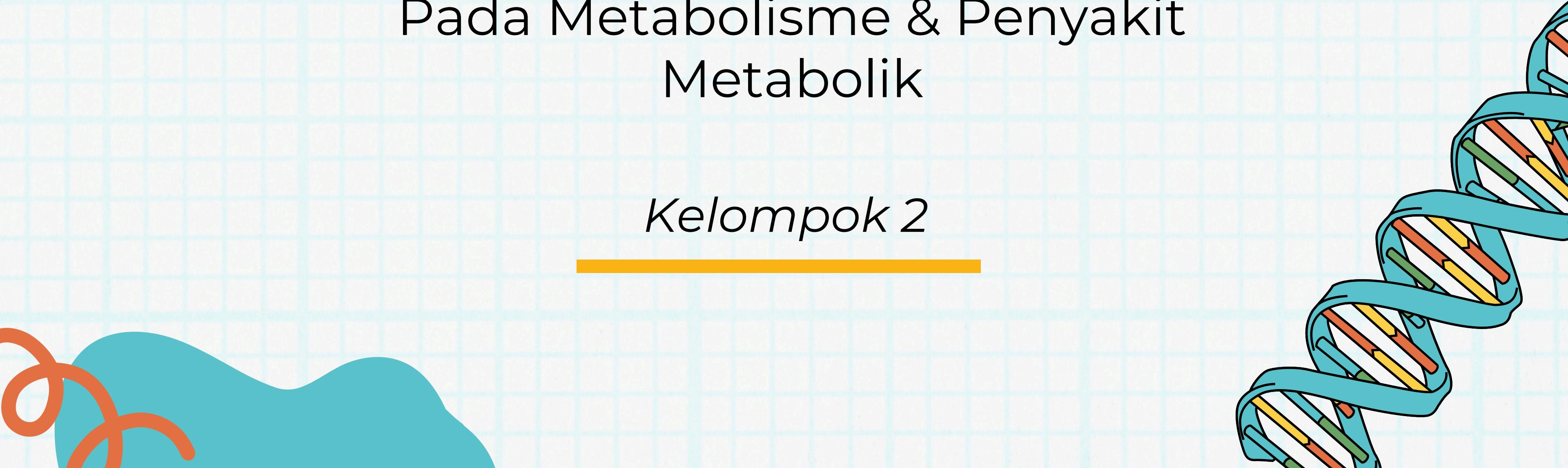




Implementasi Random Forest Untuk Klasifikasi & Menganalisis Aktivasi AMPK (*AMP-Activated Protein Kinase*) Pada Metabolisme & Penyakit Metabolik

Kelompok 2



Our Great Team 02



GHOZI ALVIN KARIM

121450123



M. FAQIH

121450120



M. GILANG MARTIANSYAH M

121450056



NADILLA ANDHARA PUTRI

121450003



LIA ALYANI

121450056



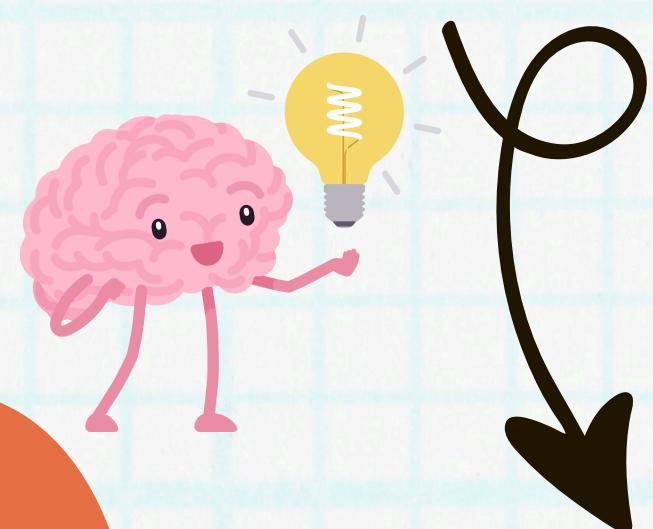
ANNISA DINI AMALIA

1214500081



PENDAHULUAN

Latar Belakang



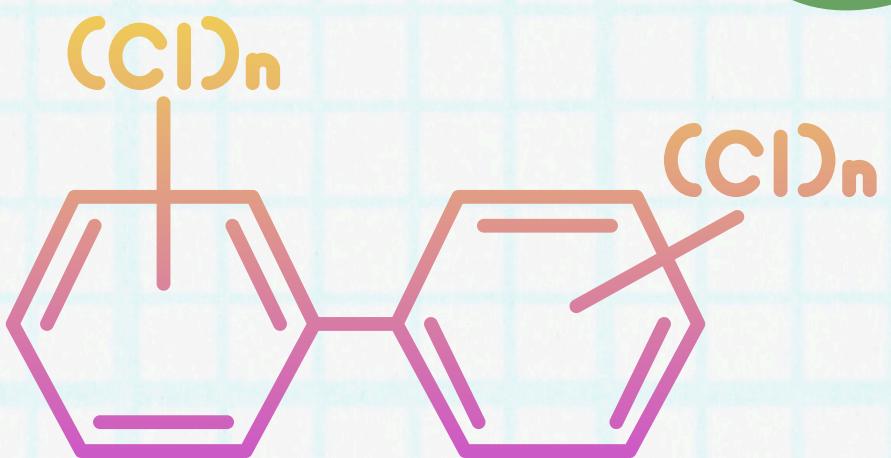
AMP-activated protein kinase (**AMPK**) merupakan enzim yang mengatur homeostasis energi seluler dan menjadi target potensial untuk terapi penyakit metabolismik seperti **Diabetes** tipe 2 dan **Obesitas**. Pendekatan bioinformatika dan algoritma Machine learning, seperti **Random Forest**, efektif dalam seleksi fitur dan klasifikasi data biologis. Teknologi ini meningkatkan pemahaman dan prediksi regulasi AMPK, terutama melalui analisis ekspresi gen di jaringan otot, adiposa, dan hati dalam konteks penyakit metabolismik.

Rumusan Masalah

Bagaimana aktivasi AMPK dapat dianalisis menggunakan algoritma Random Forest?

Apa saja pola regulasi AMPK yang dapat diidentifikasi dari data biologis yang tersedia?

Sejauh mana efektivitas algoritma Random Forest dalam seleksi fitur dan klasifikasi data ekspresi genetik yang terkait dengan AMPK?



Tujuan Penelitian



Menganalisis aktivasi AMPK dalam konteks metabolisme dan penyakit metabolik menggunakan algoritma Random Forest.

Mengidentifikasi pola regulasi AMPK berdasarkan data biologis yang relevan.

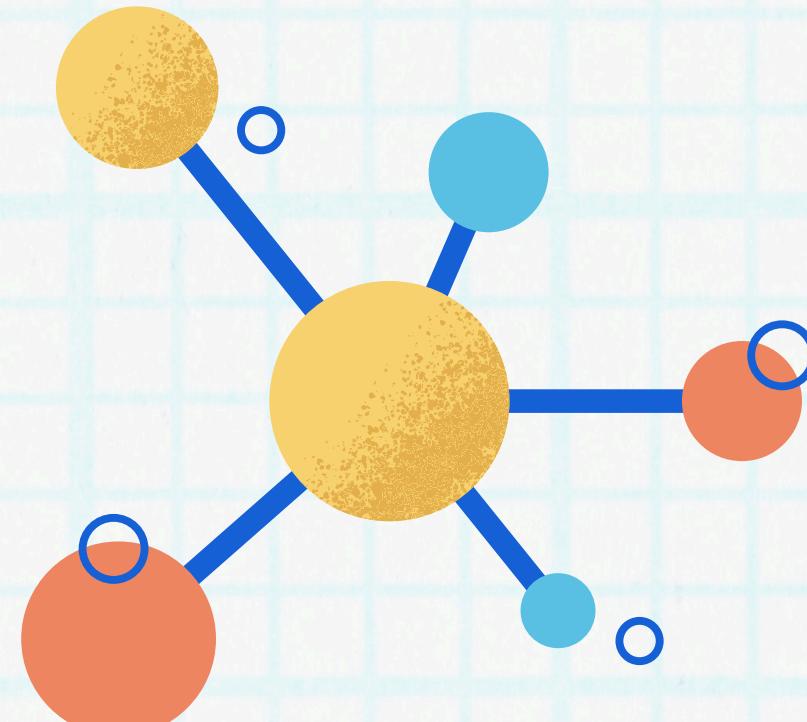
Mengevaluasi efektivitas algoritma Random Forest dalam seleksi fitur dan klasifikasi data ekspresi genetik yang terkait dengan AMPK.

