

PREDIKSI AKTIVITAS OBAT SECARA IN SILICO MENGGUNAKAN DESKRIPTOR DAN FINGERPRINT MOLEKULER

Kelompok 1 RA
Bioinformatics, Data Science
Institut Teknologi Sumatera

ANGGOTA KELOMPOK

Dara Cantika Dewi
121450127



Mahdia Nisrina
Maharani M
118140025



Della Septiani
121450109



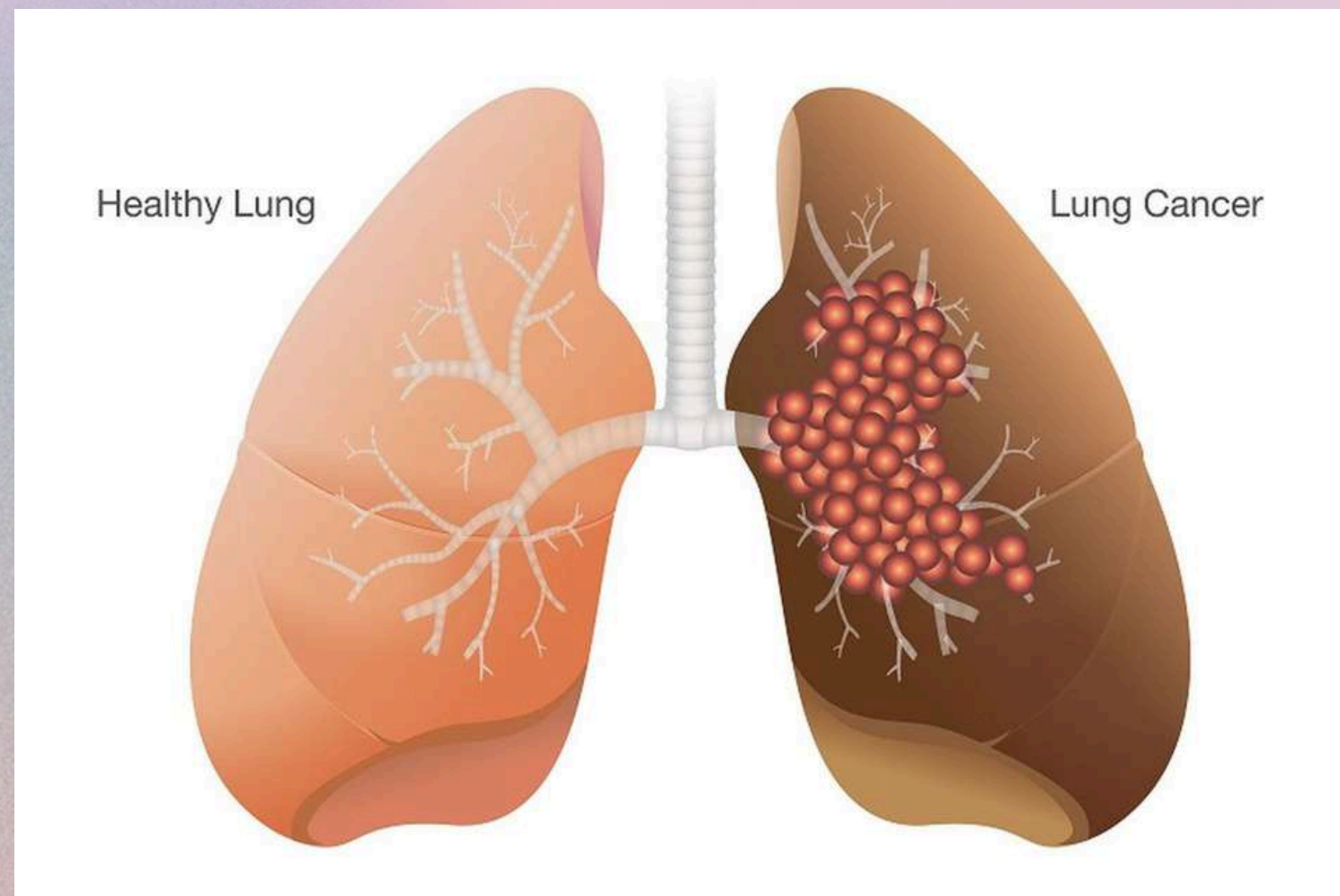
Aisyah Tiara Pratiwi
121450074



Kholisaturrohmah
120450019

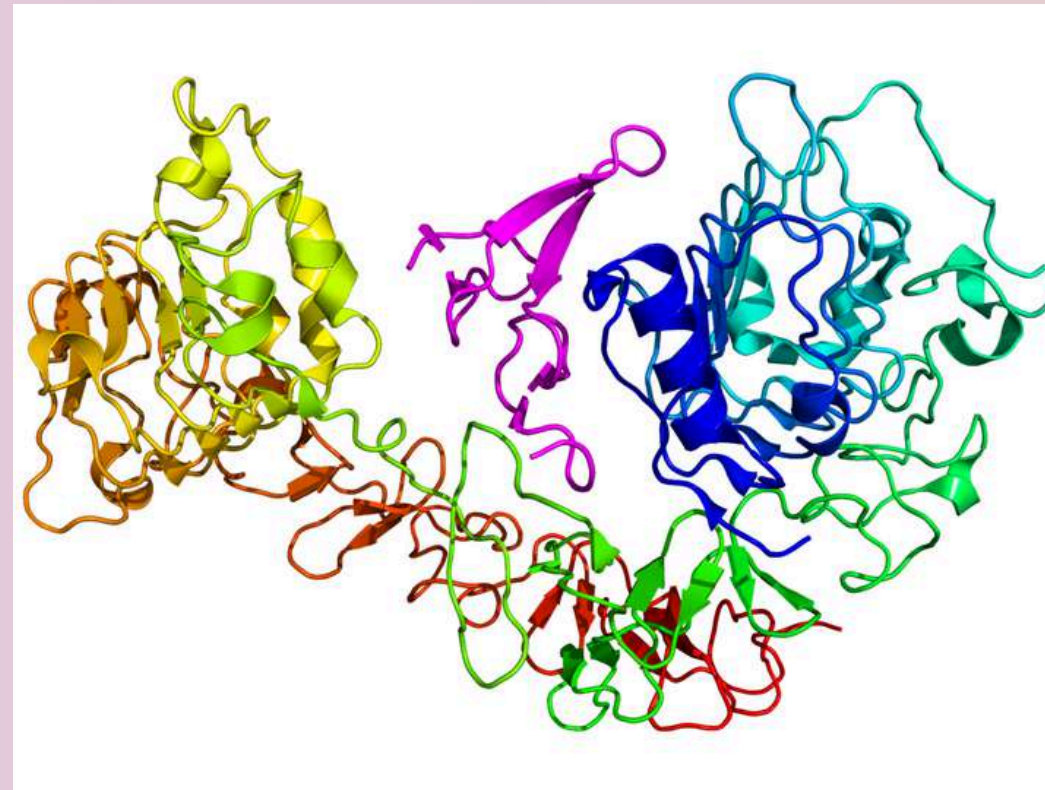


PENDAHULUAN



Kanker paru-paru merupakan penyebab utama kematian akibat kanker, terdiri dari dua kategori utama: Small Cell Lung Cancer (SCLC) dan Non-Small Cell Lung Cancer (NSCLC). Secara global, kanker paru-paru menyumbang 25% dari total kematian akibat kanker pada tahun 2020 (ACS), sementara di Indonesia, kontribusinya mencapai 11,4% dari total kematian akibat kanker (WHO). Prediksi menunjukkan angka kematian akibat kanker paru-paru di Indonesia akan meningkat hingga 83% pada tahun 2040 dibandingkan tahun 2018.

PENDAHULUAN



Epidermal Growth Factor Receptor (EGFR) adalah target terapi penting dalam pengobatan kanker karena perannya dalam regulasi pertumbuhan dan proliferasi sel, namun mutasi atau aktivitas berlebih EGFR sering memicu kanker seperti paru-paru, payudara, dan kolorektal. Oleh karena itu, penghambatan EGFR menjadi strategi utama terapi kanker. Pendekatan *in silico* berbasis pembelajaran mesin, seperti Random Forest Regresi menggunakan dataset senyawa dari ChEMBL, digunakan untuk memprediksi nilai pIC_{50} sebagai ukuran efektivitas penghambatan EGFR. Metode ini menawarkan akurasi tinggi serta efisiensi waktu dan biaya dalam proses penemuan obat, berkontribusi pada pengembangan teknologi komputasional untuk desain obat kanker berbasis pembelajaran mesin.

METODE

Deskripsi Data

Dataset yang digunakan dalam penelitian ini adalah EGFR Bioactivity Dataset, yang terdiri dari 33.727 baris dan 9 kolom. Dataset ini diperoleh dari platform ChEMBL Dataset melalui tautan [ChEMBL](#). Dataset mencakup informasi bioaktivitas senyawa, seperti pengukuran IC50, Ki, dan aktivitas lain terhadap target biologis EGFR.



