Prediksi Senyawa Kandidat Obat untuk Target KRAS (*Kirsten Rat Sarcoma Viral Oncogene*) Menggunakan Pendekatan Algoritma *K-Nearest Neighbor* (KNN)

Prediction of Drug Candidate Compounds for KRAS (Kirsten Rat Sarcoma Viral Oncogene) Targets Using the K-Nearest Neighbor (KNN) Algorithm Approach

A Rafi Paringgom Iwari*[*1], Silvia Azahrani [2], Jelli Kurnilia[3], Hermalina Sintia Putri[4], Ayu Erlinawati[5], Ditta Winanda Putri[6]

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

**E-mail: arafi.121450039@student.itera.ac.id*

Abstrak

Penelitian ini bertujuan untuk mengembangkan model prediksi berbasis *K-Nearest Neighbors* (KNN) untuk mengidentifikasi senyawa kandidat obat yang potensial menargetkan KRAS (Kirsten Rat Sarcoma Viral Oncogene). Data penelitian yang digunakan adalah dataset yang memuat informasi senyawa kimia dan sifat bioaktifnya. Proses penelitian meliputi *preprocessing* data, seperti normalisasi dan pembagian dataset menjadi 80% data pelatihan dan 20% data pengujian. Untuk mengatasi ketidakseimbangan kelas dalam dataset, diterapkan teknik oversampling menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Hasil penelitian menunjukkan bahwa model terbaik mencapai akurasi sebesar 84% pada data uji, dengan dengan *precision*, *recall*, dan *F1-score* masing-masing juga sebesar 84%, mencerminkan kinerja prediksi yang baik. Pendekatan ini memberikan kontribusi dalam proses penemuan obat, dengan menawarkan metode berbasis *machine learning* yang efisien untuk mengevaluasi potensi senyawa obat dalam menargetkan KRAS.

Kata kunci: *K-Nearest Neighbors* , KRAS, SMOTE, Senyawa Obat, *Machine Learning*

Abstract

This study aims to develop a prediction model based on K-Nearest Neighbors (KNN) to identify potential drug candidates targeting KRAS (Kirsten Rat Sarcoma Viral Oncogene). The research uses a dataset containing information on chemical compounds and their bioactive properties. The research process includes data preprocessing, such as normalization and splitting the dataset into 80% training data and 20% testing data. To address class imbalance in the dataset, Synthetic Minority Oversampling Technique (SMOTE) is applied. The results show that the best model achieves an accuracy of 84% on the test data, with precision, recall, and F1-score all also at 84%, reflecting good prediction performance. This approach contributes to drug discovery by offering an efficient machine learning method to evaluate the potential of drug candidates in targeting KRAS.

Keywords:* *K-Nearest Neighbors* , KRAS, SMOTE, *Drug Compounds, Machine Learning

PENDAHULUAN

Kanker merupakan salah satu tantangan kesehatan global yang signifikan, di mana mutasi genetik memainkan peran penting dalam perkembangan penyakit dan resistensi terhadap terapi. Salah satu onkogen yang paling sering bermutasi adalah KRAS (Kirsten Rat Sarcoma Viral Oncogene). Mutasi KRAS ditemukan pada sekitar 25% kasus kanker manusia, termasuk kanker paru-paru, kolorektal, dan pankreas [1]. Mutasi ini sering dikaitkan dengan prognosis yang buruk dan resistensi terhadap berbagai terapi yang ada.

Gen KRAS mengkode protein yang berperan dalam jalur pensinyalan sel untuk mengatur pertumbuhan dan pembelahan sel. Mutasi pada gen ini menyebabkan aktivasi jalur pensinyalan yang terus-menerus, sehingga mendorong pertumbuhan sel kanker yang tidak terkendali. Contoh, dalam kanker kolorektal, mutasi KRAS sering ditemukan pada tahap awal penyakit dan dapat mempengaruhi respons terhadap terapi target, seperti antibodi monoklonal anti-EGFR seperti cetuximab dan panitumumab [2].

Pendekatan tradisional dalam penemuan obat untuk menargetkan KRAS telah menghadapi berbagai tantangan, terutama karena sifat protein KRAS yang dianggap "*undruggable*", sulit dijadikan target obat [3]. Namun, kemajuan dalam pembelajaran mesin dan kecerdasan buatan telah membuka peluang baru dalam penemuan senyawa obat yang efektif. Algoritma seperti *K-Nearest Neighbor* (KNN) telah digunakan untuk memprediksi respon tumor terhadap obat berdasarkan gen, menunjukkan potensi dalam mengidentifikasi senyawa yang mungkin efektif melawan mutasi KRAS [4].

Studi terbaru telah menerapkan model pembelajaran mesin untuk mengidentifikasi inhibitor baru untuk mutasi spesifik KRAS, seperti G12D. Dengan menggunakan basis data seperti ZINC, peneliti dapat menyaring senyawa yang memenuhi aturan Lipinski dan menunjukkan afinitas tinggi terhadap target KRAS [1]. Pendekatan ini memungkinkan identifikasi senyawa kandidat obat yang lebih efisien dibandingkan metode tradisional.