Manfaat Penelitian

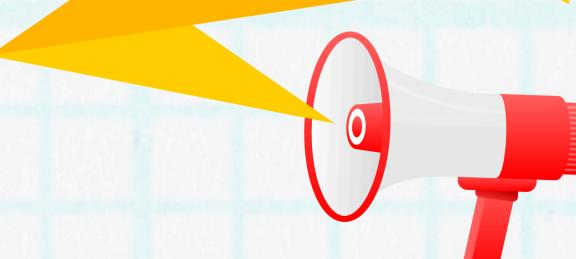
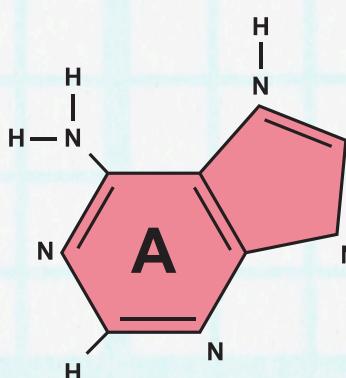
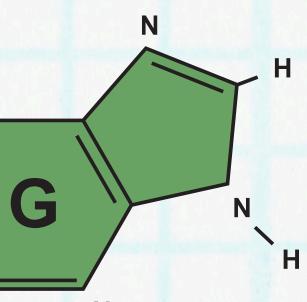
Menyediakan wawasan baru mengenai mekanisme regulasi AMPK yang dapat diterapkan dalam pengembangan terapi penyakit metabolismik.

Memberikan bukti empiris tentang keunggulan algoritma Random Forest dalam analisis data biologis dengan dimensi tinggi.

Menyediakan metode analisis yang dapat digunakan dalam penelitian molekuler lainnya.



BENEFITS

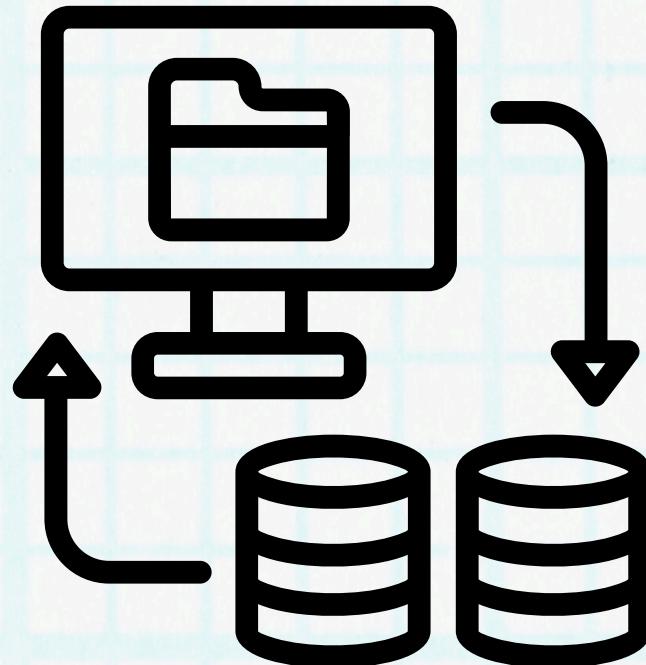




METODE PENELITIAN

Dataset

Dataset yang digunakan pada penelitian ini adalah hasil dari screening dengan bantuan dari code **Python** bioactivity AMPK dari [Chembli](#), data mencakup beberapa fitur yaitu ***MW, LogP, NumHDonors, NumHAcceptors, pIC50*** dan Class Activity sebagai kolom target yang berisikan status aktivasi senyawa AMPK Active dan Inactive. Total dataset terdiri dari 798 baris senyawa dan beberapa fitur molekul serta mencakup deskriptor fisikokimia dan fingerprint molekul.



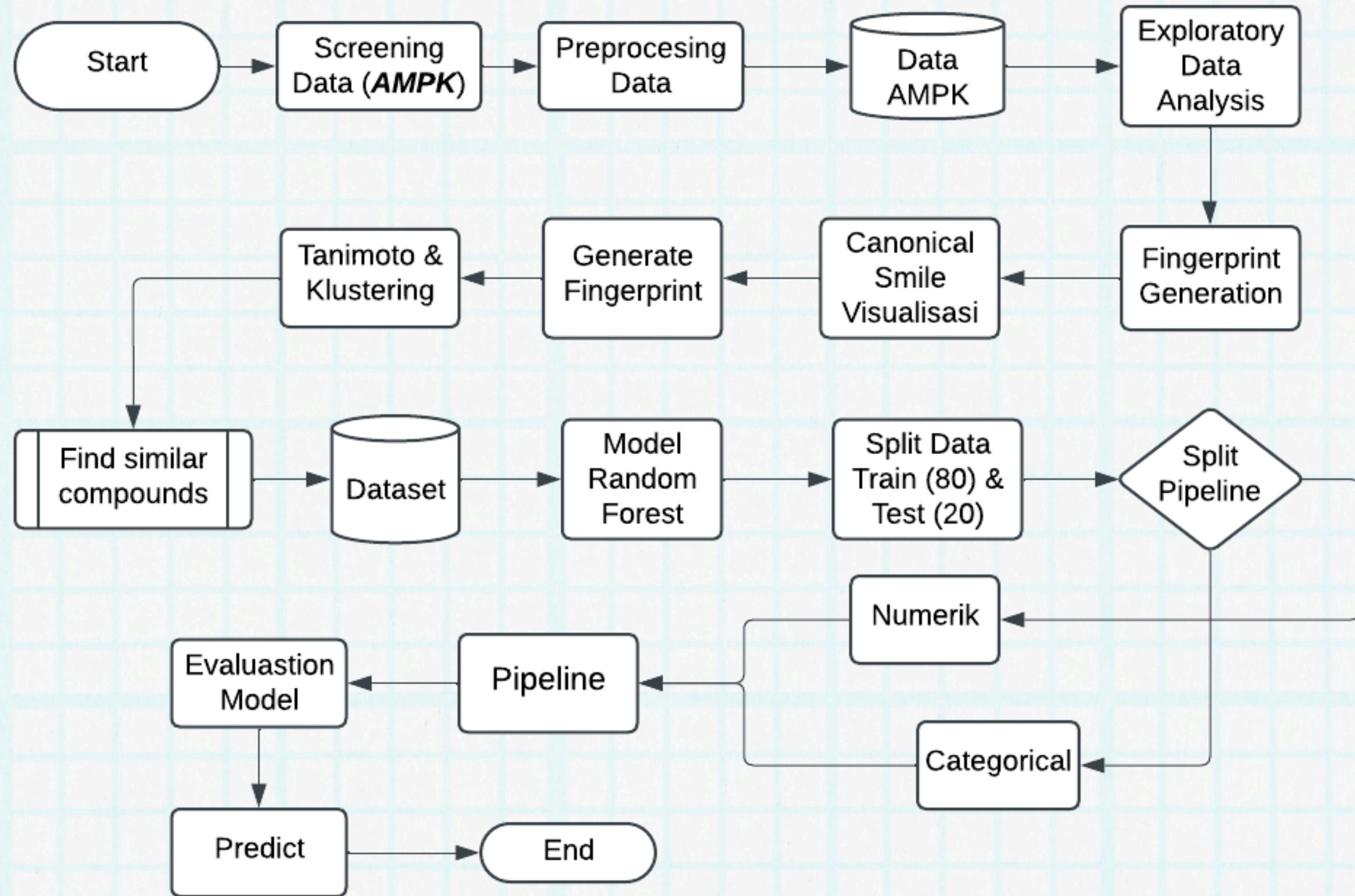
Dataset

molecule_chembl_id	class	canonical_smiles	MW	LogP	NumHDonors	NumHAcceptors	pIC50
1	CHEMBL535	CCN(CC)CCNC(=O)c1c(C)[nH]c(/C=C2\C(=O)Nc3ccc(F...)	398.482	3.33494	3	3	7.207608
2	CHEMBL281957	CCN(CC)C/C=C/c1nc(O)c2c(ccc3nc(Nc4c(Cl)cccc4Cl...)	484.431	6.54092	2	6	6.000000
3	CHEMBL437885	COc1cc2ncnc(Nc3cccc(Cl)c3F)c2cc1CN1CCCC1	386.858	4.77030	1	5	5.000000
4	CHEMBL207815	COCCN1CC(C(N)=O)(N(C)Cc2cc3c(Nc4cccc(Cl)c4F)nc...	502.978	2.79240	2	8	5.000000
5	CHEMBL425601	COc1cc2ncnc(Nc3cccc(Cl)c3F)c2cc1CN(C)C1(C(N)=O...	486.979	3.55440	2	7	5.000000
...
192	CHEMBL4569508	O=C(NCC(F)(F)F)c1cc(-c2cnn3cc(-c4ccc(OCCN5CCCC...)	529.588	5.28160	1	7	8.000000
193	CHEMBL4550702	Cn1cc(-c2cnc3c(-c4csc(C(=O)NCC(F)(F)F)c4)cnn3c...	406.393	3.15040	1	7	7.585027
194	CHEMBL4568087	Cn1cc(-c2cnc3c(-c4csc(C(=O)N[C@@H]5CCCC[C@@H]5...)	421.530	2.85800	2	8	8.337242
195	CHEMBL4552628	Cc1sc(C(=O)N[C@@H]2[C@H](N)CCCC2(F)F)cc1-c1cnn...	425.892	3.66452	2	6	7.275724
196	CHEMBL4634634	C[C@@H]1C[C@H]1C(=O)N1CCN(c2cnc(C#N)c(-c3cnn(C...	365.441	1.44188	0	7	4.000000

	MW	LogP	NumHDonors	NumHAcceptors	pIC50	class_numeric	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	...	PubchemFP873
0	398.482	3.33494	3	3	7.207608	0	1	1	1	1	0	...
1	484.431	6.54092	2	6	6.000000	0	1	1	1	1	0	...
2	386.858	4.77030	1	5	5.000000	1	1	1	1	1	0	...
3	502.978	2.79240	2	8	5.000000	1	1	1	1	1	0	...
4	486.979	3.55440	2	7	5.000000	1	1	1	1	1	0	...

