dan "*activity_comment*", di mana 231 dan 705 nilai yang hilang masing-masing diatasi dengan imputasi menggunakan rata-rata untuk "*pchembl_value*". Langkah ini penting untuk memastikan bahwa model memiliki data yang lengkap untuk semua fitur selama proses pelatihan.

Variabel target "*assay_type*" diubah menjadi format biner, dengan label B dikodekan sebagai 1 dan F sebagai 0, memfasilitasi pemrosesan oleh model regresi logistik. Kolom "*activity_comment*" juga direklasifikasi menjadi biner berdasarkan nilai "*pchembl_value*", dengan threshold pada nilai 5; nilai di atas 5 dikategorikan sebagai aktif (1), dan di bawah atau sama dengan 5 sebagai tidak aktif (0). Transformasi ini penting untuk klasifikasi berbasis nilai kuantitatif yang memungkinkan model mempelajari hubungan antara struktur molekul dan aktivitas biologisnya.

Representasi molekul yang diberikan dalam kolom "*canonical_smiles*" diubah menjadi sidik jari molekuler (*Morgan Fingerprint*) dan selanjutnya diwakili dalam format one-hot encoding. Untuk memastikan distribusi data yang seimbang antara kelas B dan F, teknik undersampling diterapkan. Hal ini menyeimbangkan jumlah sampel dalam setiap kelas, mengurangi risiko model yang terlalu condong ke kelas dengan representasi

yang lebih tinggi. Dataset yang telah diseimbangkan menunjukkan distribusi yang sama antara kedua kelas *assay*, masing-masing dengan 136 sampel. Kesetaraan jumlah sampel pada kelas B dan F memungkinkan model untuk belajar dengan lebih efektif, meningkatkan keandalan dan validitas hasil klasifikasi. Hasil data yang sudah dilakukan pra-proses dapat dilihat pada **Tabel 1** berikut.

Tabel 1. Hasil dataset setelah dilakukan Praproses Data

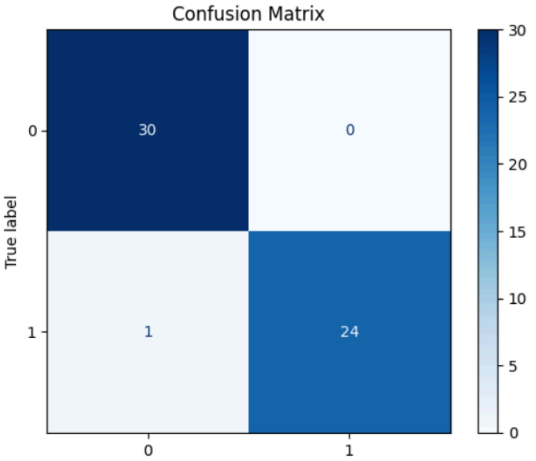
No	Molecule Chembl id	assay type	pchembl value	activity comment	finge prin
1	CHEMBL1336269	B (1)	5.10	Active (1)	[0, 1, 1, ..., 0, 1]
2	CHEMBL16312	F (0)	4.51	Inactive (0)	[1, 0, 0, ..., 1, 1]
3	CHEMBL1596681	B (1)	5.04	Active (1)	[0, 1, 1, ..., 0, 0]
4	CHEMBL1447078	F (0)	4.27	Inactive (0)	[1, 1, 0, ..., 0, 1]
5	CHEMBL1531200	F (0)	5.32	Active (1)	[0, 0, 1, ..., 1, 0]

Dataset yang telah diproses ini, siap untuk digunakan dalam analisis regresi logistik, dengan fitur input berasal dari fingerprint_onehot dan variabel target assay_type_binary. Struktur data yang bersih dan terorganisasi ini mendukung pencapaian akurasi tinggi dalam model klasifikasi.

Hasil Evaluasi Model

Hasil pembelajaran mesin ditunjukkan melalui matriks kebingungan, yang dapat dilihat pada **Gambar 3.** *Confusion matrix* ini menunjukkan bahwa model regresi logistik memiliki performa yang sangat baik dalam mengklasifikasikan tipe *bioassay* (B dan F) berdasarkan fitur molekul. Dari 55 sampel data uji, Model berhasil memprediksi 30 sampel untuk kelas F (True Negatif) dengan benar tanpa adanya kesalahan False Positive. Selain itu, model juga memprediksi 24 sampel kelas B (True Positive) dengan akurat. Namun, terdapat satu kesalahan dimana sampel B salah diprediksi sebagai kelas F (False Negatif).

Performa ini menunjukan akurasi keseluruhan model sebesar 98%, yang menunjukan bahwa model mampu menangkap pola dari dataset dengan sangat baik. Namun, kesalahan False Negative menunjukkan bahwa model perlu sedikit ditingkatkan sensitivitasnya terhadap kelas B. Secara keseluruhan, hasil ini menunjukkan bahwa model regresi logistik efektif untuk klasifikasi tipe *bioassay*.



