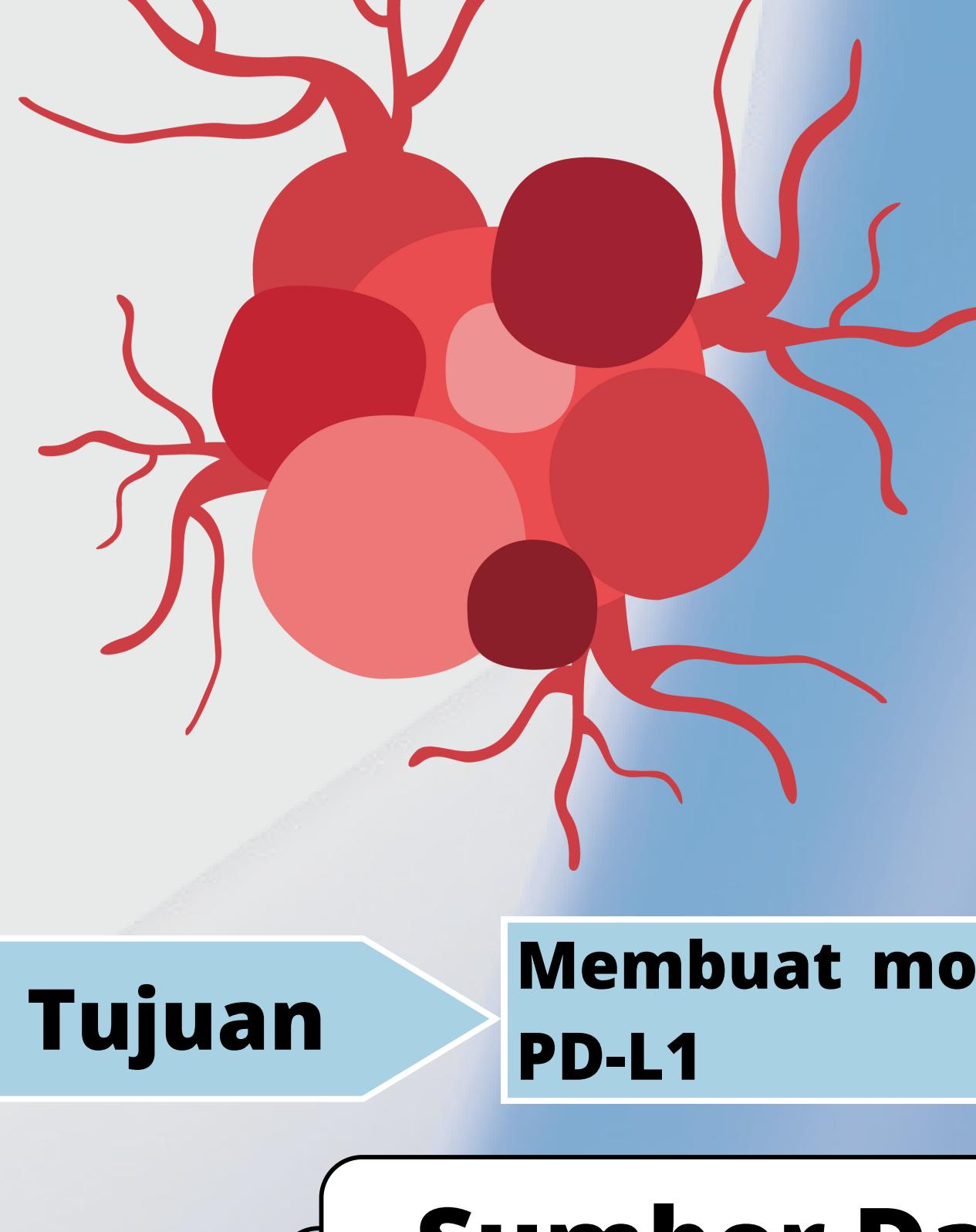


Klasifikasi Aktivitas Molekul Kecil terhadap PD-L1 Menggunakan Algoritma Random Forest



Programmed Death-Ligand 1 (PD-L1) merupakan salah satu protein penting yang terlibat dalam mekanisme penghindaran sistem imun oleh sel kanker.