Visualisasi Dataset AMPK

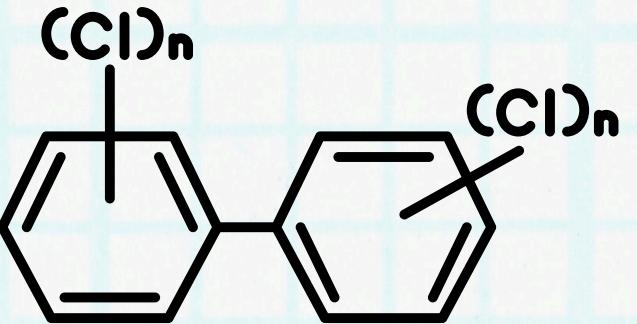
Flowchart



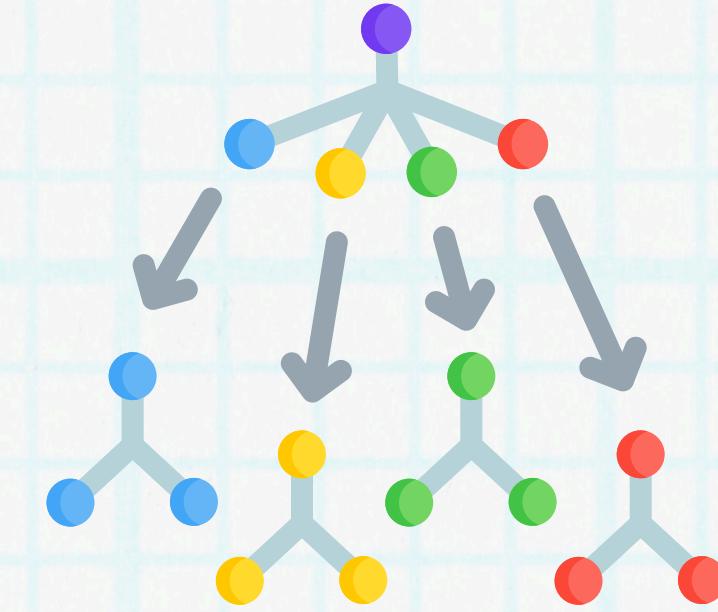
Metode



Fingerprint adalah teknik analisis untuk mengidentifikasi individu berdasarkan pola unik urutan DNA mereka. Pola ini tercipta dari variasi jumlah dan urutan basa nitrogen, sehingga berfungsi sebagai "sidik jari" genetik.



Tanimoto similarity mengukur kesamaan antara dua fingerprint molekuler dan umum digunakan dalam kimia komputasi untuk membandingkan struktur molekul. Dalam penelitian ini, metrik ini membantu penemuan obat dengan membandingkan fingerprint molekul.



Random Forest merupakan metode berbasis koleksi pohon keputusan yang meningkatkan akurasi dengan membangkitkan atribut acak untuk setiap node, mengklasifikasikan data hingga menghasilkan keputusan akhir.

Metrik Evaluasi

Accuracy:

Akurasi adalah metrik yang digunakan untuk menunjukkan tingkat ketepatan suatu model prediksi dalam memprediksi kejadian yang benar.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

Presisi :

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

$$Presisi = \frac{TP}{TP + FP}$$

Recall:

Recall mengevaluasi kemampuan model dalam secara akurat mengidentifikasi semua sampel yang termasuk dalam kelas positif.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score:

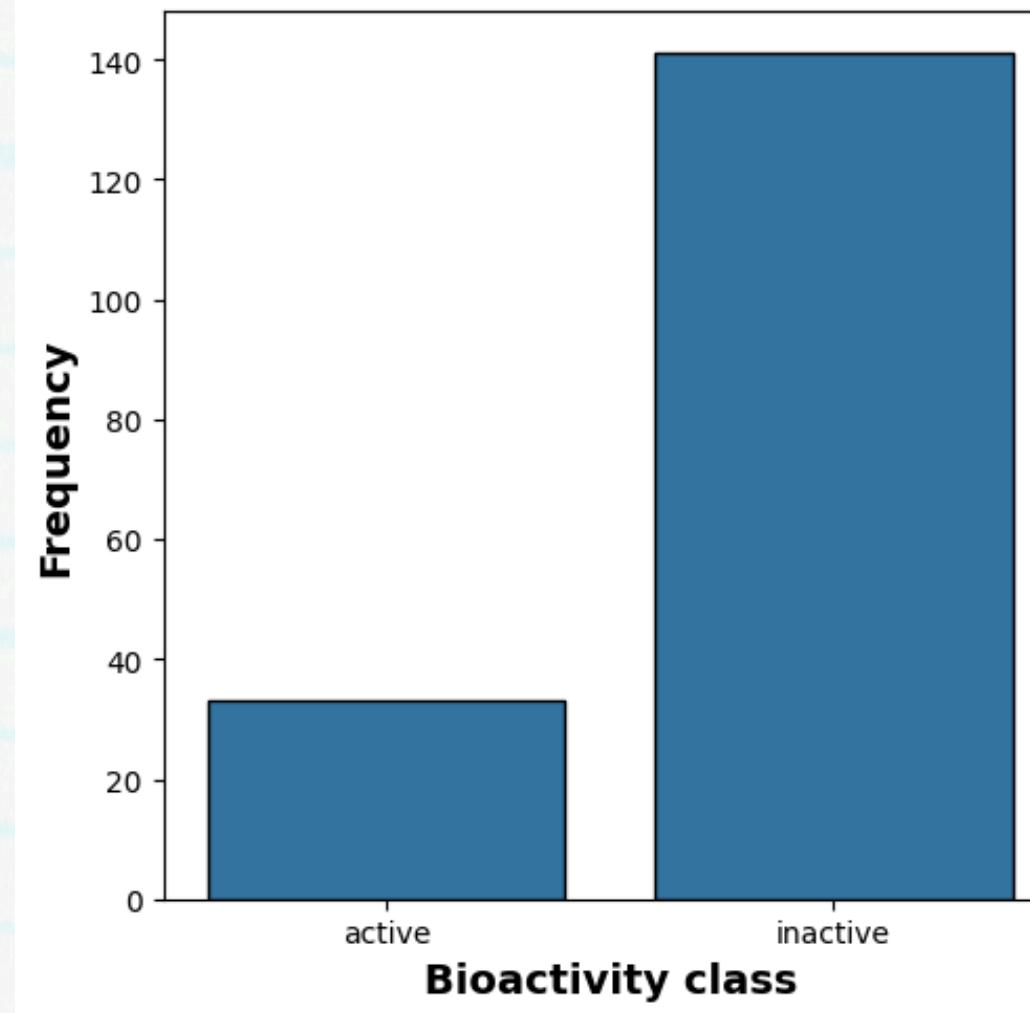
F1-score adalah metrik yang menggabungkan presisi dan recall untuk memberikan keseimbangan antara keduanya.

