Implementasi Random Forest untuk Klasifikasi dan Menganalisis Aktivasi AMPK (AMP-activated Protein Kinase) pada Metabolisme dan Penyakit Metabolik

Implementation of Random Forest for Classification and Analysis of AMPK (AMP-activated Protein Kinase) Activation in Metabolism and Metabolic Diseases.

M. Gilang Martiansyah M[1 *], Ghozi Alvin Karim[2 *], Lia Alyani[3 *], Nadilla Andhara Putri[4 *], Anisa Dini Amalia[5 *], M Faqih[6 *],

[1]Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
**E-mail: mgilang.121450056@student.itera.ac.id*

Abstrak

AMP-activated protein kinase (AMPK) adalah enzim kunci dalam pengaturan homeostasis energi seluler dan menjadi target potensial untuk terapi penyakit metabolik seperti diabetes tipe 2 dan obesitas. Penelitian ini bertujuan untuk menganalisis aktivasi AMPK menggunakan algoritma Random Forest pada data bioaktivitas molekul dari database ChEMBL. Dataset yang mencakup 174 molekul ini dianalisis menggunakan metode fingerprint molekuler dan Tanimoto similarity untuk mengevaluasi fitur kimia spesifik yang berhubungan dengan aktivitas biologis. Penelitian ini melibatkan eksplorasi data, seleksi fitur, dan pembangunan model Random Forest untuk klasifikasi molekul menjadi kelas aktif dan tidak aktif berdasarkan nilai pIC50. Hasil menunjukkan bahwa molekul dengan massa molekul (MW) tinggi dan logaritma partisi lipofilik (LogP) lebih cenderung aktif, sedangkan distribusi donor dan akseptor hidrogen memberikan pola unik untuk masing-masing kelas bioaktivitas. Evaluasi model menggunakan mengklasifikasi molekul aktif AMPK menunjukkan AUC sebesar 0.95 dengan fitur penting yang berkontribusi dalam model yaitu MW, PubchemFP568, PubchemFP194, PubchemFP195, pIC50.

Kata kunci: AMPK, Random Forest, Pubchem

Abstract

AMP-activated protein kinase (AMPK) is a key enzyme in the regulation of cellular energy homeostasis and is a potential target for the therapy of metabolic diseases such as type 2 diabetes and obesity. This study aims to analyze AMPK activation using the Random Forest algorithm on molecular bioactivity data from the ChEMBL database. This dataset, which includes 174 molecules, was analyzed using the molecular fingerprint method and Tanimoto similarity to evaluate specific chemical features related to biological activity. This research involves data exploration, feature selection, and building a Random Forest model for classifying molecules into active and inactive classes based on pIC50 values. The results show that molecules with a high molecular mass (MW) and logarithm of lipophilic partition (LogP) are more likely to be active, while the distribution of hydrogen donors and acceptors provides unique patterns for each class of bioactivity. Model evaluation using classifying active AMPK molecules shows an AUC of 0.95 with important features that contribute to the model, namely MW, PubchemFP568, PubchemFP194, PubchemFP195, pIC50.

Keywords: AMPK, Random Forest, Pubchem

PENDAHULUAN

AMP-activated protein kinase (AMPK) adalah enzim yang memainkan peran kunci dalam pengaturan homeostasis energi dalam sel. Aktivasi AMPK memiliki dampak signifikan pada metabolisme energi, termasuk peningkatan oksidasi asam lemak dan pengurangan sintesis lipid, glukosa, serta protein [1]. Dalam konteks penyakit metabolik seperti diabetes tipe 2 dan obesitas, AMPK telah dianggap sebagai target terapeutik potensial karena kemampuannya untuk mengembalikan keseimbangan energi seluler [2]. Oleh karena itu, pemahaman lebih mendalam mengenai aktivasi AMPK diperlukan untuk membuka peluang pengembangan intervensi terapeutik baru.

Teknologi bioinformatika telah menjadi alat penting dalam memahami kompleksitas molekuler, termasuk regulasi AMPK. Salah satu pendekatan yang semakin berkembang adalah penggunaan algoritma machine learning, seperti Random Forest. Random Forest merupakan algoritma berbasis ensemble yang sangat efektif untuk menangani data dengan dimensi tinggi, seperti data genetik dan proteomik [3]. Penelitian sebelumnya menunjukkan bahwa algoritma ini dapat digunakan untuk seleksi fitur dan klasifikasi data biologis dengan akurasi tinggi [4].

