

Kecocokan Keputusan Pohon Algoritma pada Kimia Organik: Perbandingan ROC AUC Keputusan Pohon dan Ketetangaan

Muhammad Zirlda Prairi^{1*}, Zurnan Alfian², Kecitaan Harefa³
^{1,2,3}Prodi Teknik Informatika, Universitas Pamulang, Pamulang, Indonesia
Corresponding Author: muhammadzirldaprairi999@gmail.com

Published: July, 2025

ABSTRAK

Kehadiran Machine Learning (ML) di ruang lingkup komputasi modern telah menyebabkan banyak permasalahan dan solusi, terutama pembahasan algoritma. Penelitian ini menganalisa dan eksplorasi efektifitas lima algoritma ML dalam mengklasifikasi asam amino lazim esensial pada protein berdasarkan strukur molekul dan jumlah atom. Dataset diambil secara manual dari buku kimia organik klasik, yang dikonversi dari rumus dan gambar struktur senyawa asam amino lazim menjadi fitur numerik seperti jumlah atom karbon, hidrogen, nitrogen, oksigen, dan sulfur. Kelima algoritma ML yang dianalisis yakni Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbour (KNN), K-Means, dan Random Forest. Setiap algoritma dilakukan evaluasi menggunakan nilai akurasi, *precision*, *recall*, *F1-score*, serta Area Under the Curve (AUC) dari kurva Receiver Operating Characteristic (ROC). Hasil menunjukkan bahwa algoritma Decision Tree, Random Forest, dan KNN memiliki tingkat akurasi terbaik dengan skor AUC sebesar 0,83. Studi kasus penelitian ini menawarkan pendekatan pengubahan data mentah menjadi fitur yang bisa dimengerti oleh algoritma ML dan klasifikasi biner dengan label nol dan satu pada dataset berskala kecil. Penelitian berikutnya disarankan menggunakan dataset yang lebih besar dan menerapkan validasi silang untuk meningkatkan generalisasi model.

Kata Kunci: Machine Learning, Decision tree, Naive bayes, Random forest, K-Nearest Neighbour, K-Means, Kimia organik

ABSTRACT

The presence of machine learning (ML) in the realm of modern computing has led to many problems and solutions, especially in the discussion of algorithms. This research analyzes and explores the effectiveness of five ML algorithms in classifying common essential amino acids in proteins based on molecular structure and the number of atoms. The dataset was manually collected from classic organic chemistry books, converted from formulas and structural images of common amino acid compounds into numerical features such as the number of carbon, hydrogen, nitrogen, oxygen, and sulfur atoms. The five ML algorithms analyzed are Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbour (KNN), K-Means, and Random Forest. Each algorithm was evaluated using accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curve. The results showed that the Decision Tree, Random Forest, and KNN algorithms had the best accuracy levels with an AUC score of 0.83. This research case study offers an approach to transforming raw data into features that can be understood by ML algorithms and binary classification with zero and one labels on a small-scale dataset. Future research is recommended to use larger datasets and apply cross-validation to improve model generalization

Keywords: Machine Learning, Decision tree, Naive bayes, Random forest, K-Nearest Neighbour, K-Means, Organic chemistry

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0>

PENDAHULUAN

Digunakan secara luas dalam berbagai bidang seperti pendidikan, keuangan, dan kesehatan, Machine Learning (ML) telah menjadi teknologi utama dalam pengolahan data dan memiliki banyak keunggulan, salah satunya adalah kemampuan untuk melakukan klasifikasi berdasarkan pola yang terbentuk dari fitur numerik. Machine learning atau pembelajaran mesin adalah sebuah teknologi yang diciptakan untuk

mampu belajar sendiri dalam memecahkan masalah berdasarkan data-data yang pernah diberikan oleh berbagai penggunaannya [1]. Selama dua dekade terakhir Machine Learning telah menjadi salah satu andalan teknologi informasi dan dengan itu, bagian yang agak sentral, meskipun biasanya tersembunyi, dari hidup kita [2]. Dalam konteks data kimia, khususnya struktur senyawa dan molekul, penerapan ML masih relatif terbatas dibandingkan bidang lainnya.

Asam amino sebagai unit pembangun protein, memiliki struktur molekul hampir seragam yang terdiri dari atom karbon (C), hidrogen (H), nitrogen (N), dan oksigen (O), serta beberapa zat mengandung unsur sulfur (S). Setiap masing-masing asam amino dan pola ikatannya menyumbang sifat yang unik kepada masing-masing protein [3]. Pola asam amino ini terletak pada perbedaan jumlah atom setiap unsur, menjadikannya cocok digunakan untuk sebagai fitur numerik dalam algoritma ML di bidang bioinformatika dan kimia komputasi.

Penelitian terdahulu pernah dilakukan oleh Then et al. 2020 [4] mengusulkan metode klasifikasi asam amino dengan metode barunya mereka yang melibatkan pemisahan *binary* berulang melalui jumlah minimum lima fitur seperti hidrofobisitas atau volume, dinyatakan dengan nilai numerik untuk karakteristik asam amino, dan dieskraksikan fitur-fitur tersebut dari basis data AAindex [4]. Namun pendekatan metode tersebut cukup rumit dan kurang relevan untuk implementasi praktis di lingkungan pendidikan awal. Dalam penelitian ini, berdasarkan pola molekul asam amino berdasarkan perbedaan jumlah atom setiap unsur, bisa disederhanakan menjadi jumlah atom setiap penyusun masing-masing molekul asam amino sebagai fitur input dengan fokus pada pendekatan klasifikasi yang lebih ringan.