Menurut data Kemenkes RI tahun 2022, angka kejadian penyakit kanker di Indonesia sebesar **136 orang per 100.000 penduduk** dan menempati urutan **ke-8** di Asia Tenggara.

Perempuan merupakan kelompok dengan risiko tinggi terkena kanker, tercatat kanker payudara sebanyak 65.858 kasus, **kanker Leher Rahim sebanyak 36.633 kasus**. Kanker pada laki-laki paling banyak **kanker paru 25.943 kasus**, dan **kanker kolorektal 21.764 kasus**.

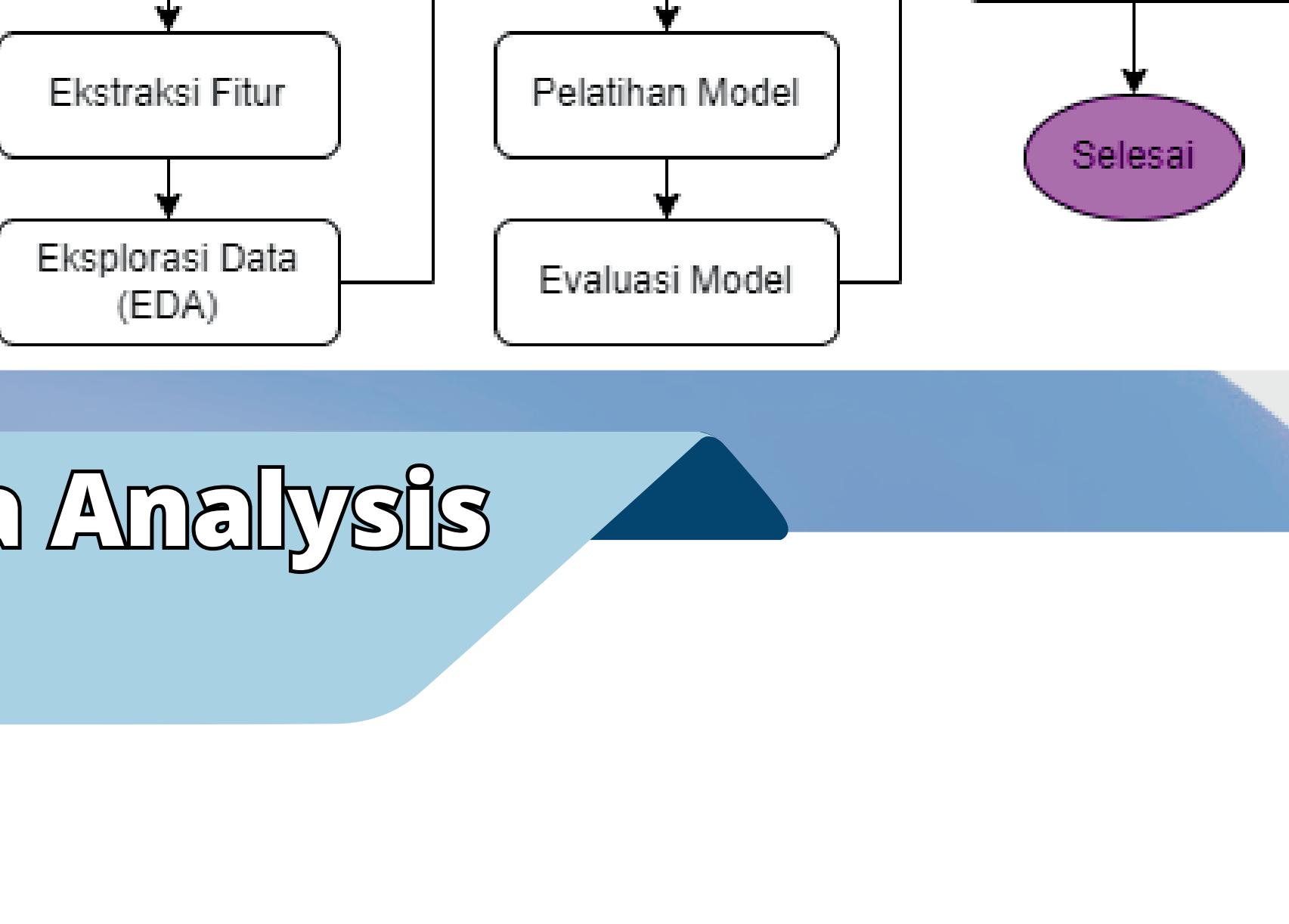
Tujuan

Membuat model Random Forest yang Dapat Mengklasifikasikan Aktivitas Molekul PD-L1