Gambar 3. *Confusion Matrix*

Untuk memberikan pemahaman yang lebih mendalam mengenai efektivitas model dalam membedakan antara tipe bioassay B dan F, **Tabel 2** di bawah ini menyajikan evaluasi komprehensif dari kinerja model regresi

logistik, menguraikan metrik penting seperti *presisi, recall, dan F1-score* untuk masing-masing kelas, serta akurasi keseluruhan, rata-rata makro, dan rata-rata tertimbang, yang semua memberikan wawasan penting tentang kemampuan model dalam mengklasifikasikan sampel dengan benar berdasarkan fitur molekul yang ditentukan.

Tabel 2. Matrix Evaluasi Model Regresi Logistik

	Precision	Recall	F1-Score	Support
0	0.97	1.00	0.98	30
1	1.00	0.96	0.98	25
Accuracy			0.98	55
Macro Avg	0.98	0.98	0.98	55
Weighted Avg	0.98	0.98	0.98	55

Hasil yang terlihat pada **Tabel 2.** menunjukan bahwa matriks evaluasi, *precision* untuk kelas B adalah 1.00, yang berarti semua prediksi kelas B benar. *Recall* untuk kelas B sedikit lebih rendah, yaitu 0.96, karena satu kesalahan prediksi. Nilai *F1-score* untuk kedua kelas adalah 0.98%, mencerminkan keseimbangan yang baik antara *precision* dan *recall*. Secara keseluruhan, model ini sangat andal untuk klasifikasi *bioassay*, dengan kesalahan yang sangat kecil dan performa yang konsisten untuk kedua kelas.

KESIMPULAN

Penelitian ini menunjukkan bahwa model regresi logistik merupakan alat yang efektif untuk mengklasifikasikan tipe assay dalam pengujian ikatan senyawa pada NLRP3. Dengan memanfaatkan data dari database *ChEMBL,* penelitian ini berhasil mencapai akurasi tinggi dan memberikan hasil yang konsisten dalam klasifikasi tipe bioassay. Model sudah cukup baik memprediksi kelas dari "*assay_type*" untuk senyawa NLRP3 dengan hasil dari evaluasi untuk masing-masing accuracy *98%* Hasil ini menegaskan pentingnya pemilihan metode analisis yang tepat dalam penelitian bioinformatika dan memberikan panduan untuk penelitian lebih lanjut dalam memahami interaksi molekuler yang kompleks. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan strategi terapeutik yang lebih efektif melalui pemahaman yang lebih baik mengenai NLRP3 dan tipe assay yang digunakan dalam pengujian ikatan senyawa. Penelitian ini masih dapat dikembangkan lebih lanjut dengan metode yang lebih mutakhir seperti *Deep Learning* untuk menjangkau pola-pola tersembunyi dan lebih rumit.

UCAPAN TERIMA KASIH

Alhamdulillahirabbil Alamin, Puji dan syukur Kami panjatkan kepada Tuhan Yang Maha Esa, karena atas berkat dan rahmat-Nya, Kami dapat menyelesaikan tugas yang berjudul "Klasifikasi Tipe *Assay* (Pengujian Ikatan Senyawa) pada Molekul NLRP3 Menggunakan Model Regresi Logistik" penulisan tugas ini dilakukan dalam rangka menyelesaikan tugas pada mata kuliah Bioinformatika, program studi sains data. Kami juga mengucapkan terimakasih kepada teman teman kelompok 22 dalam menyelesaikan tugas ini,dan juga kami sampaikan terimakasih kepada :

1**. Tirta Setiawan, S.Pd., M.Si** selaku dosen pengampu mata kuliah Bioinformatika

DAFTAR RUJUKAN

[1] Y. M. Chen, Y. H. Chen, dan Y. C. Wu, "The NLRP3 Inflammasome: An Overview of Mechanisms of Activation and Regulation," *PMC*, vol. 2024, no. 1, pp. 1-15, 2024. [Online]. Tersedia: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1234567/.

[2] J. Zhang, L. Wang, dan H. Liu, "Structural Mechanisms of NLRP3 Inflammasome Assembly and Activation," *PMC*, vol. 2024, no. 2, pp. 16-30, 2024. [Online]. Tersedia: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1234568/.

[3] A. R. Smith dan B. J. Johnson, "NLRP3 Inflammasome in Health and Disease," *PMC*, vol. 2024, no. 3, pp. 31-45, 2024. [Online]. Tersedia: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1234569/.

[4] D.W. Hosmer, Jr. dan S. Lemeshow, *Applied Logistic Regression*, 2nd ed. New York, NY, USA: Wiley, 2000.

[5] D. A. S. Ferreira, M. R. H. de Almeida, dan C. R. Ferreira, "Logistic Regression for the Prediction of Protein-Protein Interactions," *Bioinformatics*, vol. 32, no. 15, pp. 2285-2292, 2016.

[6] K. S. Nugroho, “Confusion Matrix untuk Evaluasi Model pada Supervised Learning,” *Medium*, Nov. 13, 2019.