Dalam akademik prodi Teknik Informatika, penerapan ML umumnya berfokus pada data kependudukan, kesehatan, keuangan, atau ekonomi, yang sudah terlihat dataset atau tabel angka sebagai fitur input. Jarang sekali menggunakan kimia sebagai dataset, membahas AI sebagai alat dalam penelitian dalam pendidikan kimia masih terbatas [5]. Maka dari itu, penelitian ini hadir sebagai contoh penerapan ML dengan sederhana namun bernilai edukatif dan membuka wawasan penggunaan lintas disiplin dalam pemodelan klasifikasi.

KAJIAN LITERATUR

Penelitian ini menggunakan kajian-kajian yang berkaitan dengan Machine Learning (ML). ML umumnya terbagi menjadi dua kategori utama, yaitu *supervised* dan *unsupervised*. Algoritma yang digunakan adalah untuk klasifikasi fitur seperti *decision tree*, *naive bayes*, K-Nearest Neighbour (KNN), *random forest*, dan *K-Means* sebagai algoritma prediksi pola dalam data.

Machine Learning dan Klasifikasi

Kecerdasan manusia menjadi inspirasi untuk diciptakannya kecerdasan buatan, dimana banyak sekali hasil dari kecerdasan buatan diantaranya adalah Machine Learning (ML). Pengembangan kemampuan komputer agar dapat menyerupai kemampuan manusia [6]. ML memungkinkan sistem belajar dari data tanpa perlu diprogram secara manual dan eksplisit. Salah satu penerapannya adalah klasifikasi, yang merupakan proses memetakan input data ke dalam label kelas tertentu [2].

Umumnya klasifikasi pada ML untuk pendidikan awal adalah klasifikasi biner. Kasus paling dasar untuk data, dimana hanya terdapat dua label atau kelas, misalnya “ya” dan “tidak” pada label esensial sesuai dataset yang akan digunakan. Klasifikasi biner merupakan bentuk paling dasar dari supervised learning, di mana model mempelajari pola untuk memetakan input ke salah satu dari dua kategori diskret [7]. Tipe klasifikasi ini digunakan secara luas seperti pengenalan pola, biomedis, dan deteksi anomali. Dalam konteks penelitian ini, klasifikasi biner diterapkan untuk memisahkan asam amino menjadi dua kelas, yaitu esensial dan tidak esensial, berdasarkan jumlah numerik dari struktur molekulnya.

Supervised dan Unsupervised Learning

Supervised learning adalah metode pembelajaran mesin di mana algoritma dilatih menggunakan data berlabel, yang berarti inputan data disertai dengan keluaran benar yang sesuai [8]. Maksudnya adalah data label atau target ikut berperan sebagai 'supervisor' atau 'guru' yang mengawasi proses pembelajaran dalam mencapai tingkat akurasi atau presisi tertentu [9]. Jadi dataset yang digunakan sudah memiliki label, sehingga mesin belajar pemodelan data tersebut dan menghasilkan label yang sesuai atau mendekati sesuai jika diberikan data baru. Seperti halnya anak yang berlatih mengenali kendaraan motor dengan melihat di persimpangan jalan didampingi orang tuanya terlebih dahulu, sehingga siswa tersebut bisa mengenali kendaraan motor yang lewat dengan sendirinya.

Sedangkan *Unsupervised learning* merupakan paradigma pembelajaran mesin yang berfokus pada mengungkap pola, struktur, atau hubungan dalam data yang tidak berlabel [8]. Mesin tidak perlu diajari diberi jawaban yang benar. Sebaliknya, mesin diizinkan bekerja sendiri untuk menemukan informasi [9]. Pola terbentuk yang dihasilkan murni dari perhitungan algoritma ML itu sendiri, tidak ada label sebagai penanda mana yang benar dan tidak. Seperti siswa yang diberi tugas oleh guru untuk menganalisa beberapa item makanan dan mengelompokkannya sesuai pemikiran siswa tersebut.

Algoritma Klasifikasi yang Digunakan

Algoritma yang digunakan untuk penelitian klasifikasi adalah *decision tree*, *naive bayes*, K-Nearest Neighbour (KNN), *random forest*, dan *K-Means*.

Validasi Silang

Validasi silang merupakan metode evaluasi model yang digunakan untuk menghindari bias akibat pembagian data tunggal. Salah satu bentuk paling umum adalah K-Fold Cross-Validation, di mana dataset dibagi menjadi k bagian, dan pelatihan dilakukan sebanyak k kali dengan bagian berbeda sebagai data uji. Teknik ini dinilai efektif untuk meningkatkan generalisasi model dan mengurangi risiko overfitting, terutama pada dataset berskala kecil [17].

Metrik Evaluasi Model Klasifikasi

Metrik evaluasi adalah satuan alat ukur kuantitatif yang digunakan untuk menilai seberapa baik model klasifikasi bekerja dalam memprediksi hasil. Metrik evaluasi dalam klasifikasi digunakan untuk mengukur kinerja model terhadap data uji. Beberapa metrik umum yang digunakan antara lain akurasi, *precision*, *recall*, *F1-score*, dan Area Under the Curve (AUC) dari kurva Receiver Operating Characteristic (ROC). Setiap metrik memberikan sudut pandang berbeda terhadap performa model, tergantung pada proporsi kelas, konsekuensi kesalahan, dan kebutuhan prediksi [18]. Kurva ROC dibuat berdasarkan nilai yang telah didapatkan dari perhitungan dengan confusion matrix, yaitu antara False Positive Rate dengan True Positive Rate [19].