Selain itu, kombinasi pembelajaran mesin dengan metode lain, seperti docking molekuler dan simulasi dinamika molekuler, telah digunakan untuk memprediksi interaksi antara senyawa dan protein target. Pendekatan ini meningkatkan akurasi dalam memprediksi efektivitas senyawa kandidat obat [5]. Maka, integrasi berbagai teknik komputasional dapat mempercepat proses penemuan obat dan meningkatkan peluang keberhasilan dalam menargetkan onkogen seperti KRAS. Dengan memanfaatkan data genetik dan ekspresi gen, serta kemajuan dalam pembelajaran mesin, pendekatan ini dapat mengidentifikasi senyawa potensial yang dapat dikembangkan lebih lanjut menjadi terapi efektif untuk kanker yang terkait dengan mutasi KRAS.

METODE

Dataset

Penelitian ini menggunakan dua dataset yang berbeda untuk analisis dan prediksi senyawa kandidat obat untuk target KRAS.

a) Dataset Pertama: 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.

Tabel 1. Dataset Pertama

molecule_chembl_id	standard_value	class,canonical_smiles	…	pIC50
CHEMBL2396992	155000.0	inactive	….	3.8096683018297086
…	…	…	…	…
CHEMBL5274815	14.2	active	…	7.847711655616943

Berikut adalah penjelasan mengenai masing-masing kolom dalam dataset tersebut:

- **molecule_chembl_id**: ID unik untuk setiap senyawa yang diambil dari database ChEMBL, berfungsi sebagai pengenal untuk mengakses informasi lebih lanjut tentang senyawa tersebut.
- **standard_value**: Nilai standar yang menunjukkan aktivitas senyawa, biasanya dalam bentuk IC50 (konsentrasi inhibisi setengah). Nilai ini menunjukkan seberapa efektif senyawa tersebut dalam menghambat target KRAS dan dapat mengkategorikannya sebagai "*active*" (aktif) atau "*inactive*" (tidak aktif).
- **class**: Kategori senyawa, yang menunjukkan apakah senyawa tersebut aktif atau tidak berdasarkan hasil pengujian terhadap KRAS. Informasi ini penting untuk pelatihan model klasifikasi.
- **standard_value_norm**: Nilai standar yang dinormalisasi, memungkinkan analisis lebih lanjut dan perbandingan aktivitas di antara senyawa.
- **canonical_smiles :** Representasi teks dari struktur kimia molekul dalam format SMILES (*Simplified Molecular Input Line Entry System*) yang standar dan unik.
- **MW (Molecular Weight) :** Berat molekul senyawa
- **LogP** : Logaritma dari koefisien partisi (*partition coefficient*) oktanol-air, yang menunjukkan tingkat kepolaran atau lipofilisitas molekul.
- **NumHDonors (Number of Hydrogen Donors)** : Jumlah atom hidrogen dalam molekul yang dapat membentuk ikatan hidrogen.
- **NumHAcceptors** (*Number of Hydrogen Acceptors*) : Jumlah atom dalam molekul yang dapat menerima ikatan hidrogen
- standard_value_norm : Nilai **standard_value** yang telah dinormalisasi
- **pIC50**: Logaritma dari nilai IC50, memberikan skala yang lebih mudah untuk membandingkan potensi senyawa; nilai yang lebih tinggi menunjukkan potensi penghambatan yang lebih besar terhadap target.

b) Dataset Kedua : Fingerprint Senyawa

Dataset kedua berisi representasi fingerprint senyawa yang diambil dari database PubChem, yang terdiri dari 881 kolom dan 606 baris.

Tabel 2. Dataset Kedua

PubchemFP0	**PubchemFP1**	**PubchemFP2**	**…**	**PubchemFP880**
1	1	0	…	0
…	…	…		….
1	1	1	…	0

Berikut adalah penjelasan mengenai kolom-kolom dalam dataset ini:

PubchemFP0 hingga PubchemFP880: Setiap kolom merepresentasikan fitur fingerprint biner dari senyawa. Fingerprint ini adalah representasi struktural yang menggambarkan keberadaan atau ketidakhadiran substruktur tertentu dalam molekul. Dengan menggunakan fingerprint, setiap senyawa dapat direpresentasikan dalam bentuk vektor yang memungkinkan analisis kesamaan antara senyawa.

Kedua dataset ini digunakan secara sinergis dalam penelitian ini. Dataset pertama memberikan label yang diperlukan untuk melatih model klasifikasi, sedangkan dataset kedua menyediakan informasi struktural yang esensial untuk analisis karakteristik senyawa. Kombinasi ini memungkinkan penerapan algoritma KNN untuk memprediksi senyawa kandidat obat yang berpotensi menjadi inhibitor KRAS.

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

Metode *K-Nearest Neighbors* (KNN) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Dalam penelitian ini, KNN

diterapkan untuk memprediksi senyawa kandidat obat yang dapat menghambat target KRAS. KNN bekerja dengan prinsip dasar bahwa senyawa yang memiliki karakteristik serupa cenderung memiliki label kelas yang sama. Algoritma ini tidak memerlukan asumsi distribusi yang kuat, sehingga sangat fleksibel dan dapat digunakan pada berbagai jenis data, termasuk data kompleks [6].