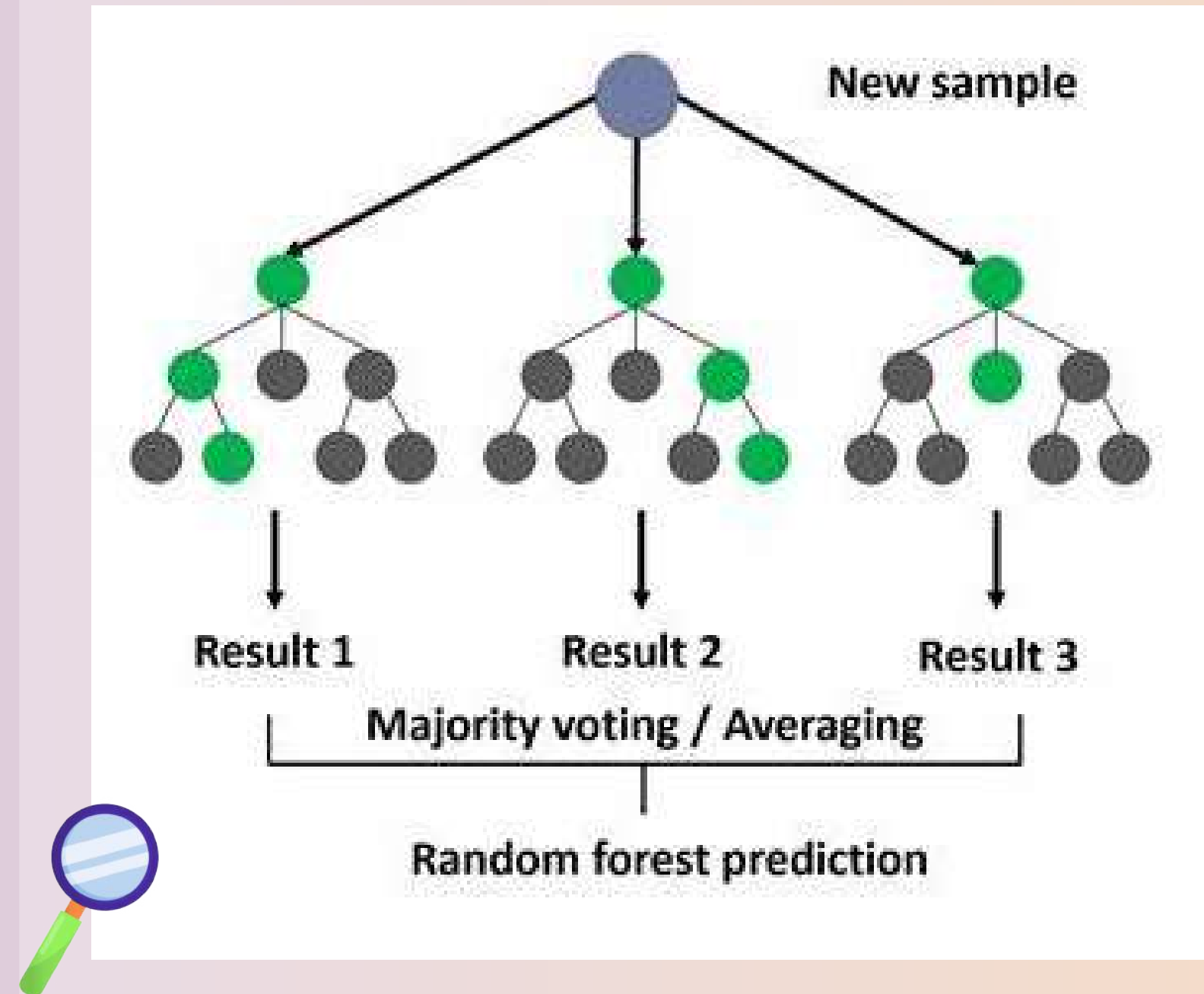
Tabel 1. Dataset

No	Molecule ChEMBL ID	Assay ChEMBL ID	Activity Type	Value	Units	Standard Type	Standard Value	Standard Units	SMILES
1	CHEMBL68920	CHEMBL674637	Active	0.041	uM	IC50	41	nM	<chem>Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c32)[nH]1</chem>
2	CHEMBL68920	CHEMBL621151	Intermediate	0.3	uM	IC50	300	nM	<chem>Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c32)[nH]1</chem>