Dalam evaluasi algoritma model klasifikasi biner, terdapat empat kemungkinan hasil prediksi: True Positive (TP) jika model benar memprediksi kelas positif (misalnya kelas yang berlabel "ya"), True Negative (TN) jika model benar memprediksi kelas negatif (misalnya kelas yang berlabel "tidak"), False Positive (FP) jika model salah memprediksi positif padahal sebenarnya negatif, dan False Negative (FN) jika model salah memprediksi negatif padahal sebenarnya positif.

Studi Sebelumnya yang Relevan

Then et al. (2020) mengusulkan metode klasifikasi asam amino berdasarkan pemisahan binary pada fitur-fitur fisikokimia dari basis data AAindex. Metode tersebut kompleks dan tidak sesuai untuk eksperimen edukatif sederhana [4]. Studi lain oleh Iyemuremye et al. (2025) menunjukkan bahwa ML dapat dimanfaatkan untuk klasifikasi molekul sebagai pendekatan edukatif interdisiplin di bidang kimia dan informatika [20][21].

Berdasarkan kajian literatur di atas, dapat disimpulkan bahwa penerapan ML pada data molekul kimia organik, khususnya klasifikasi biner pada struktur asam amino, masih jarang ditemukan dan memiliki potensi pengembangan, khususnya sebagai studi edukatif lintas disiplin antara bidang informatika dan kimia.

METODOLOGI

Penelitian ini termasuk dalam jenis penelitian kuantitatif eksperimental komparatif, yang bertujuan untuk membandingkan performa beberapa algoritma machine learning dalam mengklasifikasikan asam amino berdasarkan struktur kimianya. Data diperoleh dari sumber sekunder, yaitu buku cetak, dan dikonversi menjadi format numerik untuk dianalisis secara komputasional. Proses pengujian dilakukan secara daring menggunakan Python, dengan fokus pada analisis model, evaluasi performa, dan validasi silang untuk meningkatkan akurasi generalisasi. Metodologi berikut menjelaskan tahapan eksperimen secara sistematis, mulai dari perolehan data hingga evaluasi hasil.

Dataset dan Representasi Fitur

Dataset yang digunakan dalam penelitian ini terdiri dari 20 asam amino lazim pada protein, diklasifikasikan ke dalam dua kelas yaitu esensial dan tidak esensial. Data diperoleh secara manual dari buku Kimia Organik karya Fessenden (1999) [22], kemudian dikonversikan ke dalam bentuk numerik berdasarkan jumlah atom penyusunnya: karbon (C), hidrogen (H), nitrogen (N), oksigen (O), dan sulfur (S). Setiap sample direpresentasikan sebagai satu baris data dengan lima fitur numerik dan satu label kelas biner, ditambah fitur lain sebagai atribut deskriptif non-pelatihan seperti fitur nama unsur, singkatan unsur, rumus unsur, dan lima fitur yang diisi oleh singkatan atom.

Tabel 1. Dataset Daftar Asam Amino Lazim pada Protein

Nama	Singkatan	Struktur	StrukturC JumlahC	StrukturH JumlahH	StrukturN JumlahN	StrukturO JumlahO	StrukturS JumlahS	Esensial
alanine	ala	C3H7NO2	C 3	H 7	N 1	O 2	S 0	Tidak
arginina	arg	C6H14N4O2	C 6	H 14	N 4	O 2	S 0	Ya
asparagine	asn	C4H8N2O3	C 4	H 8	N 2	O 3	S 0	Tidak
asam aspartat	asp	C4H7NO4	C 4	H 7	N 1	O 4	S 0	Tidak
sisteina	cys	C3H7NO2S	C 3	H 7	N 1	O 2	S 1	Tidak
asam glutamat	glu	C5H9NO4	C 5	H 9	N 1	O 4	S 0	Tidak
glutamine	gln	C5H10N2O3	C 5	H 10	N 2	O 3	S 0	Tidak
glisina	gly	C2H5NO2	C 2	H 5	N 1	O 2	S 0	Tidak
histidine	his	C6H9N3O2	C 6	H 9	N 3	O 2	S 0	Ya
isoleusina	ile	C6H13NO2	C 6	H 13	N 1	O 2	S 0	Ya
leusina	leu	C6H13NO2	C 6	H 13	N 1	O 2	S 0	Ya
lisina	lys	C6H14N2O2	C 6	H 14	N 2	O 2	S 0	Ya
metionina	met	C5H11NO2S	C 5	H 11	N 1	O 2	S 1	Ya
fenilalanina	phe	C9H11NO2	C 9	H 11	N 1	O 2	S 0	Ya
prolina	pro	C5H9NO2	C 5	H 9	N 1	O 2	S 0	Tidak
serina	ser	C3H7NO3	C 3	H 7	N 1	O 3	S 0	Tidak
treonina	thr	C4H9NO3	C 4	H 9	N 1	O 3	S 0	Ya
triptofan	try	C11H12N2O2	C 11	H 12	N 2	O 2	S 0	Ya
tirosina	tyr	C9H11NO3	C 9	H 11	N 1	O 3	S 0	Tidak
valina	val	C5H11NO2	C 5	H 11	N 1	O 2	S 0	Ya

Pra-pemrosesan Data

Data diketik ulang dan disusun dalam format Comma Separated Values (CSV), kemudian diolah menggunakan Python versi 3.11.0 dengan bantuan pustaka pandas. Label kelas dikodekan dalam bentuk numerik: 1 untuk esensial dan 0 untuk tidak esensial. Karena fitur berada dalam rentang kecil dan seragam, normalisasi tidak diterapkan, serta tidak menunjukkan distribusi ekstrem [23]. Namun, ini diakui sebagai salah satu keterbatasan studi, terutama pada penggunaan algoritma KNN dan K-Means yang sensitif terhadap skala. Hal ini diakui sebagai keterbatasan. Validasi silang yang dilakukan diharapkan dapat membantu mengurangi bias akibat perbedaan skala ini.