Dalam penelitian ini, metode K-Nearest Neighbors (KNN) diterapkan untuk memprediksi senyawa kandidat obat yang memiliki potensi sebagai inhibitor terhadap target KRAS [7]. Proses dimulai dengan menggabungkan dua dataset: satu dataset yang berisi label kelas KRAS (aktif atau tidak aktif) dan dataset kedua yang berisi fingerprint senyawa dari PubChem. Setelah penggabungan, data diproses melalui normalisasi untuk memastikan bahwa semua fitur memiliki skala yang sama, sehingga perhitungan jarak antar data dapat dilakukan secara akurat. Selanjutnya, dataset dibagi menjadi dua bagian: data pelatihan dan data pengujian, untuk mencegah *overfitting*.

KNN diimplementasikan dengan memilih nilai k yang optimal melalui validasi silang. Algoritma ini menghitung jarak antara senyawa yang belum diklasifikasikan dengan semua senyawa dalam data pelatihan, menggunakan metrik jarak *Euclidean*. KNN kemudian memilih k tetangga terdekat dan mengklasifikasikan senyawa baru berdasarkan mayoritas kelas dari tetangga tersebut. Kinerja model dievaluasi menggunakan metrik seperti akurasi, presisi, dan recall, yang memberikan wawasan tentang efektivitas model dalam mengidentifikasi senyawa aktif dan tidak aktif. Dengan pendekatan ini, KNN menjadi alat yang efektif untuk analisis hubungan antara struktur kimia dan aktivitas biologis, serta mendukung pengembangan senyawa obat baru [8].

Manhattan Distance

Manhattan distance adalah metode perhitungan jarak antara dua objek dalam ruang koordinat dengan menggunakan konsep perbedaan absolut (nilai mutlak) [10]. Nama jarak ini berasal dari tata letak jalan di Pulau Manhattan yang memiliki bentuk segi empat. Rumus Jarak Manhattan ditunjukkan pada Persamaan berikut :

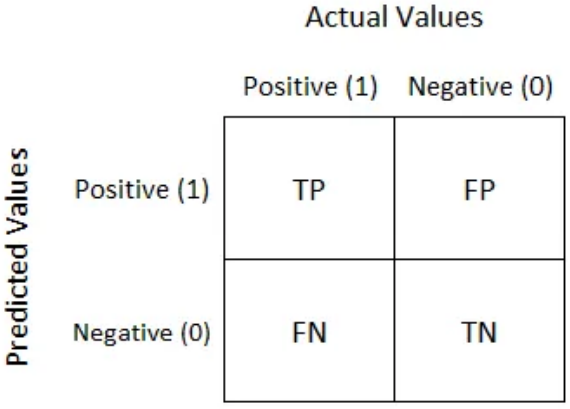
$$d(x, y) = \sum | x - y | \quad \textbf{(1)}$$

Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) adalah teknik untuk mengatasi ketidakseimbangan kelas dalam dataset, terutama ketika kelas minoritas memiliki jumlah data yang sangat sedikit. SMOTE bekerja dengan menciptakan data sintetis untuk kelas minoritas menggunakan interpolasi antara data yang ada. Teknik ini lebih efektif dibandingkan metode oversampling sederhana yang hanya menduplikasi data, karena membantu mengurangi risiko *overfitting* sekaligus mempertahankan informasi penting dalam dataset.

Evaluasi Model

Untuk mengevaluasi akurasi metode, pengujian dilakukan guna meminimalkan kesalahan dan memastikan hasil yang dihasilkan sesuai dengan harapan. Salah satu pendekatan yang digunakan adalah dengan menerapkan *Confusion Matrix* sebagai model klasifikasi. *Confusion Matrix* ini berfungsi untuk menghitung nilai *precision, recall, dan accuracy*. Biasanya, nilai-nilai dalam *Confusion Matrix* disajikan dalam bentuk persentase (%) [9].



Gambar 1. Matriks Evaluasi

$$\textbf{\textit{Accuracy}} = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(2)}$$

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP} \quad \textbf{(3)}$$

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN} \quad \textbf{(4)}$$

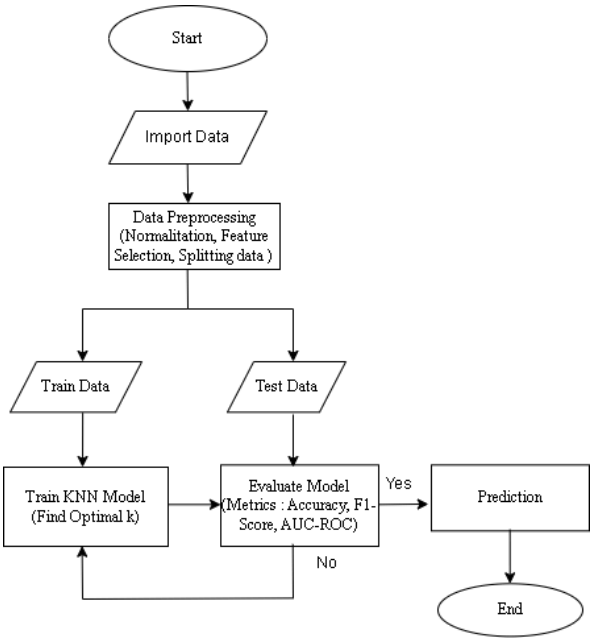
$$\textbf{\textit{F1-Score}} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad \textbf{(5)}$$