3	CHEMBL68920	CHEMBL615325	Intermediate	7.82	uM	IC50	7820	nM	<chem>Cc1cc(C)c(/C=C2\C(=O)Nc3ncnc(Nc4ccc(F)c(Cl)c4)c32)[nH]1</chem>
.....
33.728	CHEMBL4755229	CHEMBL5304266	Intermediate	100	%	% Ctrl	100	%	<chem>CNC(=O)COc1cc2cc(Nc3nc(N4CC(C(C(=O)N(C)C)CC4)nc3Cl)cc(OC)c2n(C)c1=O</chem>
33.729	CHEMBL5306577	CHEMBL5304478	Active	9	%	Inhibition	9	%	<chem>CCC[C@H]1CNc2c1[nH]c(=O)c(C#N)c2N1C(C2(CC1)C2</chem>

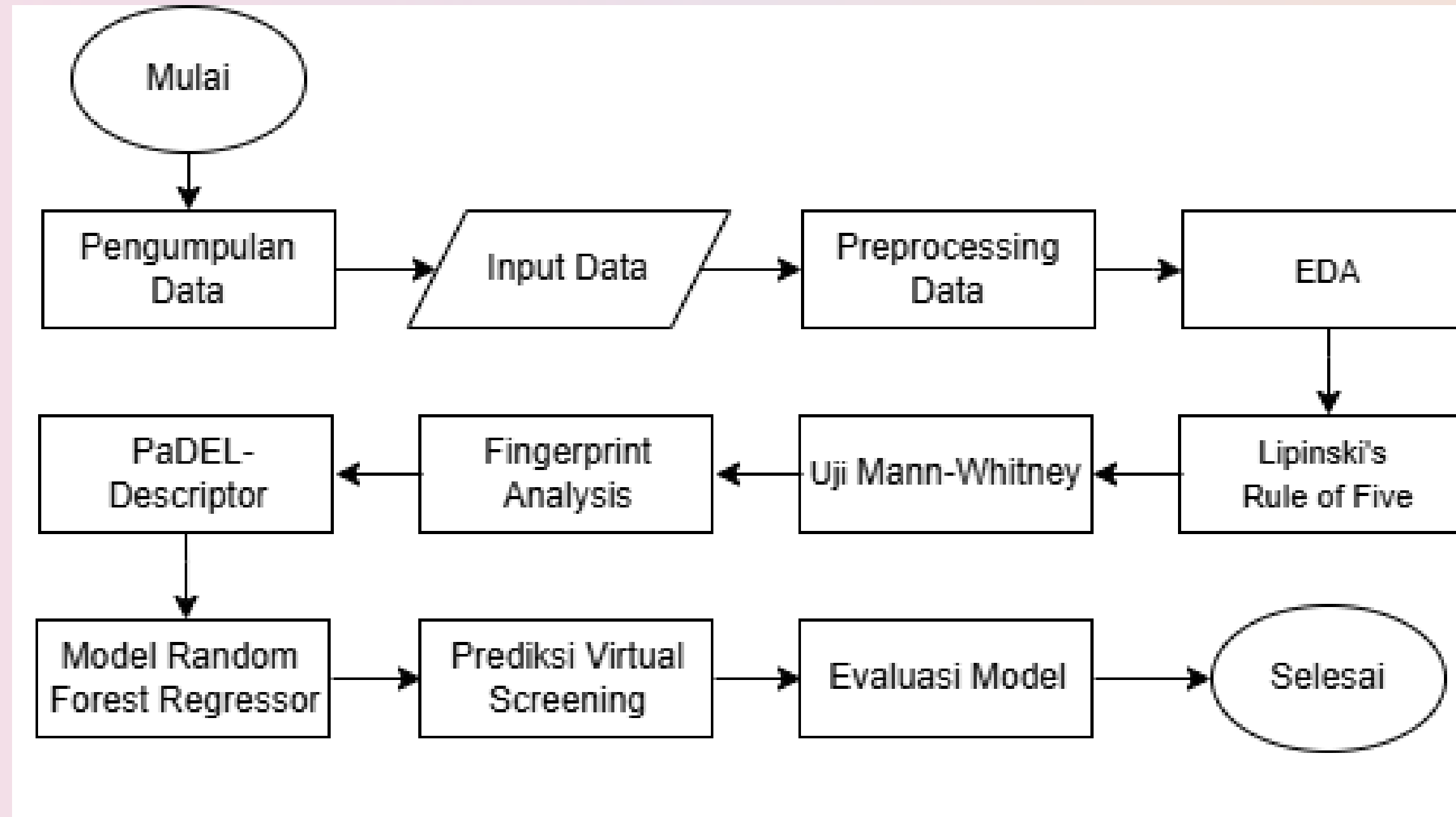
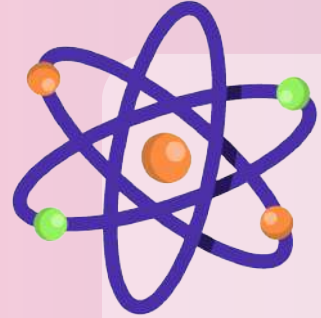
Random Forest Regressor