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

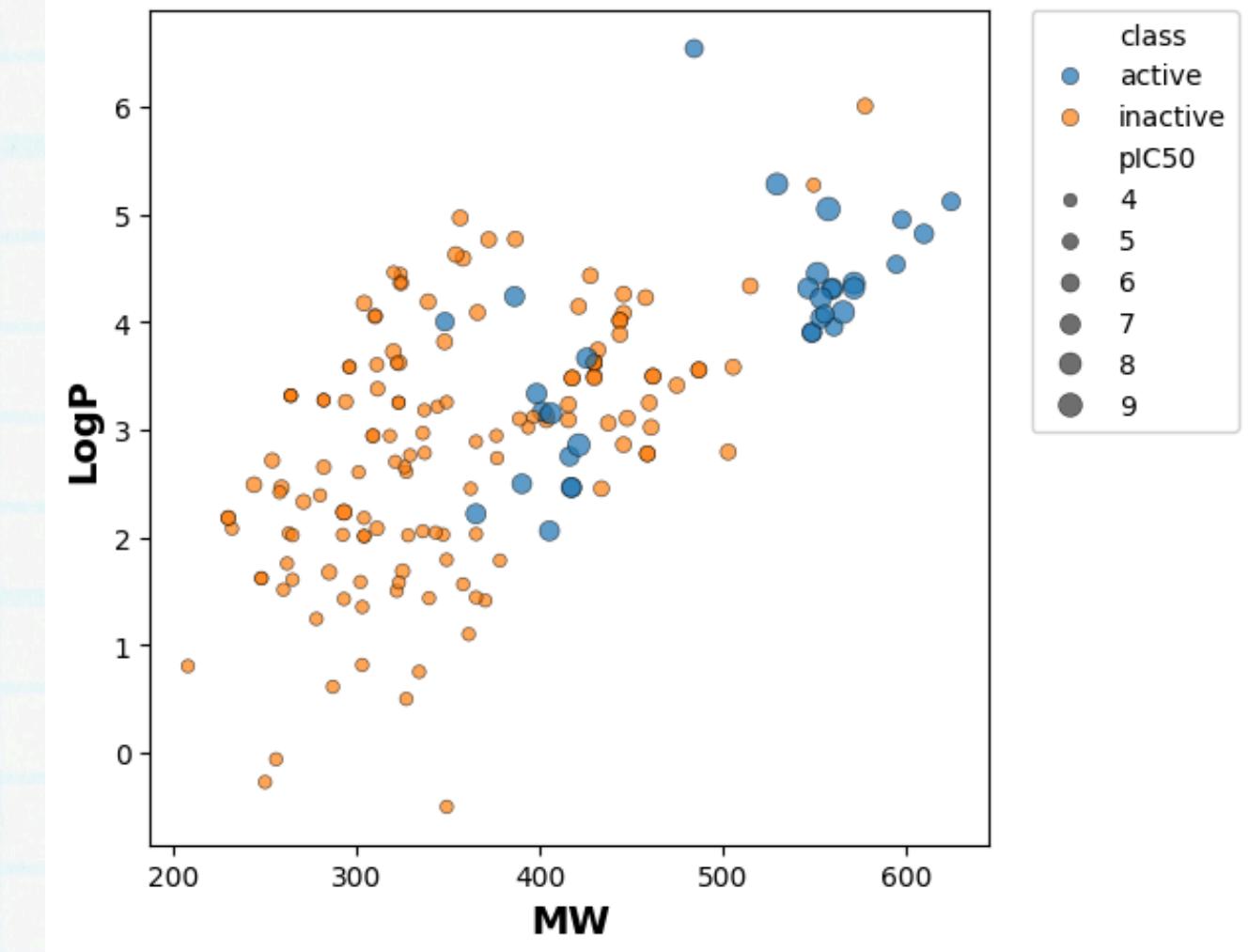


HASIL PEMBAHASAN

Exploratory Data Analysis

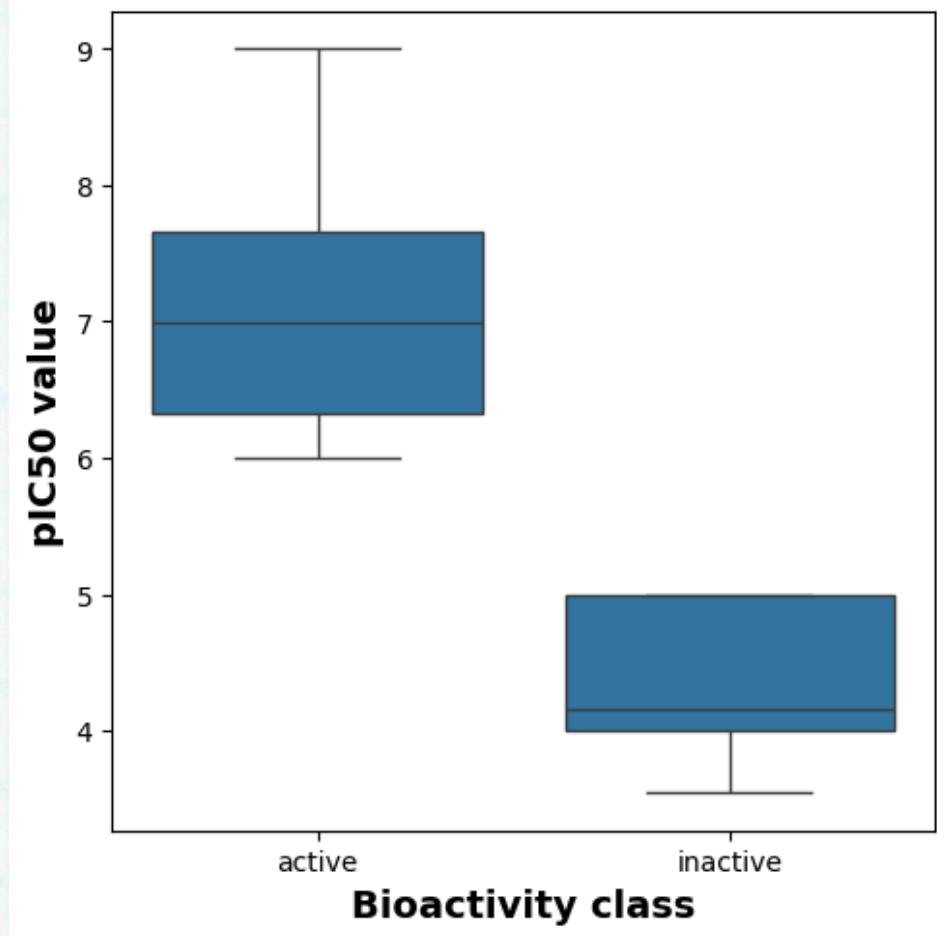


Distribusi Bioactivity class AMPK

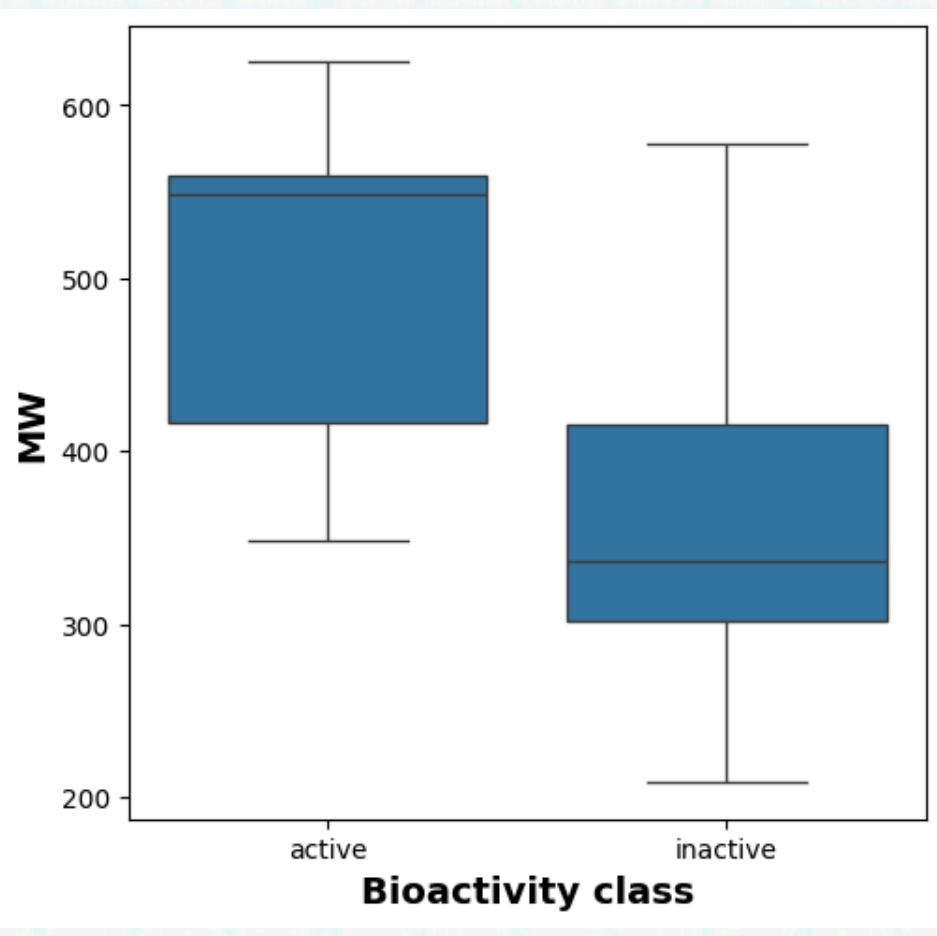


Hubungan antara Berat Molekul MW vs LogP AMPK