Keterangan:

TP *(True Positive)*: Jumlah prediksi yang benar, di mana model memprediksi positif dan nilai sebenarnya memang positif.
FP *(True Positive)*: Jumlah prediksi yang benar, di mana model memprediksi negatif dan nilai sebenarnya memang negatif.
FN *(False Negative)*: Jumlah prediksi yang salah, di mana model memprediksi positif tetapi nilai sebenarnya negatif.
TN *(True Negative)*: Jumlah prediksi yang salah, di mana model memprediksi negatif tetapi nilai sebenarnya positif.

Diagram Alir Penelitian

Berikut adalah diagram alir yang dapat digunakan untuk menggambarkan alur penelitian ini:



Gambar 2. Diagram alir penelitian

Pseudocode

```
# Instalasi dan Import Library
IMPORT sys, time, pandas AS pd, numpy AS np
IMPORT matplotlib.pyplot AS plt, seaborn AS sns
IMPORT sklearn KNeighborsClassifier, train_test_split,
cross_val_score
IMPORT sklearn.preprocessing (MinMaxScaler,
OrdinalEncoder)
IMPORT sklearn.impute (SimpleImputer)
IMPORT sklearn.compose (ColumnTransformer)
IMPORT sklearn.pipeline (Pipeline)
IMPORT sklearn.model_selection (GridSearchCV)
IMPORT sklearn.metrics (confusion_matrix,
ConfusionMatrixDisplay, accuracy_score,
precision_score, recall_score, f1_score,
classification_report)
IMPORT imbalanced_learn.pipeline AS ImbPipeline
IMPORT imbalanced_learn.over_sampling.SMOTE

# Persiapan Data
LOAD Lipinsky & Fingerprint Dataset -> `df_lipinsky`,
`df_fp`
TRANSFORM kelas kategori `class` ke numerik (0, 1, 2)
GABUNGKAN `df_lipinsky` & `df_fp` ->
`df_combined`
HANDLE missing values

# Split dan Preprocessing Data
DROP kolom non-relevan -> X, y
SPLIT `X, y` ke train/test set (80-20)
SCALE data numerik dengan MinMaxScaler()

# K-Nearest Neighbors (KNN) Modeling
INIT KNN (n_neighbors=1) -> Fit ke `X_train` dan
`y_train`
EVALUASI dengan skor train & test set
VALIDASI dengan cross_val_score

# Improvisasi Data & Model
FIT scaler -> Transformasi data `X_train_scaled,
X_test_scaled`
TUNING hyperparameter (n_neighbors, weights,
distance metric) dengan loop & plot
EVALUASI skor terbaik dari tuning

# Pipeline
DEFINISIKAN Pipeline untuk preprocessing dan KNN
TUNING hyperparameter (GridSearchCV)
FIT pipeline ke `X_train` -> EVALUASI train/test score

# Evaluasi Model
Prediksi data uji -> Hitung confusion matrix
EVALUASI metrik: Akurasi, Precision, Recall, F1-Score
TAMPILKAN Classification Report

# Pipeline dengan SMOTE
GUNAKAN SMOTE untuk menangani data imbalance
TUNING parameter -> FIT pipeline -> EVALUASI hasil

# Output
PRINT parameter terbaik, skor model, dan hasil evaluasi
```

HASIL DAN PEMBAHASAN

Pengumpulan Data

Tahap awal penelitian ini dimulai dengan pengumpulan dua jenis dataset utama yang digunakan untuk analisis prediksi senyawa kandidat obat yang menargetkan mutasi KRAS. Dataset pertama, yang disebut Dataset Lipinski, mencakup informasi sifat kimiawi senyawa, seperti berat molekul (*Molecular Weight*/MW), koefisien distribusi logaritmik (LogP), jumlah donor hidrogen (*NumHDonors*), dan jumlah akseptor hidrogen (*NumHAcceptors*). Informasi ini penting untuk menilai kesesuaian senyawa dengan aturan Lipinski, yang merupakan pedoman dasar dalam desain obat. Dataset kedua adalah representasi fingerprint molekul dalam format biner yang diperoleh dari database *PubChem*. Dataset ini memberikan representasi struktural dari senyawa, menunjukkan keberadaan atau ketidakhadiran substruktur tertentu dalam molekul. Kedua dataset ini kemudian digabungkan untuk menghasilkan dataset komprehensif yang mencakup fitur kimia dan struktural, memastikan analisis yang lebih holistik terhadap aktivitas biologis senyawa terhadap target KRAS.