Random Forest Regressor adalah algoritma pembelajaran mesin berbasis supervised learning yang populer untuk analisis data. Algoritma ini memanfaatkan data berlabel untuk melatih model, menggabungkan prediksi dari beberapa Pohon Keputusan (Decision Trees) untuk menghasilkan hasil yang lebih akurat dan andal [6].

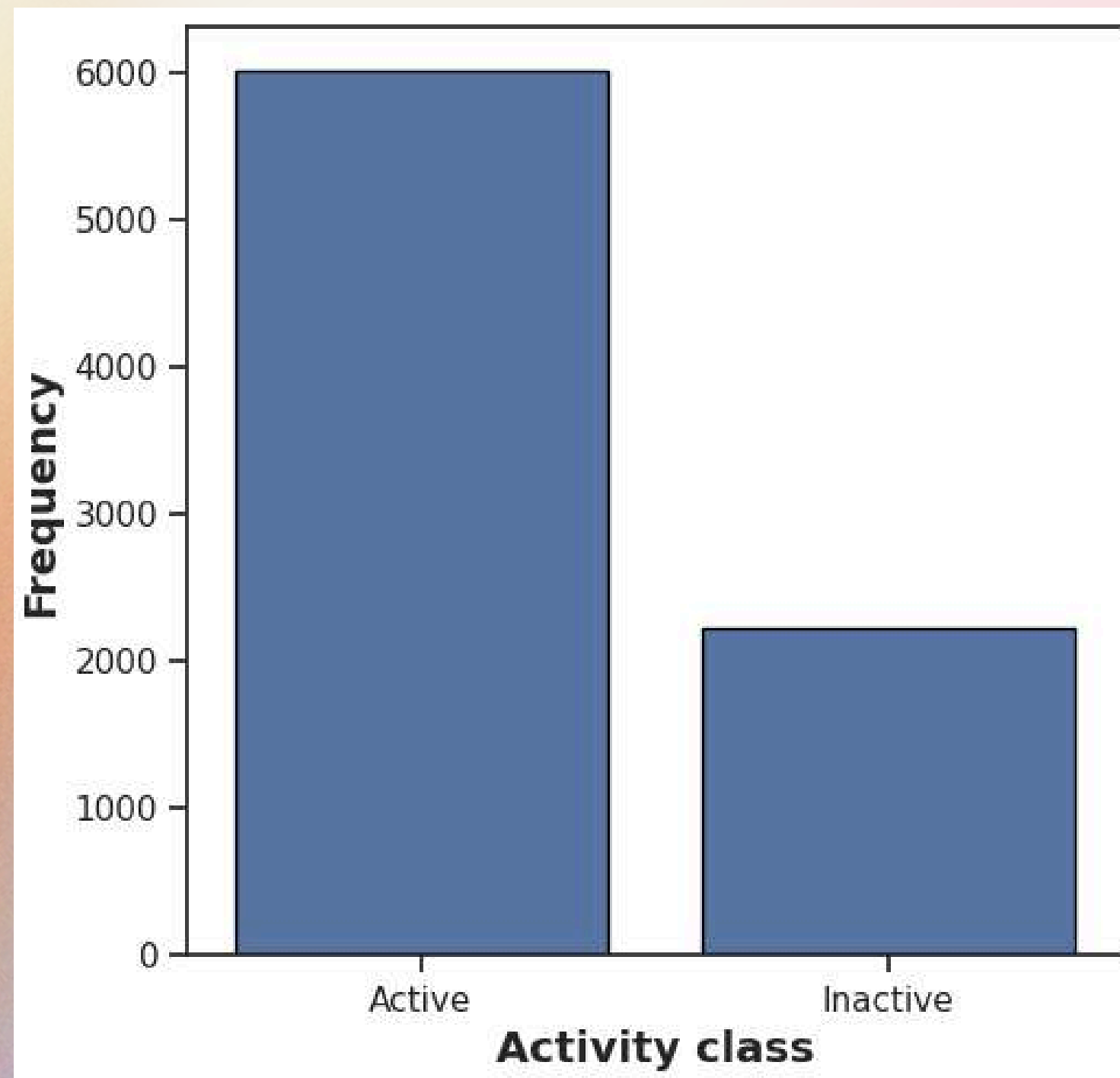


FLOWCHART



HASIL DAN PEMBAHASAN

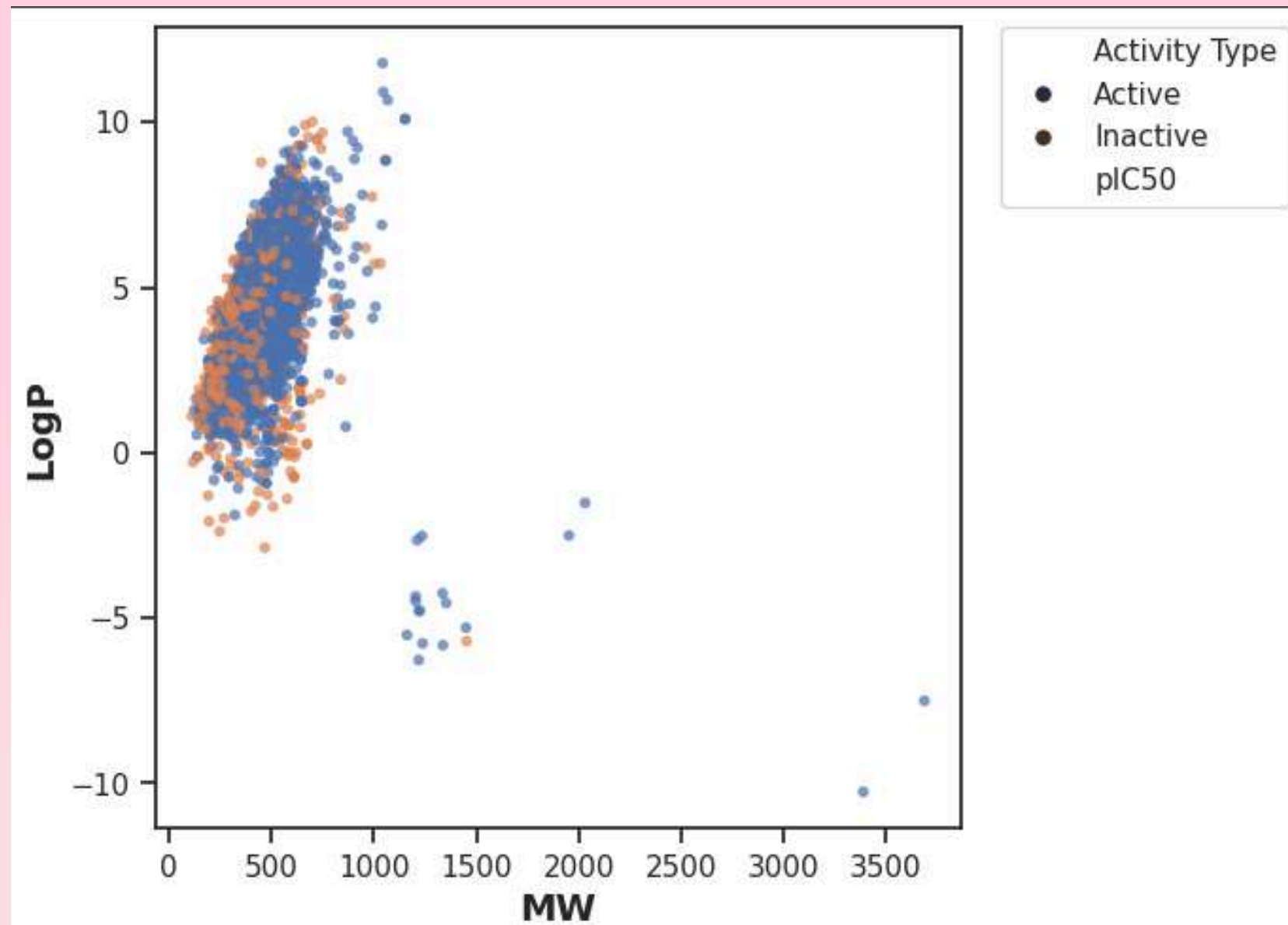
Distribusi Kelas Bioaktivitas



Dominasi signifikan senyawa aktif dengan rasio 3:1 dibandingkan senyawa inaktif, mencerminkan fokus penelitian pada optimasi senyawa dengan aktivitas inhibisi kuat terhadap EGFR. Bias ini, meskipun wajar dalam pengembangan obat, menimbulkan tantangan dalam pengembangan model prediktif yang andal.

HASIL DAN PEMBAHASAN

Persebaran dengan MW dan LogP



Senyawa aktif inhibitor EGFR cenderung memiliki berat molekul 300-500 Da dan LogP 0-5, sesuai dengan aturan Lipinski. Korelasi positif antara MW dan LogP menunjukkan "sweet spot" optimal untuk aktivitas biologis, dengan senyawa paling aktif memiliki keseimbangan MW dan LogP yang mendukung nilai pIC50 tinggi.