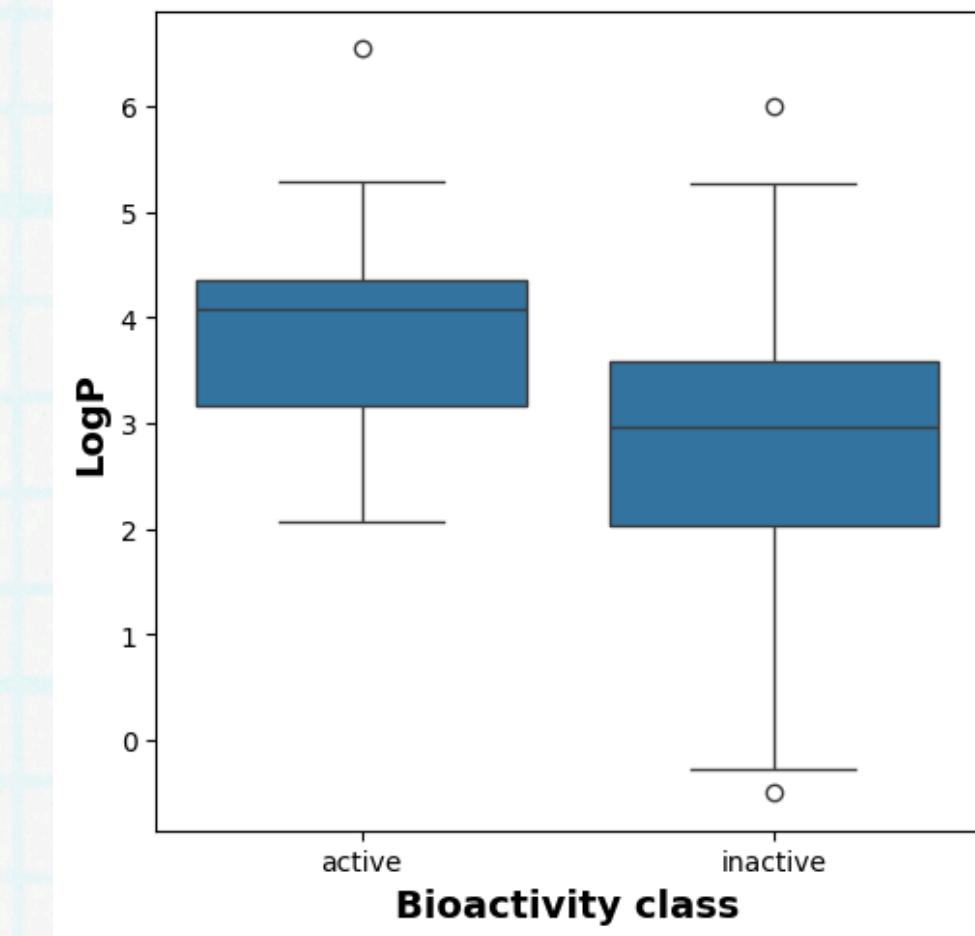
Exploratory Data Analysis



pIC50 AMPK

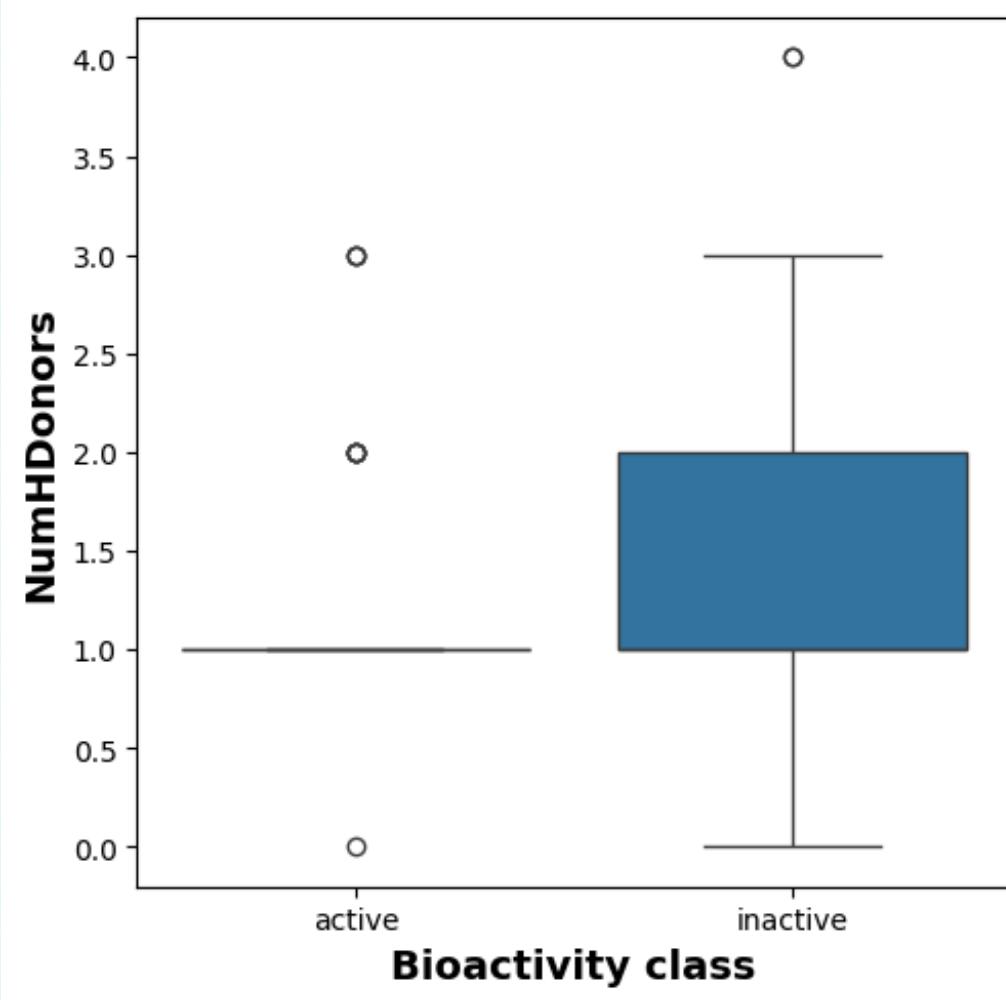


MW AMPK

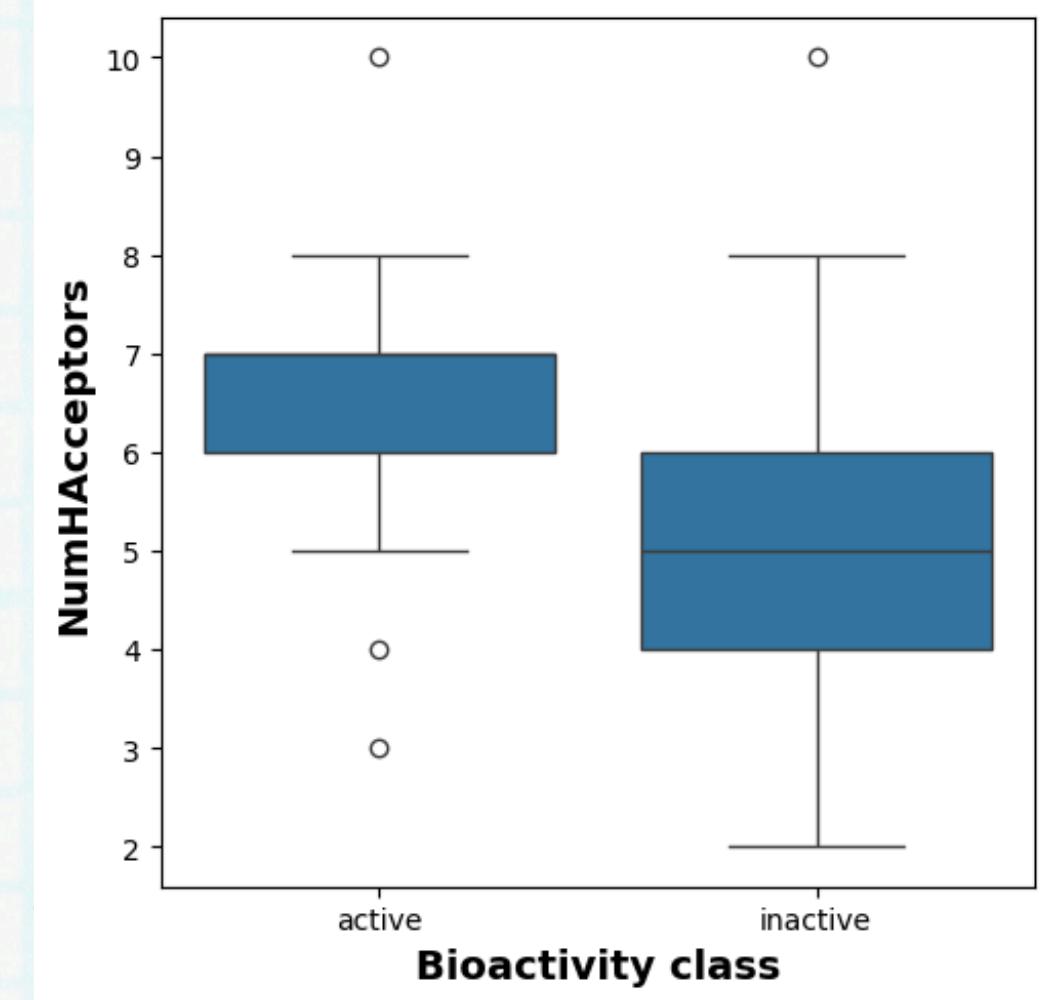


logP AMPK

Exploratory Data Analysis

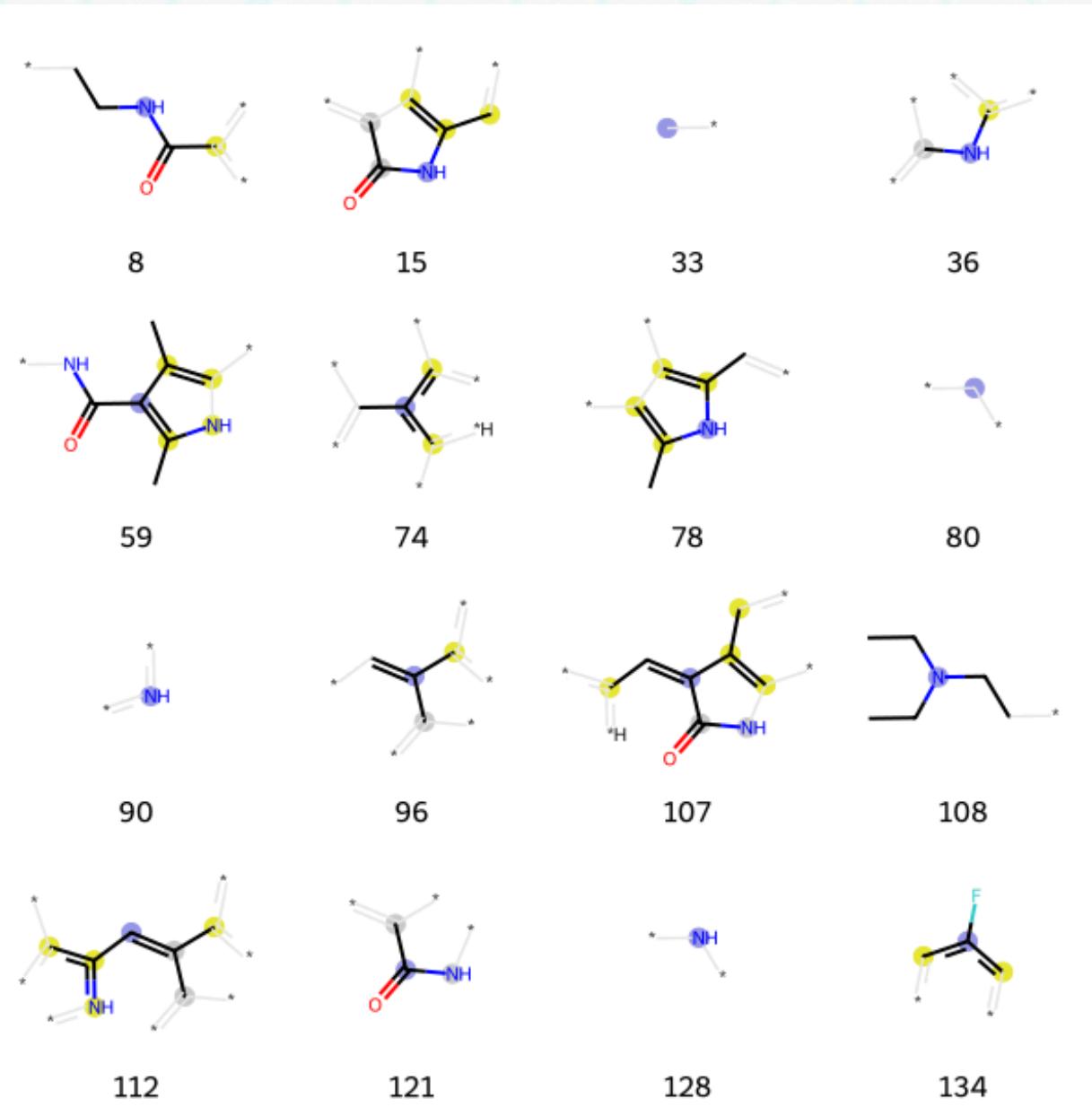


NumHDonors AMPK



