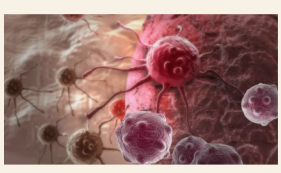
PREDIKSI SENYAWA KANDIDAT OBAT UNTUK TARGET KRAS (KIRSTEN RAT SARCOMA VIRAL ONCOGENE) MENGGUNAKAN PENDEKATAN ALGORITMA K-NEAREST NEIGHBOR (KNN)

A Rafi Paringgom Iwari, Silvia Azahrani , Jelli Kurnilia , Hermalina Sintia Putri , Ayu Erlinawati, Ditta Winanda Putri
Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung , Indonesia
Kelompok 1 RB

PENDAHULUAN



Kanker merupakan salah satu tantangan kesehatan global terbesar, di mana mutasi genetik, seperti pada **onkogen KRAS (*Kirsten Rat Sarcoma Viral Oncogene*)**, berperan signifikan dalam perkembangan penyakit dan resistensi terhadap terapi.

Mutasi KRAS ditemukan pada sekitar **25% kasus kanker manusia**, termasuk kanker paru-paru, kolorektal, dan pankreas, yang sering dikaitkan dengan prognosis buruk. Protein KRAS yang bermutasi menyebabkan jalur pensinyalan sel terus aktif, mendorong pertumbuhan sel kanker yang tidak terkendali. Namun, **sifat KRAS yang "undruggable"** telah menjadi tantangan utama dalam pengembangan terapi.

TUJUAN

Penelitian ini bertujuan untuk **memprediksi senyawa kandidat obat yang efektif dalam menargetkan mutasi KRAS menggunakan algoritma *K-Nearest Neighbor* (KNN)**. Pendekatan ini diharapkan dapat mengidentifikasi senyawa potensial untuk pengembangan terapi kanker yang terkait dengan mutasi KRAS.

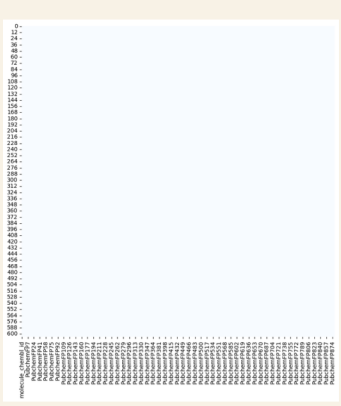
DATASET

- Sumber data : **ChEMBL**

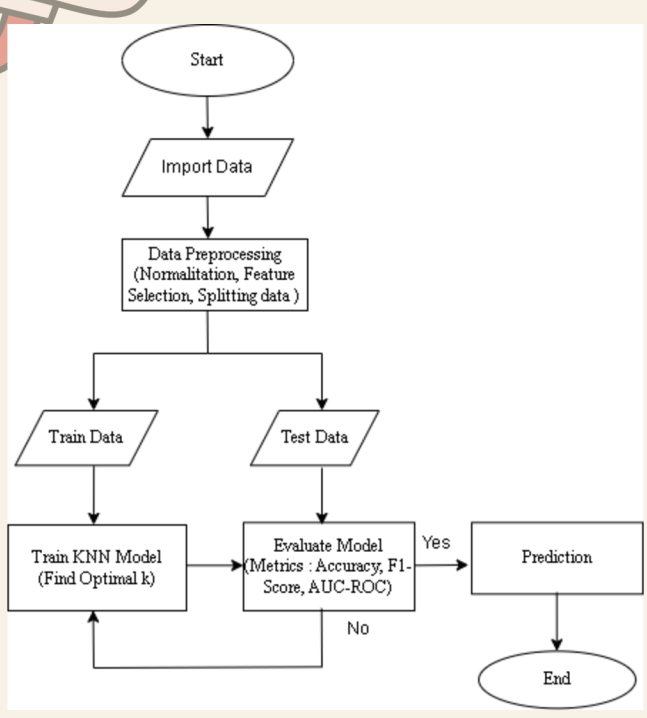
Dataset Pertama: **Lipinsky (Label Class KRAS)**
Dataset ini **berisi informasi tentang senyawa kimia dan sifat-sifatnya** yang relevan untuk analisis aktivitas biologis terhadap target KRAS yang terdiri dari **10 kolom dan 606 baris.**

Dataset Kedua : **Fingerprint Senyawa**
Dataset kedua **berisi representasi fingerprint senyawa** yang diambil dari database PubChem, yang terdiri dari **881 kolom dan 606 baris.**

Gabungan Data :



METODOLOGI



HASIL DAN PEMBAHASAN

- **Pengumpulan Data**

Dua dataset digabung: sifat kimiawi (Lipinski) dan fingerprint molekul (PubChem) untuk analisis komprehensif.