Sumber Data

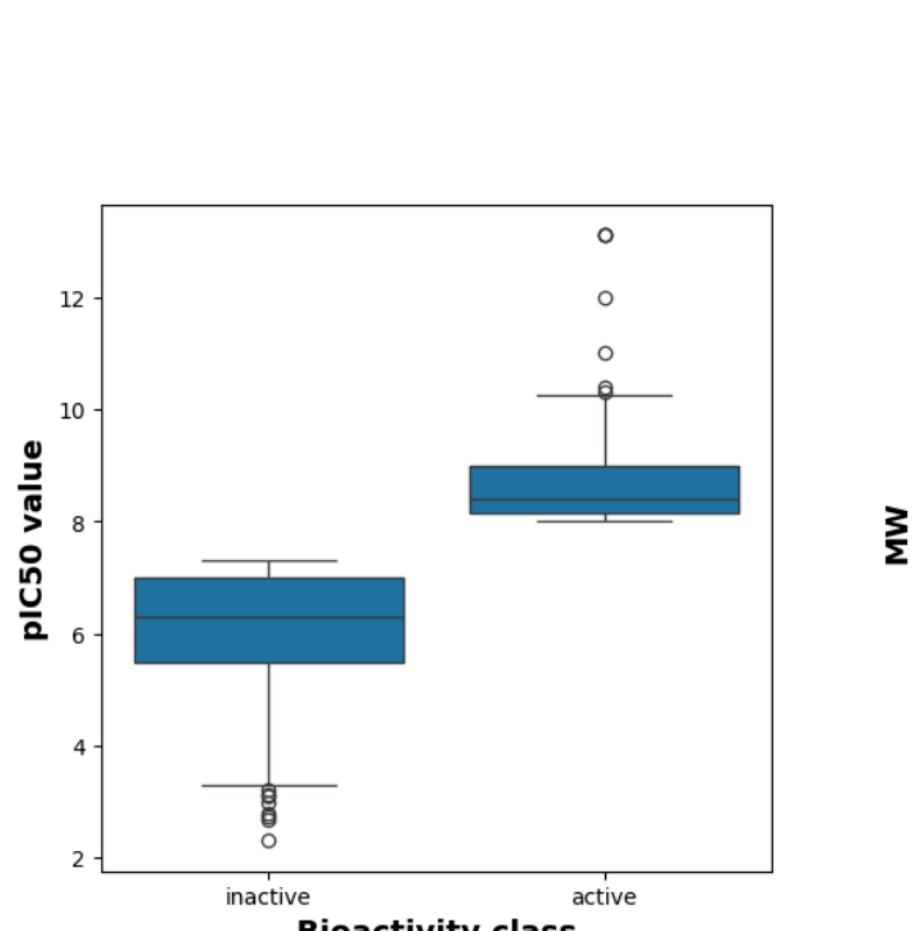
Dataset yang digunakan diunduh dari platform **ChEMBL**, sebuah basis data bioinformatika yang berisi informasi tentang aktivitas bioaktif molekul kecil terhadap target biologis. Dataset ini dipilih karena relevansinya dalam menganalisis aktivitas molekul kecil terhadap PD-L1 (Programmed Death-Ligand 1), yang merupakan target penting dalam penelitian imunoterapi kanker.