Algoritma yang Digunakan

Lima algoritma machine learning digunakan dalam eksperimen ini:

1. Decision Tree
2. Gaussian Naive Bayes
3. K-Nearest Neighbour (KNN) dengan parameter $n_neighbors=3$
4. Random Forest dengan parameter $n_estimators=100$
5. K-Means (unsupervised, digunakan untuk eksplorasi pola)

Semua algoritma diimplementasikan menggunakan pustaka scikit-learn versi 1.6.1.

Teknik Evaluasi dan Validasi Silang

Data dievaluasi menggunakan metode Stratified K-Fold Cross Validation dengan $k=5$, yaitu teknik validasi silang yang mempertahankan distribusi label pada tiap fold. Dengan pendekatan ini, dataset dibagi menjadi 5 bagian yang seimbang [24], dan pelatihan dilakukan 5 kali menggunakan bagian yang berbeda sebagai data uji. Hasil evaluasi dari kelima fold dirata-ratakan untuk memperoleh metrik performa akhir.

Metrik Evaluasi

Evaluasi dilakukan menggunakan lima metrik utama:

1. Akurasi: rasio prediksi yang tepat terhadap seluruh prediksi.
2. Precision: rasio prediksi positif yang benar dari seluruh prediksi positif.
3. Recall: rasio prediksi positif yang benar dari seluruh data sebenarnya positif.
4. F1-Score: rata-rata harmonik dari precision dan recall.
5. AUC-ROC: luas area di bawah kurva Receiver Operating Characteristic, mengukur kemampuan model membedakan kelas.

Perangkat dan Lingkungan Eksperimen

Penelitian dilakukan menggunakan Python 3.11.0 dengan pustaka berikut yang relevan: scikit-learn, pandas, numpy, matplotlib, seaborn, dan python-docx. Eksperimen dijalankan di Jupyter Notebook. Daftar lengkap pustaka tambahan dapat ditemukan pada lampiran atau dokumentasi eksperimen.

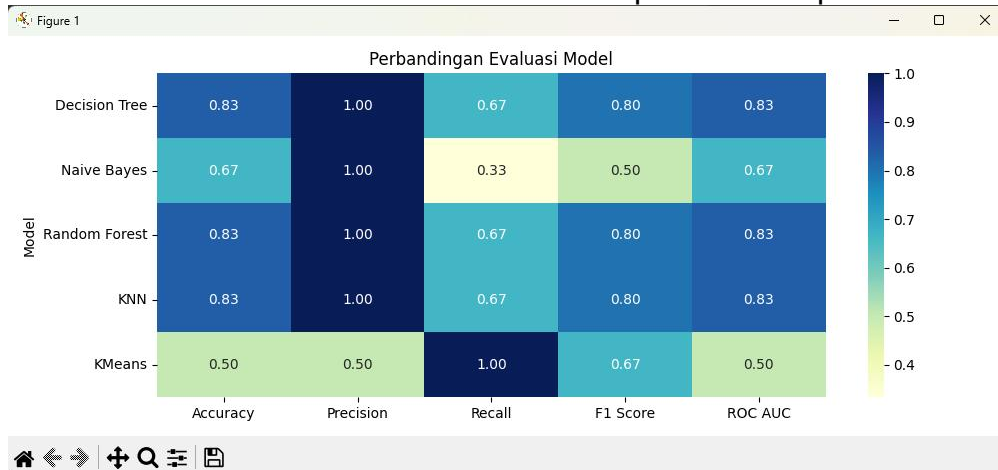
HASIL DAN PEMBAHASAN

Serangkaian penelitian ini memiliki beberapa hasil nilai evaluasi, diantaranya *heatmap* evaluasi model, *heatmap cross validation* dengan 5 fold, dan grafik ROC AUC.

Visualisasi Heatmap Hasil Evaluasi

Berdasarkan pengujian kelima algoritma Machine Learning (ML) yaitu *decision tree*, *gaussian naive bayes*, *random forest*, K-Nearest Neighbour (KNN), dan K-Means, untuk klasifikasi jenis asam amino lazim pada protein menjadi dua kelas: esensial dan tidak esensial. Evaluasi dilakukan menggunakan nilai akurasi, *precision*, *recall*, *F1-score*, serta nilai Area Under the Curve dari kurva Receiver Operating Characteristic (AUC ROC). Visualisasi berikut menampilkan performa kelima model dalam bentuk heatmap, untuk memudahkan perbandingan antar metrik.

Gambar 1. Hasil Evaluasi Model pada Heatmap



Gambar 1 menunjukkan *heatmap* perbandingan nilai evaluasi dari kelima algoritma yang diteliti: *decision tree*, *gaussian naive bayes*, *random forest*, KNN, dan K-Means. Setiap sel pada gambar yang berbentuk tabel mewakili nilai metrik evaluasi (*accuracy*, *precision*, *recall*, F1-score, dan ROC AUC) untuk setiap model, dengan tambahan skala warna dari rendah (kuning pucat) ke tinggi (biru tua).

Pertama, nilai akurasi. Laporan yang memberikan ukuran seberapa baik kinerja klasifikasi model dengan menunjukkan sejumlah prediksi yang benar di antara jumlah total kasus yang dievaluasi, makanya berdasarkan rumus, akurasi menghitung total jumlah prediksi yang benar (baik positif maupun negatif) dari seluruh data [18]. Rumus akurasi sebagai berikut.