Preprocessing Data

Tahap ini mencakup serangkaian langkah penting untuk mempersiapkan dataset sebelum diterapkan pada model pembelajaran mesin. Label kelas pada dataset pertama, yang mengklasifikasikan senyawa sebagai *active*, *inactive*, atau *intermediate*, diubah menjadi nilai numerik menggunakan teknik pemetaan (0 untuk *active* dan 1 untuk *inactive*). Selanjutnya, fitur numerik seperti MW dan LogP dinormalisasi menggunakan metode *Min-Max Scaling* untuk menyamakan skala data, sehingga perhitungan jarak antar data menjadi lebih akurat. Dataset Lipinski dan fingerprint digabungkan untuk membentuk dataset akhir yang memiliki 886 fitur. Selain itu, *missing values* diatasi menggunakan metode imputasi sederhana, memastikan tidak ada data yang hilang yang dapat mempengaruhi hasil analisis. Proses *preprocessing* ini penting untuk memastikan model pembelajaran mesin dapat bekerja secara optimal.

Split Data

Setelah *preprocessing*, dataset dibagi menjadi dua bagian untuk memastikan model dapat dievaluasi secara adil. Sebanyak 80% data digunakan untuk melatih model, sementara sisanya sebesar 20% dialokasikan untuk pengujian. Pembagian dilakukan secara stratifikasi untuk menjaga distribusi kelas yang seimbang antara data latih dan uji. Hal ini penting mengingat ketidakseimbangan kelas dalam dataset, di mana jumlah senyawa aktif jauh lebih sedikit dibandingkan senyawa tidak aktif. Stratifikasi membantu model untuk lebih memahami pola data pada setiap kelas selama pelatihan dan pengujian.

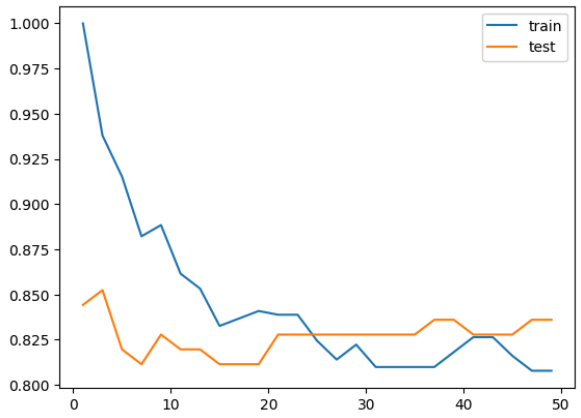
Klasifikasi K-Nearest Neighbor

Model ini dipilih untuk memprediksi aktivitas senyawa terhadap target KRAS. Model ini bekerja berdasarkan prinsip bahwa senyawa dengan karakteristik yang mirip cenderung memiliki label kelas yang sama. Dalam implementasinya, nilai parameter k ditentukan melalui validasi silang, dengan k optimal ditemukan pada nilai 3. Senyawa baru kemudian diklasifikasikan berdasarkan mayoritas kelas dari k tetangga terdekatnya.

Improvisasi Model

1. Parameter Tuning

Proses tuning parameter bertujuan untuk menemukan konfigurasi terbaik dari model *K-Nearest Neighbor* (KNN), yang mencakup jumlah tetangga terdekat (n_neighbors) dan metrik jarak. Dengan menggunakan tqdm untuk memantau progres, dilakukan iterasi pada nilai k dari 1 hingga 51. Model KNN dilatih pada data latih dan diuji pada data uji.

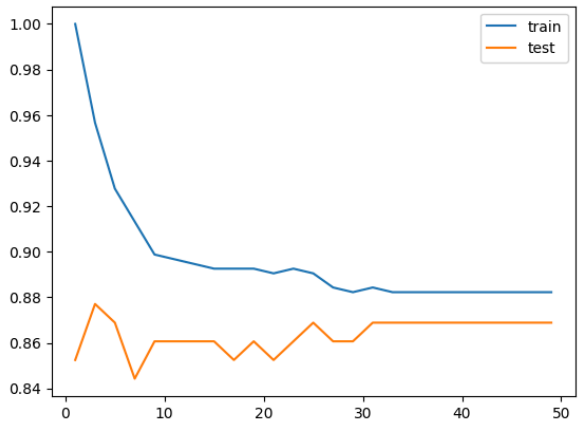


Gambar 3. Plot Akurasi Training dan Testing terhadap Epoch

Hasil *tuning* menunjukkan bahwa nilai k = 3 menghasilkan performa terbaik dengan akurasi pada data uji sebesar 85.25%. Grafik antara nilai k dengan akurasi menunjukkan adanya keseimbangan antara performa pada data latih dan uji, mengindikasikan bahwa model tidak mengalami *overfitting* atau *underfitting* secara signifikan.

2. *Tuning* dengan *Scaling*

Untuk memastikan bahwa setiap fitur memiliki skala yang sebanding, dilakukan normalisasi data menggunakan *MinMaxScaler*. Proses ini dilakukan dalam pipeline *preprocessing* sebelum melatih model.