- **Preprocessing Data**

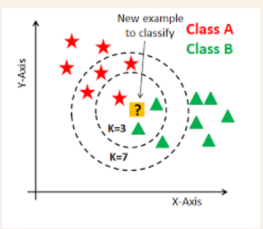
Data diklasifikasikan, dinormalisasi, dan diperbaiki dari missing values, **menghasilkan 886 fitur**.

- **Split Data**

Dataset dibagi **80% latih** dan **20% uji** secara stratifikasi untuk menjaga keseimbangan kelas.

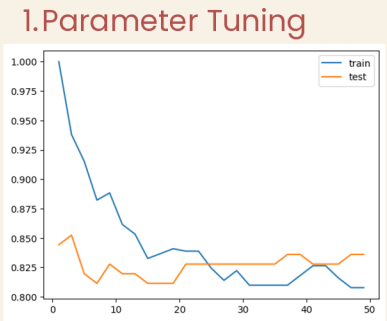
- **Klasifikasi K-Nearest Neighbor (KNN)**

Model KNN digunakan untuk memprediksi aktivitas senyawa terhadap KRAS berdasarkan kemiripan karakteristik. Nilai optimal k = 3 diperoleh melalui validasi silang, dan klasifikasi dilakukan berdasarkan mayoritas kelas dari 3 tetangga terdekat.



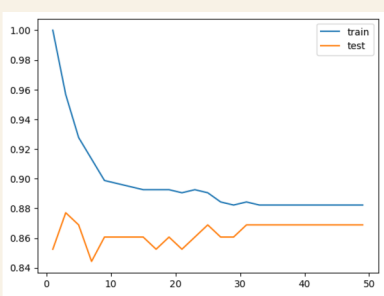
- **Improvisasi Model**

1. Parameter Tuning



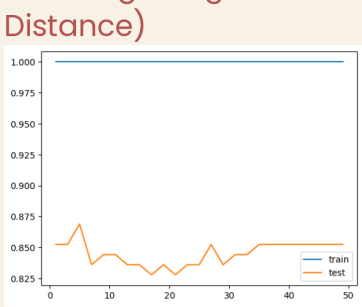
Tuning KNN dilakukan dengan menguji **nilai k dari 1 hingga 51**. Nilai optimal **k = 3** memberikan akurasi terbaik sebesar **85.25%.**

2. Tuning dengan Scaling



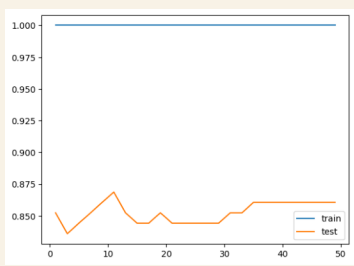
Normalisasi meningkatkan akurasi KNN menjadi **87.70%** dengan **k = 3**, memaksimalkan perhitungan jarak data.

3. Tuning dengan Pembobotan Jarak (Weighted Distance)



Dengan weights = distance dan **k = 5**, KNN mencapai **akurasi 86.88%**, unggul pada data dengan distribusi tidak seragam atau outlier.

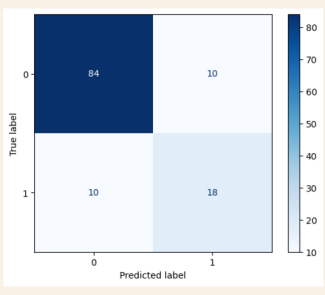
4. Menggunakan Manhattan Distance



Dengan **p = 1** (Manhattan Distance) dan **k = 11**, model mencapai **akurasi 86.88%**, lebih robust terhadap outlier dan stabil untuk dataset berdimensi tinggi.

5. Klasifikasi KNN Dengan SMOTE

6. Evaluasi Model



Model mencapai **akurasi 84%** pada data uji, dengan **precision, recall, dan F1-score** masing-masing sebesar **84%**.

Kelas	Precision	Recall	Fl-score	Support
0	0.89	0.89	0.89	94
1	0.64	0.64	0.64	28
Accuracy			0.84	122
Micro avg	0.77	0.77	0.77	122
Weighted avg	0.84	0.84	0.84	122

Model memiliki akurasi **84%**, dengan performa lebih baik pada **kelas 0** (precision, recall, dan F1-score **0.89**). **Kelas 1** memiliki precision, recall, dan F1-score **0.64**.