NumHAcceptors AMPK

Struktur Molekul Dataset

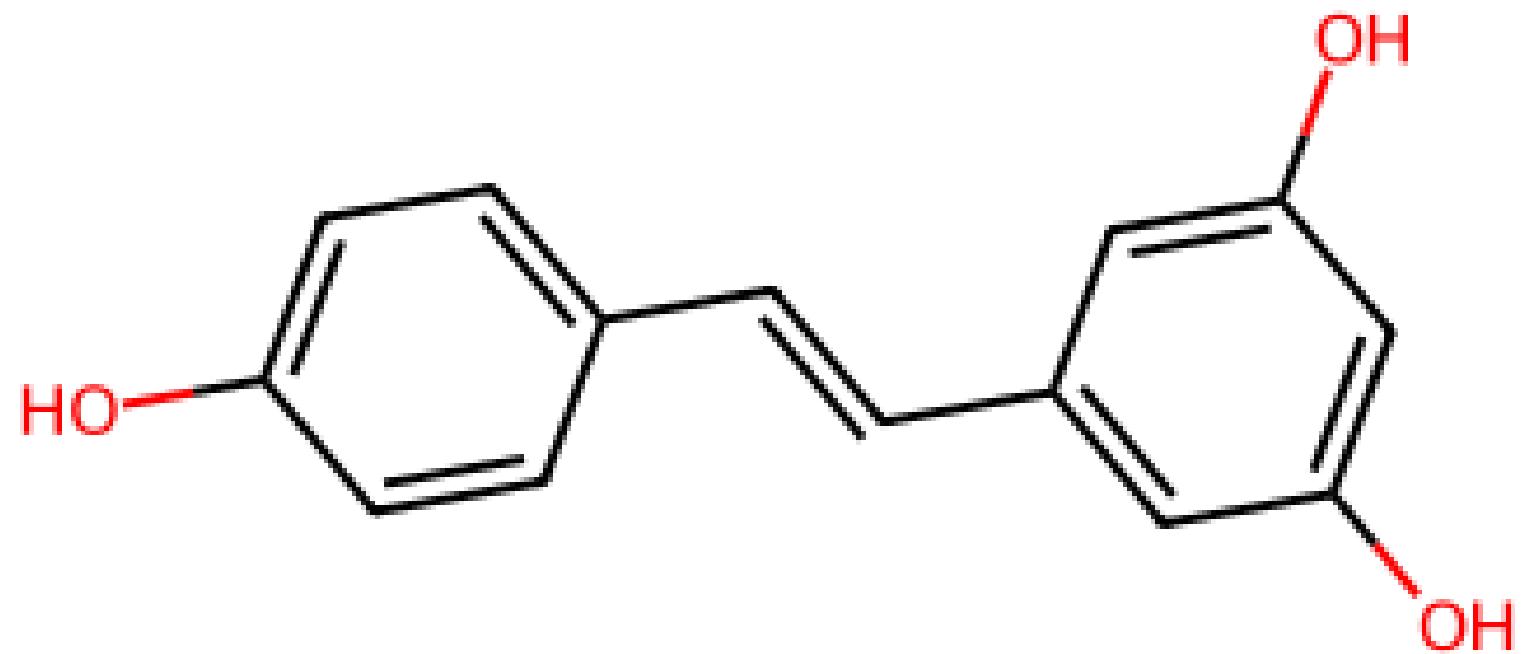


Struktur molekul kimia ditandai dengan "fingerprint" digital untuk analisis, seperti pengelompokan dan prediksi sifat biologis. Keanekaragaman ini mendukung penelitian farmasi dan pengembangan obat.