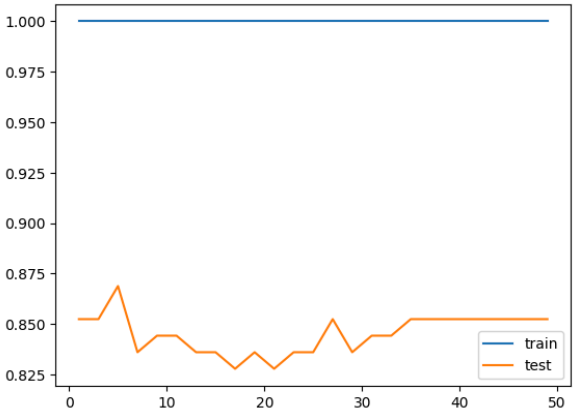


Gambar 4. Plot Akurasi Training dan Testing terhadap Epoch

Setelah *scaling*, dilakukan proses *tuning* yang sama pada parameter n_neighbors, menghasilkan akurasi yang lebih tinggi, yaitu 87.70% pada data uji dengan nilai k = 3. Peningkatan ini mengindikasikan bahwa normalisasi data membantu KNN dalam menghitung jarak antar data secara lebih akurat, terutama untuk dataset yang memiliki fitur dengan skala yang sangat berbeda.

3. *Tuning* dengan Pembobotan Jarak (*Weighted Distance*)

Untuk meningkatkan kemampuan model dalam mengklasifikasikan senyawa, dilakukan eksperimen dengan menambahkan pembobotan jarak pada parameter weights. Dengan menggunakan opsi *distance*, KNN memberikan bobot lebih tinggi pada tetangga yang lebih dekat dengan data uji, dibandingkan tetangga yang lebih jauh.

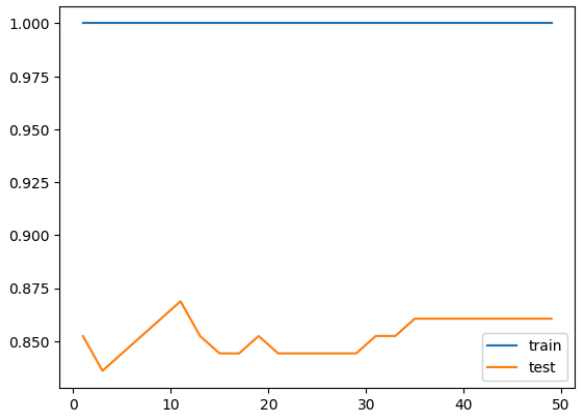


Gambar 5. Plot Akurasi Training dan Testing terhadap Epoch

Hasil *tuning* menunjukkan bahwa dengan k = 5 dan pembobotan jarak, model mencapai akurasi 86.88% pada data uji. Pendekatan ini memberikan keunggulan pada prediksi untuk data dengan pola distribusi yang tidak seragam, terutama jika ada *outlier* atau senyawa dengan kemiripan tinggi di area tertentu.

4. Menggunakan *Manhattan Distance*

Selain metrik *Euclidean*, diterapkan metrik *Manhattan Distance* (parameter p=1) untuk menghitung jarak antara data. Metrik ini lebih *robust* terhadap *outlier* karena menghitung jarak berdasarkan perbedaan absolut antara fitur.



Gambar 6. Plot Akurasi Training dan Testing terhadap Epoch

Pada tuning ini, nilai k = 11 menghasilkan akurasi terbaik sebesar 86.88%. Grafik akurasi terhadap nilai k menunjukkan bahwa *Manhattan Distance* lebih stabil pada nilai k yang lebih besar, membuatnya cocok untuk dataset dengan dimensi tinggi seperti fingerprint molekul.

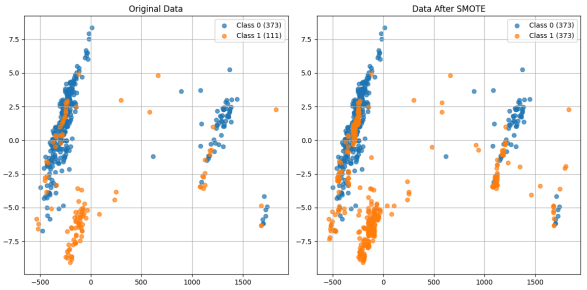
Klasifikasi KNN Dengan SMOTE

Penggunaan SMOTE diterapkan untuk mengatasi ketidakseimbangan kelas dalam dataset, khususnya pada kelas minoritas, yaitu senyawa aktif. SMOTE bekerja dengan menghasilkan data sintetis melalui interpolasi antara sampel-sampel yang ada, sehingga distribusi kelas menjadi lebih seimbang. Hal ini memungkinkan model pembelajaran mesin, seperti KNN, untuk lebih sensitif terhadap kelas minoritas yang sebelumnya sulit terdeteksi akibat ketimpangan jumlah data.

Setelah penerapan SMOTE, model KNN dilatih ulang menggunakan konfigurasi parameter optimal yang diperoleh sebelumnya melalui proses optimisasi GridSearchCV, yaitu *n_neighbors = 33*, *weights = 'distance'*, dan *p = 1* (Manhattan Distance). GridSearchCV mengevaluasi 100 kombinasi parameter menggunakan validasi silang 3-fold untuk memastikan model tetap menggunakan konfigurasi terbaik, meskipun dataset telah mengalami perubahan karena penerapan SMOTE. Dengan konfigurasi ini, model KNN menunjukkan performa yang baik pada data yang telah diseimbangkan.