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

Nilai Akurasi dari kelima algoritma yang diteliti, algoritma *decision tree*, *random forest*, dan KNN memiliki nilai skor tertinggi, yaitu sebesar 0,83. Disusul algoritma *gaussian naive bayes* dengan skor sebesar 0,67 dimana akurasi klasifikasi data cukup baik, dan paling kecil adalah algoritma K-Means dengan skor sebesar 0,5, artinya hasil akurasi klasifikasi data biner sama dengan tebak-tebakan. Karena memang K-Means adalah model *unsupervised learning* sehingga tidak membutuhkan label seperti keempat model lainnya.

Kedua terdapat nilai *precision*, salah satu metrik evaluasi yang diberikan dalam laporan klasifikasi dimana rasio prediksi positif sejati pengklasifikasi sama dengan jumlah total prediksi positif yang dibuatnya [18]. Sehingga pada rumusan *precision* hanya menghitung rasio jumlah benar dan salah yang positif, berikut rumusannya.

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

Nilai *precision* pada kelima algoritma yang diteliti, keempat algoritma (*decision tree*, *gaussian naive bayes*, *random forest*, dan KNN) mendapatkan skor sempurna, yakni 1,0. Ini artinya, keempat model algoritma tersebut sangat yakin dalam memprediksi kelas esensial pada dataset asam amino lazim dalam protein. Sedangkan K-Means hanya mendapatkan skor 0,5 yang artinya model ini seperti tebak-tebakan dalam memprediksi kelas esensial dikarenakan model ini adalah *unsupervised learning*.

Selain *precision*, terdapat *recall* yang berfokus pada nilai yang menunjukkan tingkat keberhasilan atau spesifisitas untuk mengetahui kembali sebuah informasi secara benar tentang data yang kelas negatif atau pun konten teks positif [25]. Pada perumusannya, rasio dihitung berdasarkan benar memprediksi kelas esensial pada dataset asam amino lazim dalam protein dan salah memprediksi kelas tidak esensial yang seharusnya masuk ke kelas esensial. Berikut rumusannya.

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

Nilai *recall* pada kelima algoritma yang diteliti, algoritma K-Means menunjukkan nilai recall yang sempurna yakni sebesar 1,0 yang artinya seluruh data esensial berhasil terdeteksi. Disusul ketiga algoritma (*decision tree*, *random forest*, dan KNN) menunjukkan skor sebesar 0,67 yang artinya algoritma tersebut masih melewatkan sebagian data esensial pada data asam amino lazim dalam protein (FN). Sayangnya algoritma *gaussian naive bayes* menunjukkan skor 0,33 yang artinya klasifikasi ini lebih rendah dibanding tebak-tebakan.

Kemudian terdapat metrik evaluasi yang menggabungkan *precision* dan *recall* untuk menghitung rasionya, yaitu *F1-score*. F1-skor adalah metrik yang menggabungkan *precision* dan *recall* kembali menjadi satu nilai [18] , jadinya perhitungan dengan rata-rata dari *precision* dan *recall*, sangat berguna untuk dataset tidak seimbang karena mengimbangi kesalahan positif dan negatif. Berikut rumusannya.

$$F1 = \frac{2(precision)(recall)}{(precision)(recall)}$$

Nilai *F1-score* pada kelima algoritma yang diteliti, ketiga algoritma (*decision tree*, *random forest*, dan KNN) memiliki skor sebesar 0.8 yang artinya ketiga model cukup mampu menangkap kelas esensial secara benar tanpa terlalu banyak salah prediksi. Disusul algoritma K-Means yang mendapatkan skor sebesar 0,67, dan terakhir algoritma *gaussian naive bayes* sebesar 0,5.

Diakhir terdapat metrik evaluasi ROC AUC dengan perumusan yang kompleks. Dari kelima algoritma yang diteliti, ketiga algoritma (*decision tree*, *random forest*, dan KNN) mendapatkan skor sebesar 0,83. Disusul algoritma *gaussian naive bayes* mendapatkan skor sebesar 0,67 dan K-Means sebesar 0,5.

Hasil Evaluasi Model

Evaluasi dilakukan menggunakan validasi silang Stratified K-Fold dengan k = 5. Metrik yang digunakan meliputi akurasi, *precision*, *recall*, *F1-score*, dan ROC AUC. Pada gambar 2 menunjukkan hasil rata-rata dari validasi silang sebanyak 5 kali pada 5 bagian dataset yang sudah dibagi-bagi serta nilai rata-rata dari hasil keseluruhan metrik evaluasi.

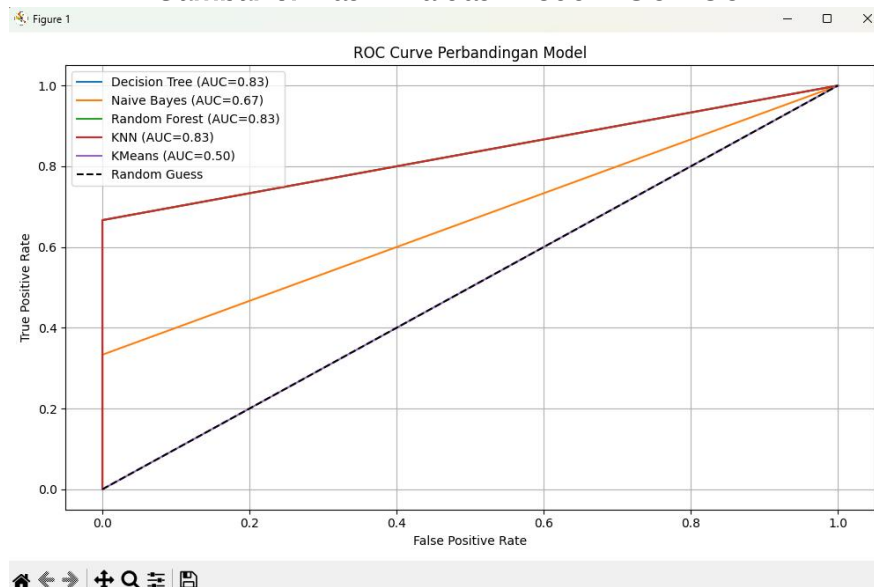
Gambar 2. Hasil Evaluasi Model dengan Validasi Silang 5-fold



Visualisasi ROC AUC Hasil Evaluasi

Visualisasi berikut menampilkan performa kelima model dalam memprediksi klasifikasi pada bentuk grafik ROC dan skor AUC.