Visual Fingerprint

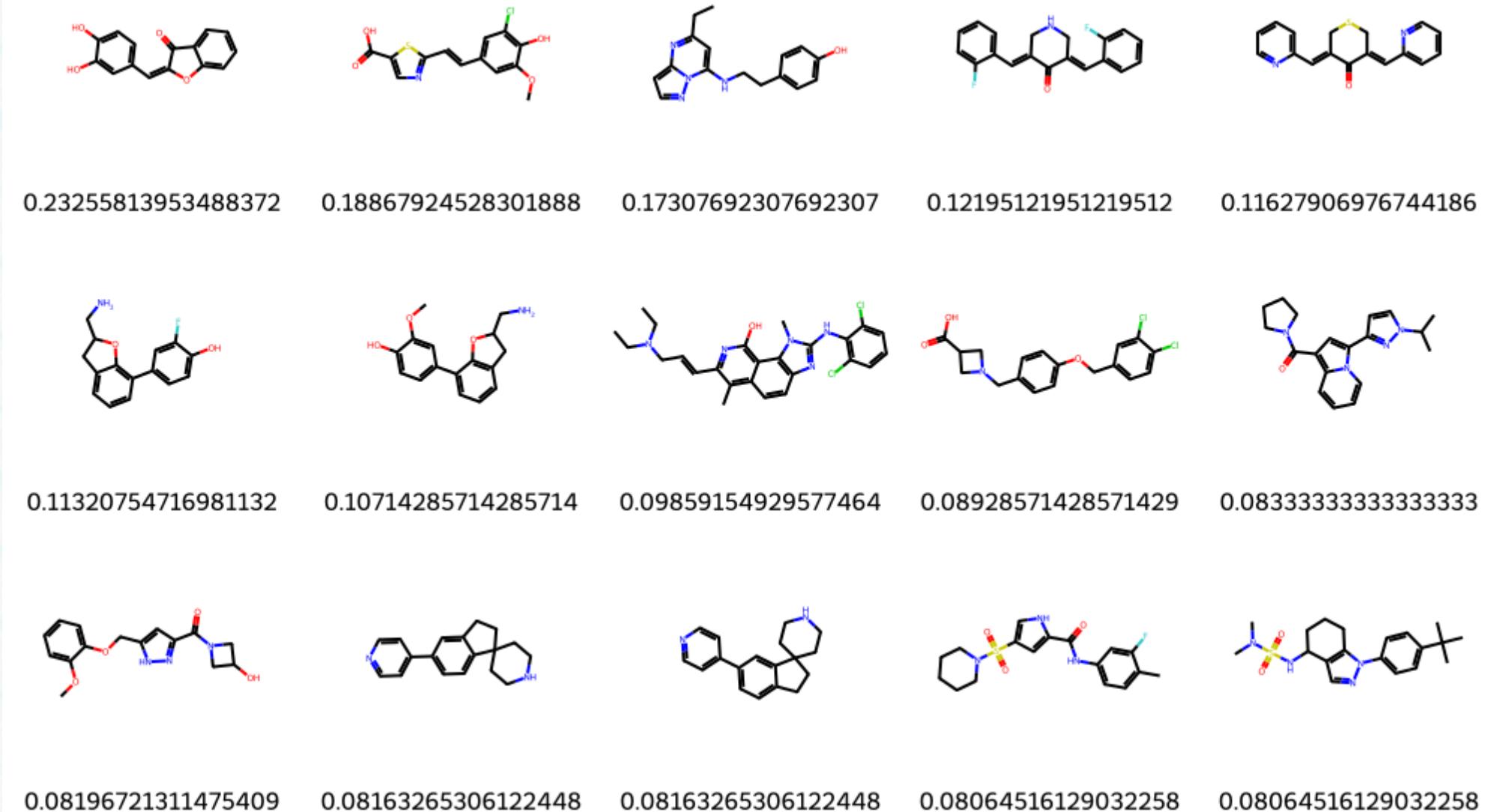
Struktur Kimia Resveratrol

Resveratrol adalah senyawa polifenol dengan aktivitas antioksidan, antiinflamasi, dan antikanker, didukung oleh gugus hidroksil pada cincin aromatik yang menangkal radikal bebas.



Tanimoto Resveratrol

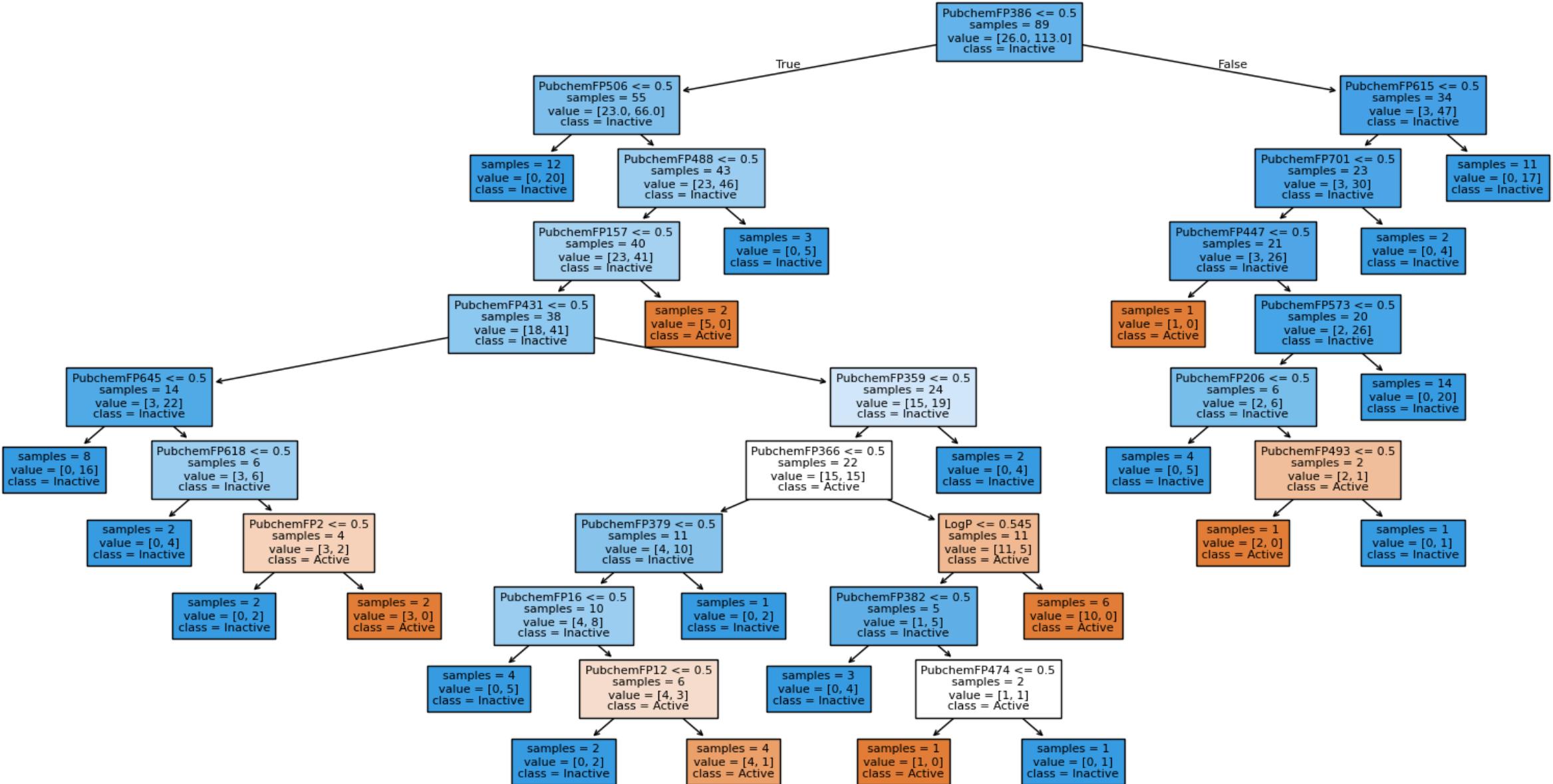
Kemiripan berdasarkan koefisien **Tanimoto** (0.134–0.220) menunjukkan senyawa dalam dataset memiliki kesamaan struktural rendah dengan resveratrol, namun senyawa dengan nilai lebih tinggi dapat menjadi kandidat untuk studi bioaktivitas.



Visual Fingerprint

Hasil Random Forest

Fitur seperti Polar Surface Area dan LogP digunakan sebagai kriteria pemisahan dalam klasifikasi, menunjukkan pengaruhnya terhadap bioaktivitas molekul. Pohon keputusan ini mengungkap parameter utama yang mendukung interaksi molekul dengan target biologis.