Hasil pelatihan ulang menunjukkan bahwa SMOTE meningkatkan sensitivitas model terhadap kelas minoritas. Model menjadi lebih mampu mengenali senyawa aktif, yang sebelumnya sulit dideteksi akibat distribusi data yang tidak seimbang. Namun, meskipun sensitivitas terhadap kelas minoritas meningkat, akurasi keseluruhan model mengalami sedikit penurunan. Penurunan ini disebabkan oleh data sintetis yang dihasilkan oleh SMOTE, yang dapat memengaruhi distribusi asli data dan menambah variasi yang tidak selalu sepenuhnya mencerminkan karakteristik nyata senyawa.

Pada data pelatihan, model mencapai skor akurasi 1.0, menunjukkan kemampuan yang sempurna dalam mengenali pola dari data pelatihan. Sementara itu, pada data uji, akurasi model mencapai 0.8524 (85,24%), yang tetap menggambarkan kemampuan generalisasi yang cukup baik meskipun terdapat sedikit penurunan dibandingkan data pelatihan..



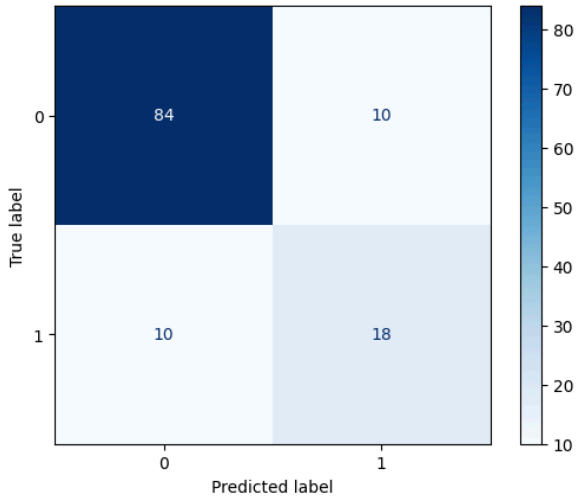
Gambar 7. Perbandingan data sebelum dan sesudah SMOTE

Gambar 7 memperlihatkan visualisasi data sebelum dan sesudah penerapan SMOTE. Pada bagian kiri, terlihat distribusi data asli di mana Class 0 (berwarna biru) mendominasi dengan 373 sampel, sedangkan Class 1 (berwarna oranye) hanya memiliki 111 sampel. Ketidakseimbangan kelas yang signifikan ini membuat model cenderung bias terhadap kelas mayoritas. Setelah SMOTE diterapkan, seperti yang ditampilkan pada bagian kanan, jumlah sampel Class 1

meningkat hingga menyamai Class 0, yaitu 373 sampel. Proses SMOTE tidak hanya menyalin sampel yang ada, tetapi juga menghasilkan data sintetis dengan interpolasi antar-sampel dalam kelas minoritas, sehingga distribusi kelas menjadi seimbang. Visualisasi ini menunjukkan keberhasilan SMOTE dalam menangani ketimpangan data dan memberikan peluang yang lebih adil bagi model untuk mempelajari fitur dari kedua kelas.

Evaluasi Hasil

Evaluasi kinerja model dilakukan menggunakan beberapa metrik utama, yaitu akurasi, *precision, recall,* dan *F1-score*. Akurasi model pada data uji mencapai 84%, dengan *precision, recall*, dan *F1-score* masing-masing juga sebesar 84%. Meskipun akurasi ini cukup baik, penerapan SMOTE meningkatkan *recall* untuk kelas aktif, yang berarti lebih banyak senyawa aktif berhasil teridentifikasi.



Gambar 8. *Confusion Matrix*

Pada tabel *confusion matrix* yang ditampilkan pada Gambar 5, hasil klasifikasi model KNN menunjukkan rincian sebagai berikut:

1. *True Positives* (TP): Sebanyak 18 sampel kelas senyawa aktif diklasifikasikan dengan benar oleh model.
2. *True Negatives* (TN): Sebanyak 84 sampel kelas senyawa tidak aktif juga diklasifikasikan dengan benar.
3. *False Positives* (FP): Sebanyak 10 sampel dari kelas senyawa tidak aktif salah diklasifikasikan sebagai kelas senyawa aktif.
4. *False Negatives* (FN): Sebanyak 10 sampel dari kelas senyawa aktif salah diklasifikasikan sebagai kelas senyawa tidak aktif.

Tabel 3. Hasil Prediksi

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

Hasil klasifikasi yang ditampilkan dalam Tabel 1 menunjukkan kinerja model dalam mengklasifikasikan senyawa aktif (kelas 1) dan senyawa tidak aktif (kelas 0). Metrik seperti *precision, recall,* dan *F1-score* memberikan gambaran rinci tentang performa model. *Precision* untuk kelas 0 adalah 0.89, yang berarti 89% dari sampel yang diklasifikasikan sebagai kelas 0 benar-benar senyawa tidak aktif. Begitu juga dengan *recall* yang mencapai 0.89 untuk kelas 0, yang menunjukkan bahwa model berhasil mengidentifikasi 89% dari semua sampel senyawa tidak aktif. Sebaliknya, untuk kelas 1, *precision* dan *recall* masing-masing adalah 0.64, mengindikasikan bahwa model hanya mengidentifikasi 64% dari sampel senyawa aktif dengan benar.