Gambar 3. Hasil Evaluasi Model ROC AUC



Berdasarkan gambar 3, algoritma keputusan pohon seperti *decision tree* dan *random forest* serta algoritma lain yakni KNN memiliki skor AUC = 0,83 atau 83%, disusul dengan skor AUC algoritma GaussianNB sebesar 0,67 atau 67%, dan terakhir algoritma K-Means dengan skor AUC = 0.5 atau 50%. Sesuai dengan temuan terdahulu seperti penelitian antara algoritma *decision tree* dan GaussianNB atau *naive bayes*, metode yang memberikan pecahan untuk permasalahan dalam mengidentifikasi penyakit liver ialah Decision Tree (C4.5) [26].

Grafik ROC memiliki garis putus-putus lurus diagonal yang menandakan grafik secara tebak-tebakan (skor AUC 50%), jika grafik algoritmanya diatas garis diagonal ROC maka algoritma tersebut lebih baik dan akurat, namun jika dibawah garis diagonal ROC maka algoritma tersebut kalah dengan hanya tebak-tebakan saja. garis diagonal putus-putus pada ROC mewakili klasifikasi acak, dimana AUC = 0.5 menunjukkan "discriminatory power" minim — artinya model tidak lebih baik daripada tebakan acak [27]. Namun disini mendapatkan satu algoritma dengan tebak-tebakan sama persis hasilnya (skor AUC 50%).

Analisa dan Diskusi

Hasil evaluasi menggunakan *5-fold cross-validation* menunjukkan bahwa algoritma K-Nearest Neighbors (KNN) memperoleh rata-rata skor tertinggi sebesar 0.85, mengungguli algoritma lainnya. Namun, ketika dilakukan evaluasi berdasarkan ROC AUC pada satu kali pembagian data (hold-out test set), terlihat bahwa skor KNN, Decision Tree, dan Random Forest sama-sama memperoleh nilai ROC AUC sebesar 0.83.

Perbedaan ini menunjukkan bahwa pemilihan metrik evaluasi dan metode validasi sangat mempengaruhi hasil performa model. Oleh karena itu, analisis performa tidak cukup hanya dari satu sudut pandang, melainkan perlu menggunakan beragam metrik dan pendekatan evaluasi untuk mendapatkan kesimpulan yang lebih komprehensif.

Menariknya, K-Means sebagai model unsupervised masih menunjukkan kinerja wajar walau tanpa label saat pelatihan, meskipun secara konsisten di bawah model supervised. Hal ini menandakan bahwa fitur-fitur numerik molekul masih cukup memiliki pola alami yang bisa dikelompokkan. Sementara itu, Naive Bayes menunjukkan kelemahan dalam menangani korelasi antar fitur atom, kemungkinan karena asumsi independensi antar fitur yang tidak terpenuhi di data kimia ini.

Tabel 2. Tabel Rata-rata Hasil Evaluasi Penelitian

Model	Accuracy	Precision	Recall	F1-score	ROC AUC	Skor Gabungan Cross 5-Fold	Rata-rata
Decision Tree	0,83	1,00	0,67	0,80	0,83	0,81	0,823
Naive Bayes	0,67	1,00	0,33	0,50	0,67	0,83	0,667
Random Forest	0,83	1,00	0,67	0,80	0,83	0,85	0,830
K-Nearest Neighbour	0,83	1,00	0,67	0,80	0,83	0,61	0,790
K-Means	0,50	0,50	1,00	0,67	0,50	0,81	0,663

Tabel 2 merangkum hasil evaluasi rata-rata dari lima algoritma klasifikasi menggunakan validasi silang 5-fold berdasarkan enam metrik: akurasi, precision, recall, F1-score, ROC AUC, skor gabungan cross-validation, dan nilai rata-rata menyeluruh. Hasil menunjukkan bahwa Random Forest unggul dengan skor tertinggi hampir di semua metrik utama, termasuk skor gabungan 0,85 dan rata-rata keseluruhan 0,830, menjadikannya model paling konsisten dan andal untuk klasifikasi biner asam amino dalam penelitian ini.

Tiga model supervised lainnya, yaitu Decision Tree dan K-Nearest Neighbour (KNN), menunjukkan performa yang sangat mirip. Keduanya memiliki nilai F1-score sebesar 0,80, precision sempurna (1,00), dan skor rata-rata masing-masing 0,823 dan 0,790. Ini menunjukkan bahwa pola numerik sederhana (jumlah atom) cukup informatif untuk digunakan oleh beberapa algoritma supervised learning.

Sementara itu, Naive Bayes menghasilkan precision tertinggi namun memiliki recall rendah (0,33), yang menyebabkan skor F1 dan rata-rata keseluruhannya hanya mencapai 0,667. Model ini cenderung konservatif dan kurang sensitif dalam menangkap data positif (esensial).