Visualisasi Random Forest

Evaluasi Model

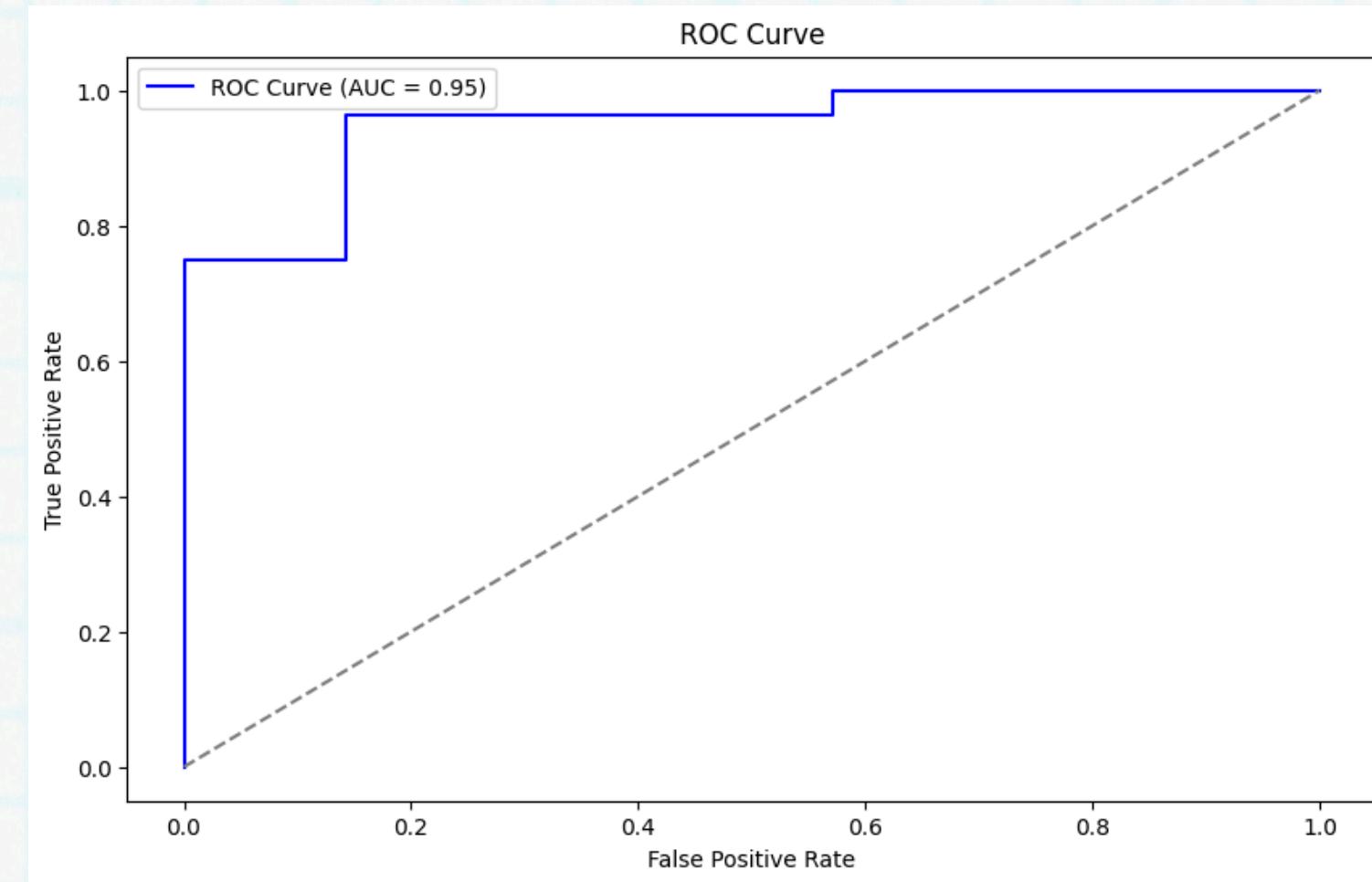
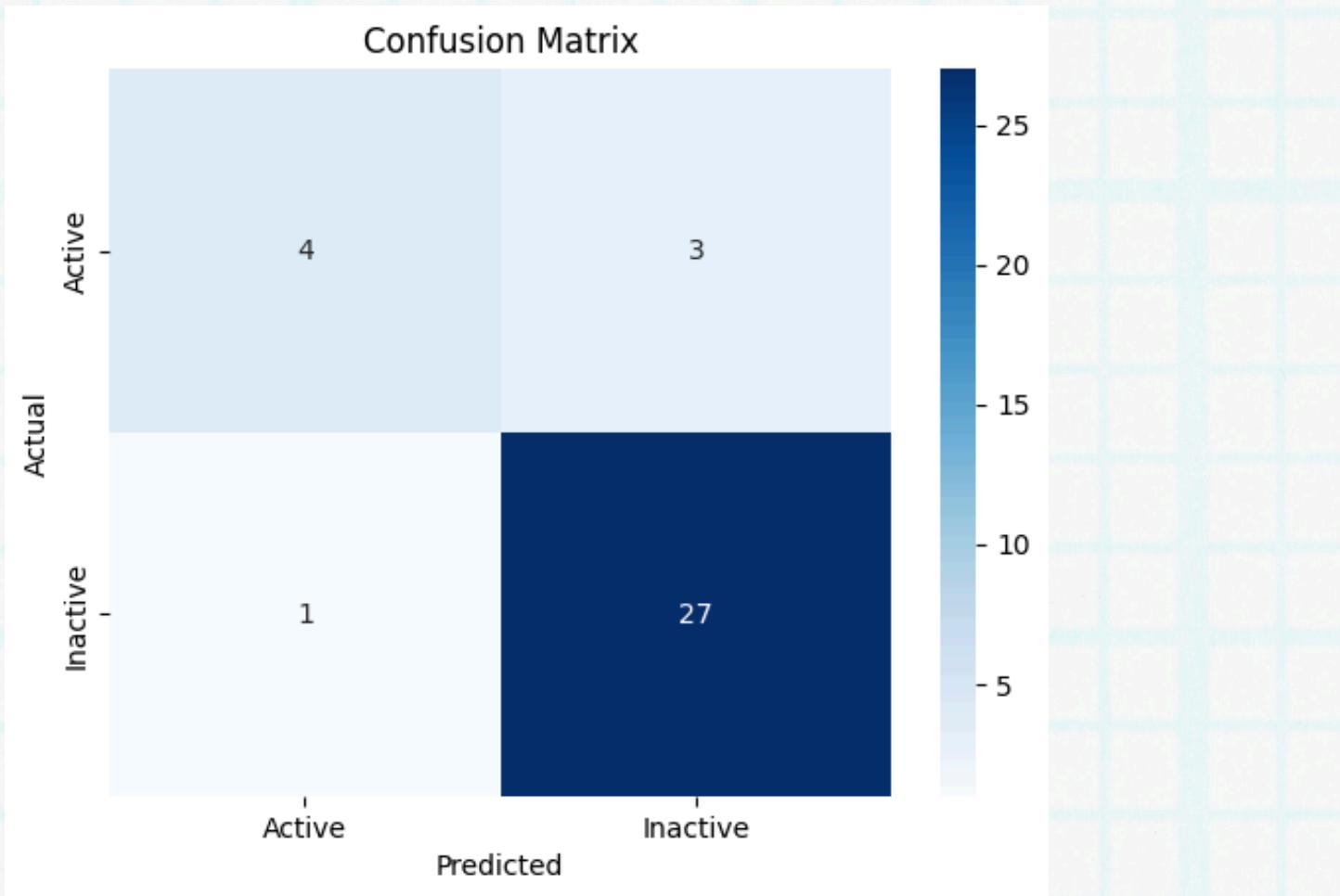
Model memiliki tingkat akurasi yang tinggi dan performa yang solid untuk kelas mayoritas ("Inactive"). Namun, terdapat performa yang lebih rendah pada kelas minoritas ("Active") dengan recall hanya 0.57, yang menandakan bahwa model kurang sensitif dalam mendeteksi kelas "Active." Hal ini berpotensi menyebabkan terlewatnya beberapa data penting.

Kemungkinan, ketidakseimbangan jumlah sampel antara kelas "Inactive" (141 sampel) dan "Active" (33 sampel) berperan dalam mempengaruhi performa ini.

Tabel Classification Report

Parameter	Presisi	Recall	F1-Score	Support
<i>Active</i>	0.80	0.57	0.67	7
<i>Inactive</i>	0.90	0.96	0.93	28
<i>Accuracy</i>			0.89	35
<i>Macro avg</i>	0.85	0.77	0.80	35
<i>Weighted avg</i>	0.88	0.89	0.88	35

Evaluasi Model



Matriks kebingungan menunjukkan kinerja model dalam mengklasifikasikan molekul secara akurat dengan beberapa kesalahan klasifikasi yang minim. Kurva ROC dengan nilai AUC 0.95 menunjukkan kemampuan model yang sangat baik dalam membedakan antara kelas aktif dan tidak aktif, dengan tingkat false positive yang rendah dan true positive rate yang tinggi.



Kesimpulan

Penelitian ini menunjukkan hubungan erat fitur molekuler (MW, LogP, NumHDonors, NumHAcceptors) dengan bioaktivitas molekul. Molekul aktif memiliki MW > 400, LogP 4–6, dan pIC₅₀ > 6,5. Tanimoto similarity menunjukkan kemiripan rendah dengan Resveratrol (0,134–0,220). Algoritma Random Forest mencapai AUC 0,95, dengan MW sebagai fitur paling berpengaruh dalam prediksi model.



Thank
You!