HASIL DAN PEMBAHASAN

Uji statistik Mann-Whitney antar kelas

Deskriptor	P-value	Interpretasi
MW	2.386505e-120	Distribusi berbeda (Tolak H0)
LogP	6.453325e-42	Distribusi berbeda (Tolak H0)
<u>NumHDonors</u>	4.655712e-09	Distribusi berbeda (Tolak H0)
NumHAcceptors	2.922315e-107	Distribusi berbeda (Tolak H0)
pIC50	0.0	Distribusi berbeda (Tolak H0)

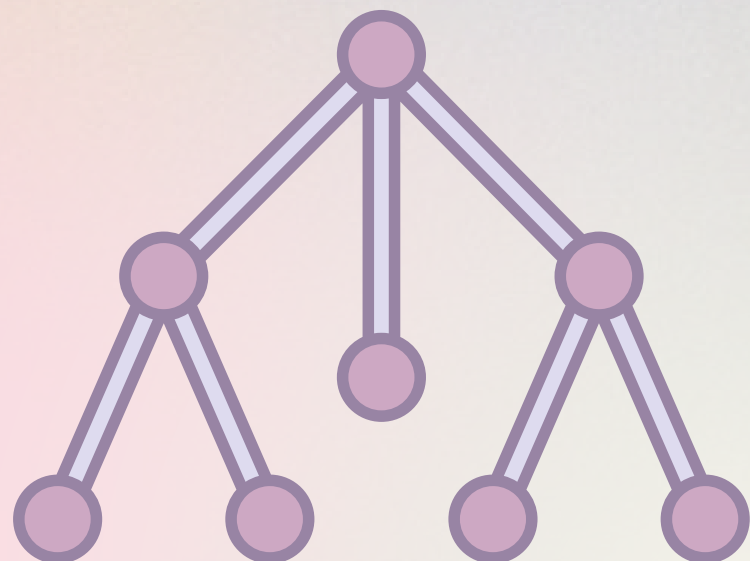
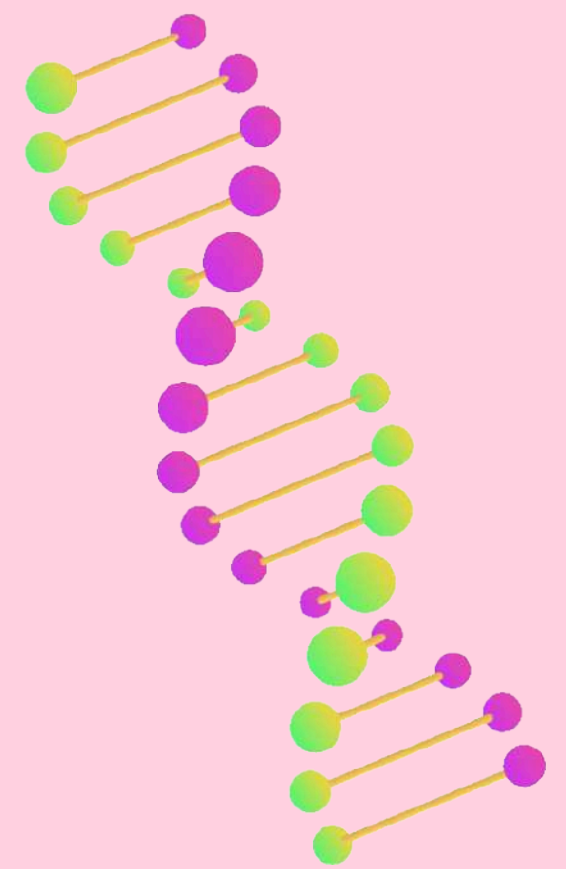
Hasil uji Mann-Whitney dengan p-value sangat kecil untuk seluruh deskriptor Lipinski. MW ($2.39\text{e-}120$) dan LogP ($6.45\text{e-}42$) menegaskan pentingnya ukuran molekul dan lipofilisitas, sementara donor ($4.66\text{e-}09$) dan akseptor ikatan hidrogen ($2.92\text{e-}107$) menunjukkan peran kapasitas pembentukan ikatan hidrogen. Temuan ini memberikan dasar kuantitatif untuk optimasi properti fisikokimia dalam pengembangan inhibitor EGFR.