Metodelogi

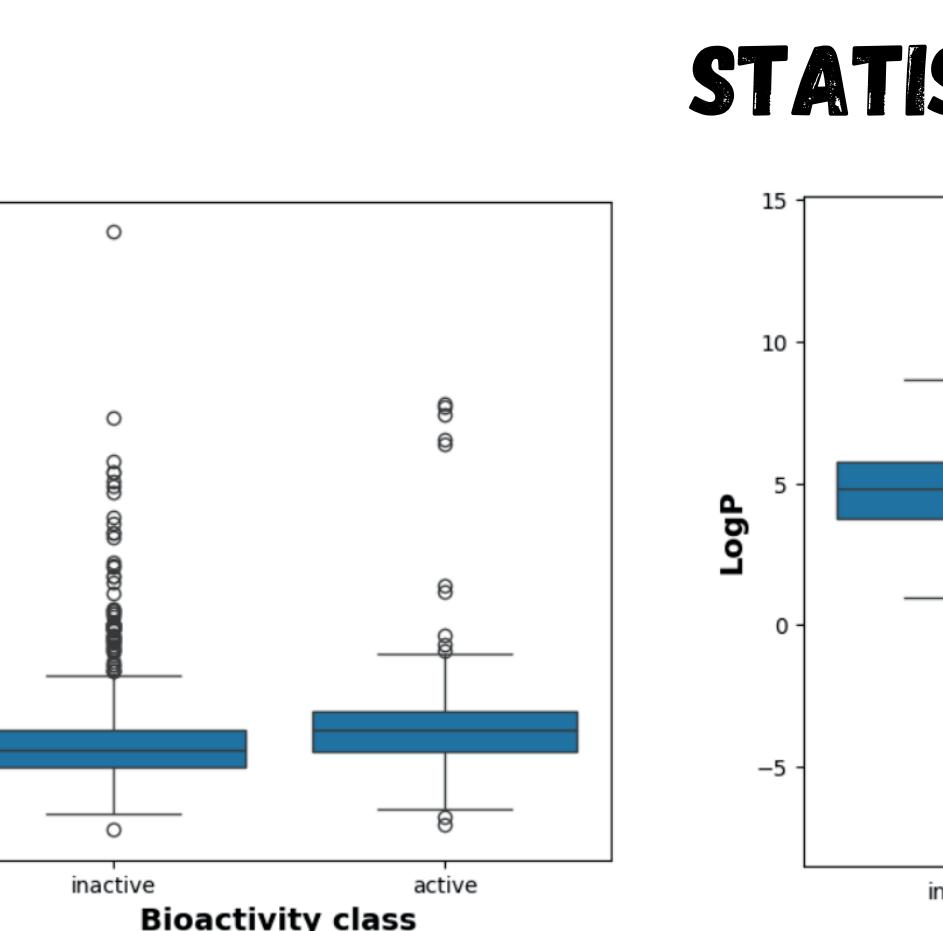


Exploratory Data Analysis (EDA)

FREQUENCY PLOT



SCATTER PLOT



- Terdapat **ketidakseimbangan** yang signifikan antara molekul aktif dan tidak aktif
- Pola **distribusi yang tumpang tindih** antara molekul aktif dan tidak aktif.