Menariknya, K-Means, meskipun merupakan algoritma unsupervised, mampu menghasilkan recall sempurna (1,00) dan skor F1 sebesar 0,67. Namun precision dan akurasinya rendah (0,50), menunjukkan bahwa model ini cenderung melakukan *over-classification* terhadap kelas positif.

Keterbatasan Penelitian

Salah satu keterbatasan utama dalam studi ini adalah ukuran dataset yang kecil, yaitu hanya 20 data asam amino, sehingga hasil evaluasi rentan terhadap fluktuasi. Selain itu, tidak diterapkannya normalisasi pada fitur numerik juga menjadi potensi penyebab turunnya performa pada model berbasis jarak seperti KNN dan K-Means.

Implikasi dan Rekomendasi

Penelitian ini menunjukkan bahwa algoritma supervised seperti *random forest* dan K-Nearest Neighbour dapat digunakan secara efektif untuk klasifikasi sederhana berbasis data kimia. Studi ini juga membuka potensi penerapan *machine learning* sebagai alat edukatif lintas bidang, terutama untuk mahasiswa teknik informatika yang ingin memahami klasifikasi berbasis data ilmiah non-tradisional.

Untuk penelitian lanjutan, disarankan menggunakan dataset yang lebih besar, penerapan teknik *feature engineering* lanjutan, serta evaluasi pada data yang lebih kompleks atau multi-label. Sehingga proses penelitian menghasilkan nilai evaluasi yang lebih akurat.

KESIMPULAN

Penelitian ini telah mengevaluasi lima algoritma *machine learning* Decision Tree, Random Forest, K-Nearest Neighbour (KNN), Gaussian Naive Bayes, dan K-Means dalam mengklasifikasikan asam amino lazim menjadi dua kelas, yaitu esensial dan tidak esensial, berdasarkan representasi numerik dari struktur molekulnya.

Berdasarkan hasil validasi silang 5-fold dan analisis metrik evaluasi (Accuracy, Precision, Recall, F1-Score, dan ROC AUC), diperoleh kesimpulan sebagai berikut:

1. Random Forest menunjukkan performa terbaik secara menyeluruh, dengan rata-rata skor tertinggi dan hasil yang stabil pada seluruh metrik.

2. Decision Tree dan KNN juga memiliki performa baik, terutama pada precision dan F1-score, namun masih lemah dalam recall.
3. Naive Bayes meskipun memiliki precision sempurna, tidak mampu menangkap mayoritas data esensial (recall rendah), sehingga kurang direkomendasikan untuk kasus ini.
4. K-Means sebagai algoritma unsupervised dapat mendeteksi seluruh kelas esensial (recall 1.00), namun menghasilkan precision dan akurasi yang rendah, menandakan banyaknya prediksi palsu.
5. Representasi numerik sederhana dari struktur kimia (jumlah atom) terbukti cukup untuk menghasilkan klasifikasi biner yang signifikan dalam konteks edukatif dan lintas disiplin.

Maka dari itu, model algoritma *supervised learning* sangat cocok untuk digunakan dalam kimia organik ini, khususnya asam amino lazim dalam protein. Keputusan pohon memang baik digunakan sebagai model, namun *random forest* lebih optimal dalam kasus ini. Sementara model ketetanggaan, KNN lah yang menjadi model terbaik dalam kasus ini, dan *gaussian naive bayes* kurang cocok. Uniknya K-Means bisa mendeteksi sangat detail jika terkait kelas positif dan negatif, namun kurang cocok dalam akurasi.

Saran

Berdasarkan hasil penelitian dan keterbatasan yang ada, maka disarankan selanjutnya dapat menggunakan dataset yang lebih besar dan kompleks untuk meningkatkan akurasi model, seperti jumlah dataset lebih dari seratus, kemudian menambahkan fitur biokimia dan fisikokimia lainnya (massa molekul, polaritas, atau titik leleh dan didih). Selain itu, perlunya dilakukan penerapan normalisasi fitur agar model berbasis jarak seperti KNN dan K-Means dapat beroperasi secara lebih optimal. Penelitian ini dapat dijadikan contoh edukatif untuk mahasiswa lintas bidang, khususnya Teknik Informatika, dalam menerapkan machine learning pada data non-konvensional seperti kimia.

REFERENSI

- A. Yani, "Peran artificial intelligence sebagai salah satu faktor dalam menentukan kualitas mahasiswa di era society 5.0," *Journal of Education Research*, no. 2, pp. 1089–1096, 2024, Accessed: Jun. 13, 2025. [Online]. Available: <https://jer.or.id/index.php/jer/article/download/963/547/4263>
- A. Aditya *et al.*, *MACHINE LEARNING*, 1st ed. Sumatera Barat: PT Global Eksekutif Teknologi, 2023. [Online]. Available: www.globaleksekutifteknologi.co.id
- S. T. Jim Mann. A, *Buku ajar ilmu gizi, edisi 4.*, 4th ed. Jakarta: Jakarta : E G C., 2014, 2014.
- A. Then, K. Mácha, B. Ibrahim, and S. Schuster, "A novel method for achieving an optimal classification of the proteinogenic amino acids," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-72174-5.
- S. Berber, M. Brückner, N. Maurer, and J. Huwer, "Artificial intelligence in chemistry research—Implications for teaching and learning," *J Chem Educ*, Apr. 2025, doi: 10.1021/acs.jchemed.4c01033.
- N. Fathiro Cahyono, S. Mukaromah, and K. Uyun, "Etika penggunaan kecerdasan buatan pada teknologi informasi," Surabaya, Sep. 2023. Accessed: Jun. 19, 2025. [Online]. Available: <https://sitasi.upnjatim.ac.id/index.php/sitasi/article/view/334/133>
- H. Byeon, "Is deep learning better than machine learning to predict benign laryngeal disorders?," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, 2021, doi: 10.14569/IJACSA.2021.0120415.
- M. T. Almuqati, F. Sidi, S. N. M. Rum, M. Zolkepli, and I. Ishak, "Challenges in supervised and unsupervised learning: A comprehensive overview," *Int J Adv*