HASIL DAN PEMBAHASAN

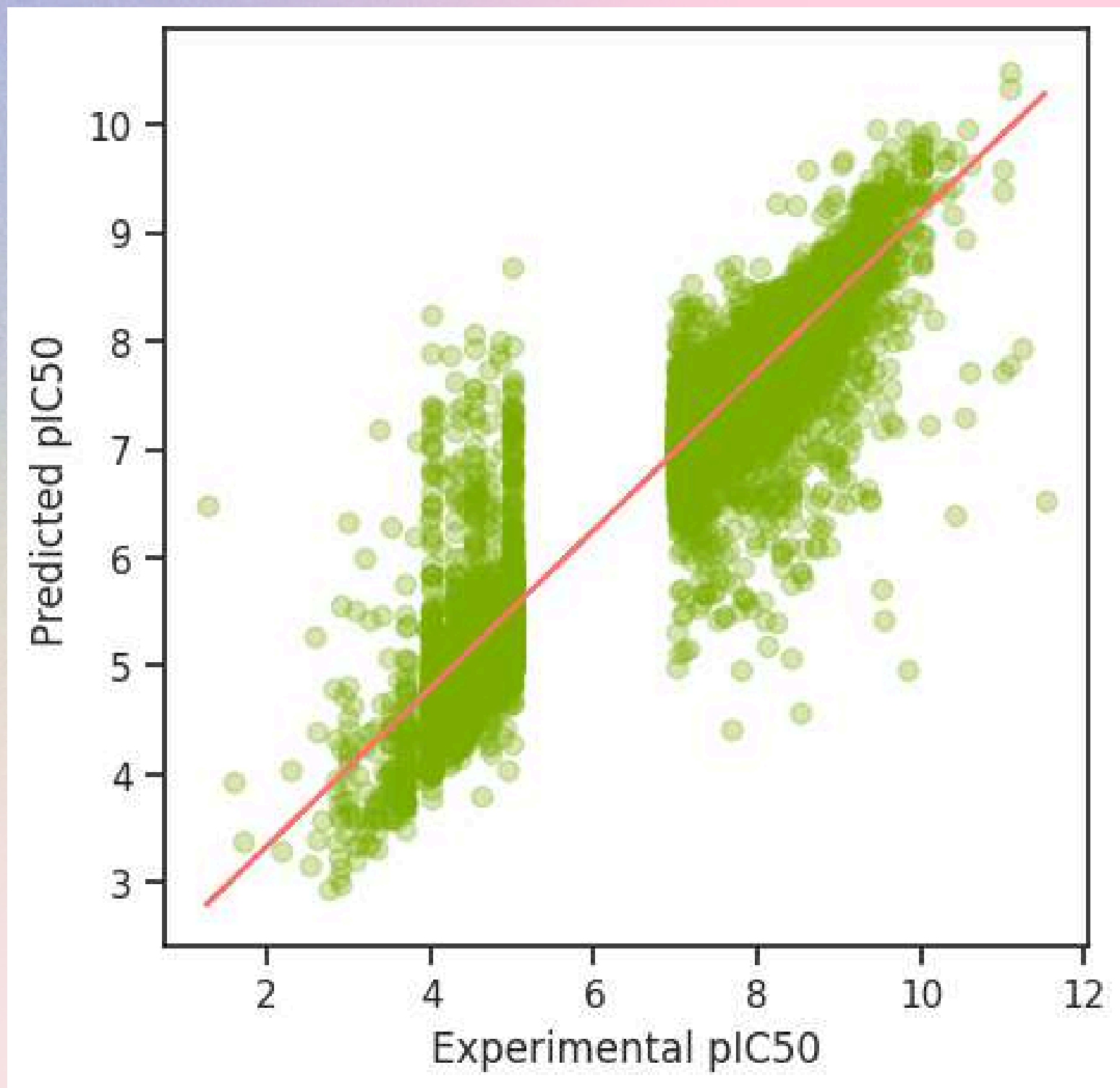
Performa Model Random Forest

- **MSE** = 1.41 (test set)
- **R²** = 0.50 (test set), 0.84 (keseluruhan data)
- Model menjelaskan 50% variabilitas data pada test set.
- Perlu optimasi untuk meningkatkan generalisasi pada data baru.



HASIL DAN PEMBAHASAN

Scatter Plot nilai pIC50


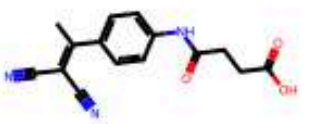
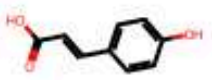


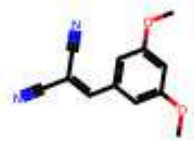
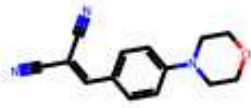
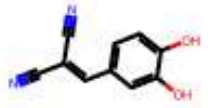
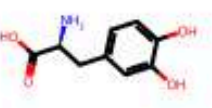

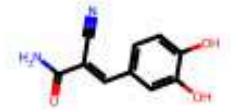

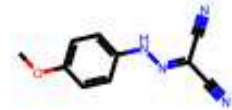

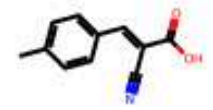
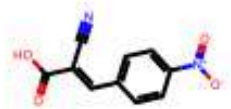
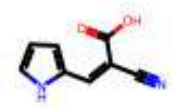
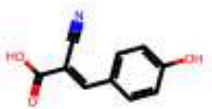
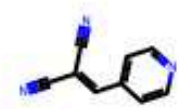
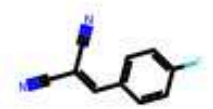


- Korelasi prediksi dan eksperimental cukup baik.
- Mayoritas data mendekati garis diagonal.
- Terdapat outlier pada nilai pIC50 ekstrem.
- Model memprediksi kluster aktivitas rendah (2-5) dan tinggi (6-10) dengan baik.
- Perlu optimasi dataset dan fitur molekuler untuk meningkatkan akurasi.