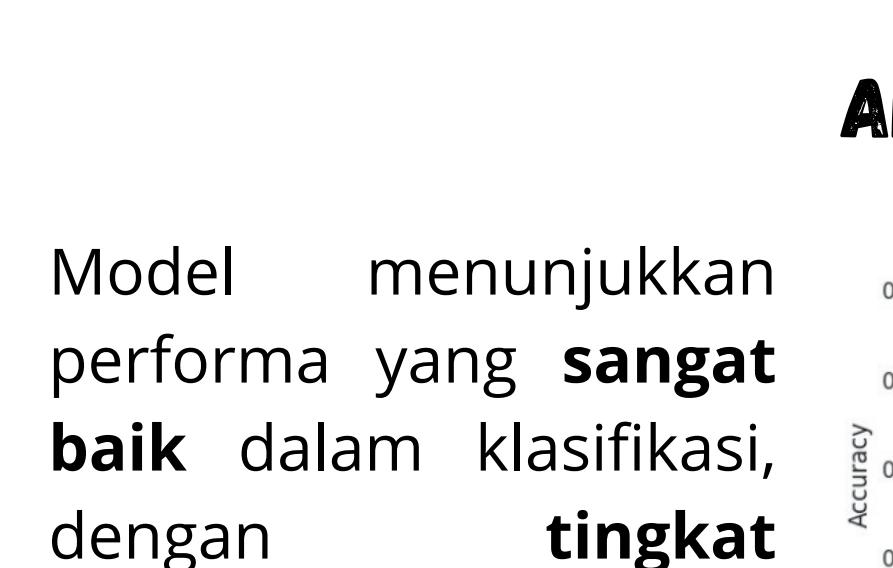
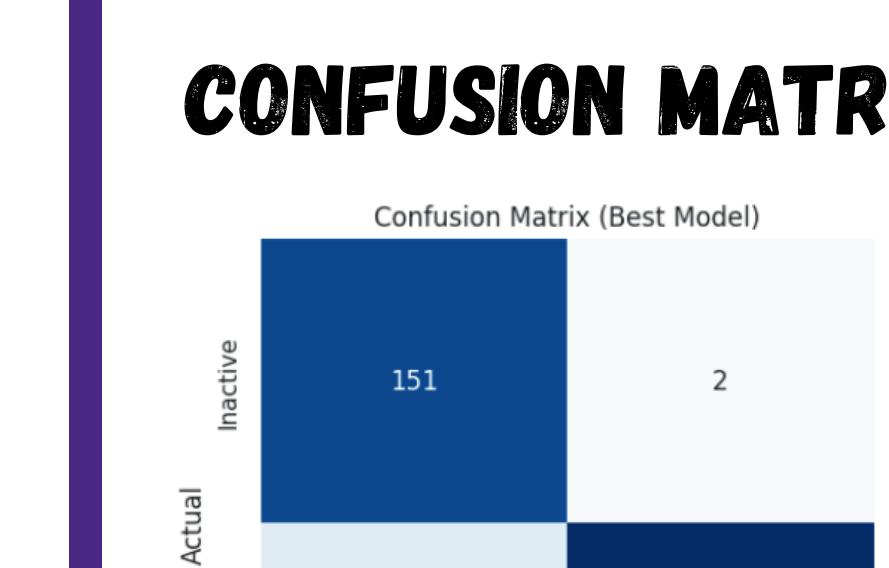
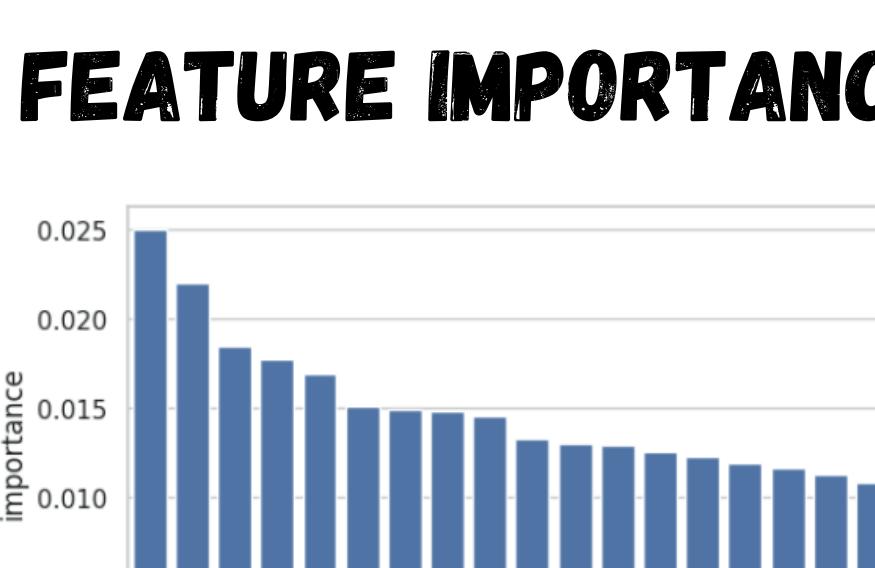
STATISTICAL ANALYSIS



- Nilai pIC50 untuk *bioactivity class*, **active lebih tinggi** menunjukkan bahwa senyawa aktif memiliki **potensi bioaktivitas yang lebih besar**.
- Distribusi berat molekul (MW) antara kelas active dan inactive **relatif mirip**, menunjukkan variabilitas yang lebih besar.
- Jumlah donor hidrogen (NumHDonors) **tidak memiliki perbedaan yang signifikan** antara kelas bioaktivitas aktif dan tidak aktif