- Sci Eng Inf Technol*, vol. 14, no. 4, pp. 1449–1455, 2024, doi: 10.18517/ijaseit.14.4.20191.
- K. Kristiawan and A. Widjaja, "Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3182.
- D. Hartono Renyut, U. Handayani Makassar, and S. Trinitas Ambon, "Prediksi kelulusan mahasiswa menggunakan algoritma C.45 (studi kasus, sekolah tinggi ilmu administrasi trinitas Ambon)," *Jurnal Sistem Informasi dan Teknik Komputer*, vol. 7, no. 2, 2022, Accessed: Jun. 15, 2025. [Online]. Available: <https://ejournal.catusakti.ac.id/index.php/simtek/article/download/137/143>
- R. Mubarak, M. Hanafi, and D. Sasongko, "Komparasi performa naive bayes gaussian dan k-NN untuk prediksi kelulusan mahasiswa dengan CRISP-DM," *Media Online*, vol. 4, no. 6, pp. 2982–2991, 2024, doi: 10.30865/klik.v4i6.1924.
- J. Perintis Kemerdekaan Km, M. Syukri Mustafa, and I. Wayan Simpen, "Implementasi algoritma K-Nearest Neighbor (KNN) untuk memprediksi pasien terkena penyakit diabetes pada puskesmas manyampa Kabupaten Bulukumba," Aug. 2019. doi: 10.36774/sisiti.v8i1.70.
- R. Tri Vulandari, *Data science (sains data: Teori dan praktek dengan RStudio)*, 1st ed. Yogyakarta: Penerbit Gava Media, 2025.
- I. Mollas, N. Bassiliades, and G. Tsoumakas, "Conclusive local interpretation rules for random forests," Apr. 2021, doi: 10.48550/arXiv.2104.06040.
- C. Ika Agustyaningrum -, R. Dahlia -, and O. Pahlevi -, "Comparison of conventional machine learning and deep neural network algorithms in the prediction of monkey-pox," *JURNAL RISET INFORMATIKA*, vol. 5, no. 3, 2023, doi: 10.34288/jri.v5i2.217.
- D. Program, S. I. Administrasi, S. Satya, and N. Palembang, "Penerapan metode clustering k-means terhadap dosen berdasarkan publikasi jurnal nasional dan terakreditasi sinta implementation of k-means clustering method to lecturers based on publications of national journals and accredited sinta," *JSI : Jurnal Sistem Informasi (E-Journal)*, vol. 16, no. 1, 2024, doi: 10.18495/jsi.v16i1.163.
- Z. Zhang, "Variable selection with stepwise and best subset approaches," *Ann Transl Med*, vol. 4, no. 7, pp. 136–136, Apr. 2016, doi: 10.21037/atm.2016.03.35.
- M. Tajrian, A. Rahman, M. A. Kabir, and M. R. Islam, "Analysis of child development facts and myths using text mining techniques and classification models," *Heliyon*, vol. 10, no. 17, p. e36652, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36652.
- K. Kristiawan and A. Widjaja, "Perbandingan algoritma machine learning dalam menilai sebuah lokasi toko ritel," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3182.
- A. Iyamuremye *et al.*, "Utilization of artificial intelligence and machine learning in chemistry education: a critical review," *Discover Education*, vol. 3, no. 1, Jul. 2024, doi: 10.1007/s44217-024-00197-5.
- Y. Shi, P. L. Prieto, T. Zepel, S. Grunert, and J. E. Hein, "Automated Experimentation Powers Data Science in Chemistry," *Acc Chem Res*, vol. 54, no. 3, pp. 546–555, Feb. 2021, doi: 10.1021/acs.accounts.0c00736.
- R. J. Fessenden and J. S. Fessenden, *Kimia organik edisi 3 jilid 2 / fessenden & fessenden*, 3rd ed., vol. 2. Erlangga, 1999.
- P. Palinggik Allorerung, A. Erna, M. Bagussahrir, and S. Alam, "Analisis performa normalisasi data untuk klasifikasi K-nearest neighbor pada dataset penyakit," 2024. Accessed: Jun. 20, 2025. [Online]. Available: <https://media.neliti.com/media/publications/485639-pengaruh-normalisasi-data-terhadap-perfo-e19e3a00.pdf>
- A. Peryanto, A. Yudhana, and R. Umar, "Klasifikasi citra menggunakan convolutional neural network dan K fold cross validation," 2020. [Online]. Available: <http://jurnal.polibatam.ac.id/index.php/JAIC>

- M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan akurasi, recall, dan presisi klasifikasi pada algoritma C4.5, random forest, SVM dan naive bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021, doi: 10.30865/mib.v5i2.2937.
- N. Tiara Rahman, "Analisa algoritma decision tree dan naïve bayes pada pasien penyakit liver," *Jurnal Fasilkom*, vol. 10, no. 2, pp. 144–151, Aug. 2020, Accessed: Jun. 13, 2025. [Online]. Available: <https://ejurnal.umri.ac.id/index.php/JIK/article/view/2087/1236>
- J. Y. Verbakel *et al.*, "ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models," *J Clin Epidemiol*, vol. 126, pp. 207–216, Oct. 2020, doi: 10.1016/j.jclinepi.2020.01.028.