HASIL DAN PEMBAHASAN

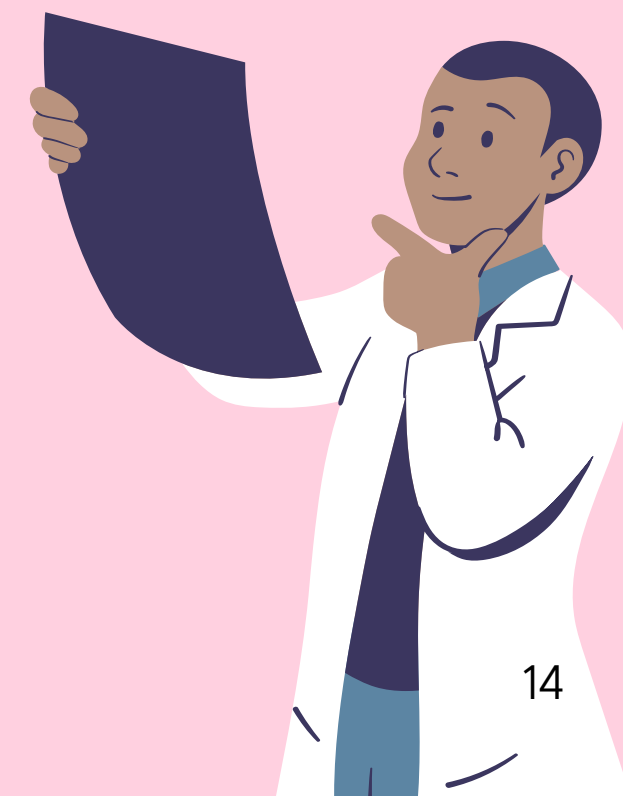
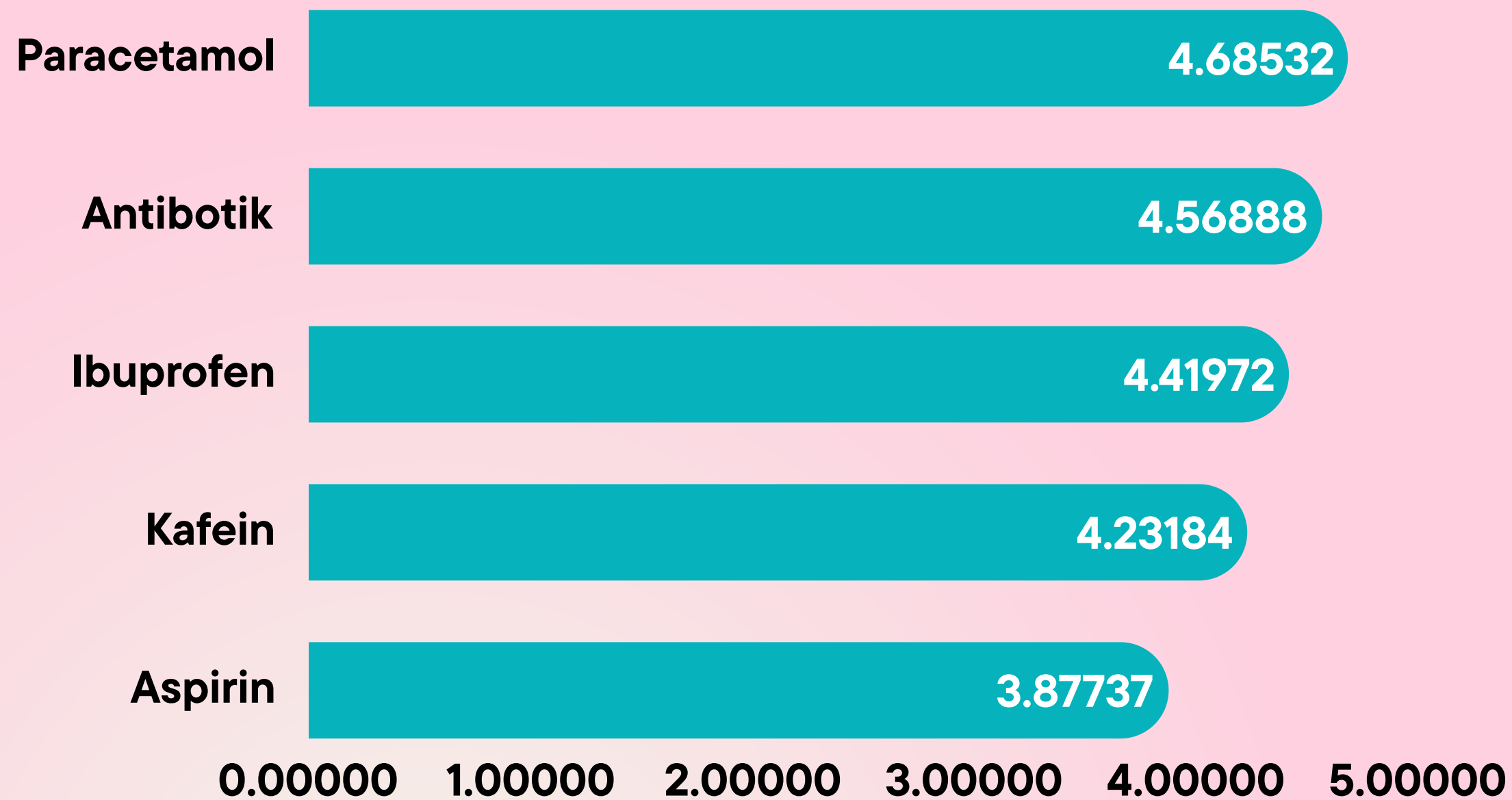
Hasil virtual
screening
senyawa



				
7.387216143280264	3.3010299956639813	2.5228787452803374	4.017728766960431	3.5783960731301687
				
2.903089986991944	3.204119982655925	4.455931955649724	3.0457574905606752	9.346787486224656
				
5.0	4.221848749616356	3.221848749616356	3.070581074285707	2.869666231504994
				
2.886056647693163	2.8664610916297826	3.7825160557860937	3.6020599913279625	2.903089986991944

HASIL DAN PEMBAHASAN

Prediksi pIC50 untuk Lima Senyawa Uji



KESIMPULAN

Penelitian ini berhasil mengembangkan model prediksi bioaktivitas senyawa terhadap EGFR menggunakan algoritma Random Forest Regressor dengan memanfaatkan dataset ChEMBL. Model yang dikembangkan menunjukkan performa prediksi yang cukup baik, dengan R^2 sebesar 0,50 pada data uji, meskipun masih terdapat ruang untuk optimasi, terutama pada senyawa ekstrem. Hasil virtual screening menunjukkan bahwa fitur molekul seperti cincin aromatik dan donor hidrogen berperan penting dalam meningkatkan nilai pIC₅₀, dengan paracetamol dan antibiotik menonjol sebagai kandidat potensial untuk pengembangan lebih lanjut. Studi ini membuktikan efisiensi pendekatan pembelajaran mesin dalam mempercepat proses penemuan obat, khususnya dalam desain inhibitor EGFR untuk terapi kanker.

THANK YOU