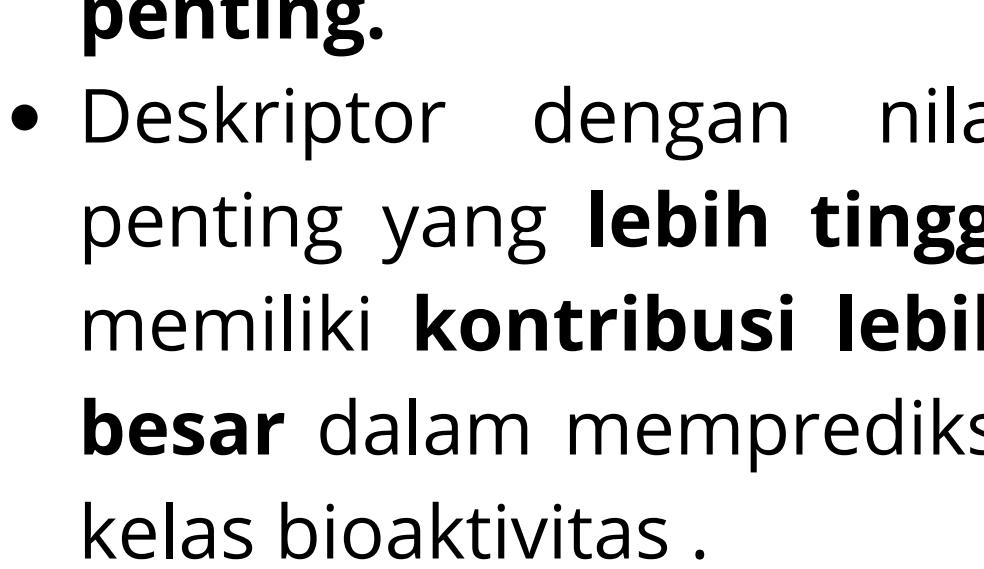
Hasil dan Pembahasan

LAZY PREDICT



- Model RandomForestRegressor, memiliki **kemampuan terbaik** dalam menjelaskan variasi data target.
- Model RandomForestRegressor, memiliki **tingkat kesalahan prediksi terendah**.
- Proses latih membutuhkan waktu mendekati **1.5 s**.

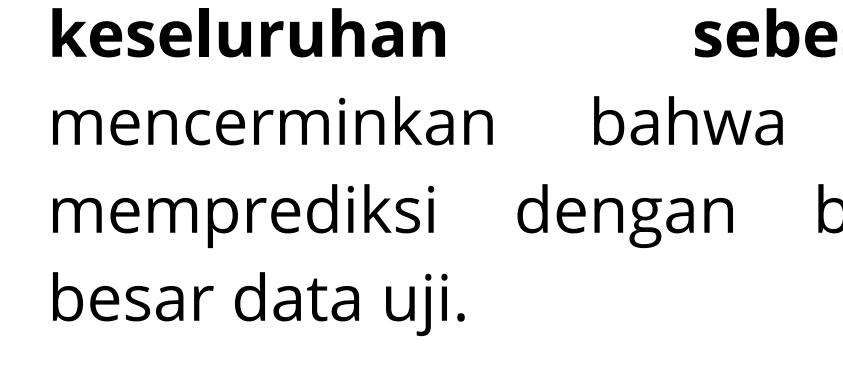
FEATURE IMPORTANCE



CONFUSION MATRIX

Model menunjukkan performa yang **sangat baik** dalam klasifikasi, dengan **tingkat kesalahan yang rendah**.

AKURASI TRAIN VS TEST



Iterasi optimal berada di sekitar iterasi 20-25, di mana perbedaan antara train set dan test set tidak besar

CLASSIFICATION REPORT

Classification Report (Best Model):

	precision	recall	f1-score	support
0	0.88	0.99	0.93	153
1	0.99	0.89	0.94	186

	accuracy	macro avg	weighted avg	
0	0.94	0.94	0.94	339

	precision	recall	f1-score	support
1	0.94	0.94	0.94	339

	accuracy	macro avg	weighted avg	
0	0.94	0.94	0.94	339

Model menunjukkan performa yang **sangat baik** dengan akurasi keseluruhan sebesar **94%**,

mencerminkan bahwa model dapat memprediksi dengan