F1-score, yang mengukur keseimbangan antara *precision* dan *recall*, menunjukkan nilai yang lebih tinggi untuk kelas 0 (0.89) dibandingkan kelas 1 (0.64), mencerminkan performa yang lebih baik dalam klasifikasi senyawa tidak aktif. Secara keseluruhan, akurasi model adalah 0.84, yang berarti model berhasil mengklasifikasikan 84% dari semua sampel dengan benar. Selain itu, support menunjukkan distribusi sampel yang tidak seimbang, dengan kelas 0 memiliki 94

sampel dan kelas 1 hanya 28 sampel. Rata-rata mikro (*micro avg*) dan rata-rata tertimbang (*weighted avg*) memberikan gambaran kinerja model secara keseluruhan, yang menunjukkan bahwa meskipun model memiliki kinerja yang baik pada kelas 0, performa untuk kelas 1 perlu ditingkatkan, terutama pada precision dan recall yang lebih rendah.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa model *K-Nearest Neighbor* dapat digunakan untuk memprediksi aktivitas senyawa kandidat obat terhadap target KRAS dengan tingkat akurasi yang cukup baik, yaitu mencapai 84% pada data uji. Proses tuning parameter, penerapan normalisasi, dan pengoptimalan model menggunakan pembobotan jarak serta metrik *Manhattan Distance* memberikan peningkatan performa yang cukup signifikan. Selain itu, penggunaan SMOTE juga berhasil meningkatkan sensitivitas model terhadap kelas minoritas (senyawa aktif), evaluasi menggunakan metrik *precision, recall,* dan *F1-score* menunjukkan bahwa model lebih unggul dalam mengklasifikasikan senyawa tidak aktif dibandingkan senyawa aktif, yang masih menjadi tantangan dalam penelitian ini.

UCAPAN TERIMA KASIH

Kami ingin mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Tirta Setiawan, S.Pd., M.Si. Dosen Bioinformatika sekaligus pembimbing kami, yang telah memberikan bimbingan, arahan, dan dukungan yang luar biasa selama proses penelitian ini. Terima kasih atas waktu, pengetahuan, dan kesabaran yang telah diberikan, sehingga kami dapat menyelesaikan penelitian ini dengan baik.

DAFTAR RUJUKAN

1. Ajmal A, et al. In silico prediction of new inhibitors for Kirsten rat sarcoma G12D cancer drug target using machine learning-based virtual screening, molecular docking, and molecular dynamic simulation approaches. *Pharmaceutics*. 2024;17(5):551. doi: 10.3390/ph17050551.
2. Dewi NNYA, Pranata AAGNS, Suksmarini NMPW. Mutasi gen KRAS pada kanker kolorektal. *Maj Kedokteran Andalas*. 2021;44(2):117-125.
3. Xie X, Yu T, Li X, Zhang N, Foster LJ, Peng C, et al. Recent advances in targeting the 'undruggable' proteins: from drug discovery to clinical trials. *Signal Transduct Target Ther*. 2023;8(335). doi: 10.1038/s41392-023-01589-z.
4. Li Y, Umbach DM, Krahn JM, Shats I, Li X, Li L. Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. *BMC Genomics*. 2021;22(272). doi: 10.1186/s12864-021-07581-7. PMID: 33858332; PMCID: PMC8048084.
5. Ajmal A, Alkhatabi HA, Alreemi RM, et al. Prospective virtual screening combined with bio-molecular simulation enabled identification of new inhibitors for the KRAS drug target. *BMC Chem*. 2024;18(57). Available from: https://doi.org/10.1186/s13065-024-01152-z.
6. Alfiyanti YD, Ratnawati DE, Anam S. Klasifikasi fungsi senyawa aktif data berdasarkan kode Simplified Molecular Input Line Entry System (SMILES) menggunakan metode Modified K-Nearest Neighbor. *J Pengembangan Teknol Inf Ilmu Komputer*. 2019;3(4):3244–3251.
7. Saputra SK, et al. Pneumonia identification based on lung texture analysis using modified k-nearest neighbour. *J Phys Conf Ser*. 2022;2193(1).
8. Munandar TA, Munir AQ. Implementasi K-Nearest Neighbor untuk prototype sistem pakar identifikasi dini penyakit jantung. *Respati*. 2022;17(2):44–50.
9. Firmanto B, Soekotjo H, Dachlan HS. Perbandingan kinerja algoritma PROMETHEE dan TOPSIS untuk pemilihan guru teladan. *J Penelitian Pendidikan IPA*. 2016;2(1).
10. Y. Miftahuddin, S. Umaroh, and F. R. Karim, "PERBANDINGAN METODE PERHITUNGAN JARAK EUCLIDEAN, HAVERSINE , (STUDI KASUS : INSTITUT TEKNOLOGI NASIONAL BANDUNG)," vol. 14, no. 2, pp. 69–77, 2020