Ketika diterapkan pada konteks regulasi AMPK, teknik data mining seperti Random Forest dapat digunakan untuk mengidentifikasi pola regulasi yang sering terjadi. Studi "Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle" menunjukkan bahwa pendekatan ini efektif dalam memahami regulasi AMPK pada otot rangka [5]. Selain itu, penelitian lainnya menunjukkan bagaimana Random Forest dapat digunakan untuk seleksi gen dan klasifikasi data microarray, yang relevan untuk analisis ekspresi gen yang berhubungan dengan AMPK [6]. Dalam bidang bioinformatika, integrasi data biologis dan algoritma pembelajaran mesin seperti Random Forest juga telah diterapkan dalam studi regulasi metabolik untuk meningkatkan pemahaman tentang mekanisme yang mendasari penyakit metabolik [7].

Studi “Integrating biological knowledge and gene expression data using pathway-guided random forests: a benchmarking study” menyoroti bagaimana algoritma ini dapat memanfaatkan jalur biologis untuk meningkatkan akurasi prediksi dalam analisis ekspresi genetik [8]. Selain itu, teknik serupa telah digunakan dalam studi regulasi metabolik pada jaringan adiposa dan hati, dua lokasi utama aktivitas AMPK [9]. Dengan meningkatnya kebutuhan untuk memahami hubungan molekuler dalam penyakit metabolik, penelitian ini menjadi semakin penting.

Bagaimana aktivasi AMPK dapat dianalisis menggunakan algoritma Random Forest? Apa saja pola regulasi AMPK yang dapat diidentifikasi dari data biologis yang tersedia? Sejauh mana efektivitas algoritma Random Forest dalam seleksi fitur dan klasifikasi data ekspresi genetik yang terkait dengan AMPK?

Penelitian ini bertujuan untuk menganalisis aktivasi AMPK dalam konteks metabolisme dan penyakit metabolik menggunakan algoritma Random Forest, mengidentifikasi pola regulasi AMPK berdasarkan data biologis yang relevan, serta mengevaluasi efektivitas algoritma Random Forest dalam seleksi fitur dan klasifikasi data ekspresi genetik yang terkait dengan AMPK.

Dengan rumusan masalah dan tujuan yang telah dibuat, penelitian ini diharapkan bisa memberikan kontribusi dalam bidang bioinformatika dan biologi molekuler dengan cara menyediakan wawasan baru mengenai mekanisme regulasi AMPK yang dapat diterapkan dalam pengembangan terapi penyakit metabolik, memberikan bukti empiris tentang keunggulan algoritma Random Forest dalam analisis data biologis dengan dimensi tinggi, serta menyediakan

metode analisis yang dapat digunakan dalam penelitian molekuler lainnya.

Penelitian terkait AMPK telah berkembang pesat dalam beberapa dekade terakhir. Hardie et al. menjelaskan peran penting AMPK dalam pengaturan metabolisme energi [1], sementara penelitian lain menyoroti potensi AMPK sebagai target terapeutik untuk penyakit metabolik [2]. Selain itu, studi yang lebih teknis seperti "Gene selection and classification of microarray data using random forest" menunjukkan bagaimana algoritma Random Forest dapat digunakan untuk menganalisis data biologis yang kompleks [6].

Di sisi lain, penerapan Random Forest dalam domain non-biologis juga menunjukkan fleksibilitas algoritma ini. Misalnya, penelitian "Implementasi Algoritma Random Forest pada Klasifikasi Dataset Credit Approval" menunjukkan bagaimana algoritma ini digunakan untuk klasifikasi data dalam konteks persetujuan kredit [10]. Studi ini menekankan keunggulan Random Forest dalam menangani data dengan dimensi tinggi dan fitur yang saling terkait. Dalam konteks bioinformatika, keunggulan ini sangat relevan untuk menganalisis data genetik dan proteomik yang sering kali memiliki kompleksitas serupa.

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam memahami aktivasi AMPK dan potensinya sebagai target terapeutik untuk penyakit metabolik, sekaligus menunjukkan efektivitas algoritma Random Forest dalam analisis data biologis yang kompleks.

METODE

Tahapan analisis data bioaktivitas ChEMBL untuk membangun model Random Forest terdiri dari beberapa langkah utama yang dijelaskan pada **Gambar 1.**

Dataset

Data yang digunakan adalah data yang berasal dari database ChEMBL tentang aktivitas biologis terhadap protein "*AMP-activated Protein Kinase*". Dataset ini berisi tentang 174 molekul yang diuji aktivitasnya terhadap protein AMPK. Setiap molekul memiliki berbagai data, seperti massa molekul (MW), tingkat lipofilisitas (LogP), serta jumlah atom yang dapat menyumbang atau menerima ikatan hydrogen. Aktivitas biologis molekul terhadap target diukur menggunakan nilai plC50, yang menunjukkan seberapa efektif molekul tersebut menghambat target. Semakin tinggi nilai plC50, semakin kuat kemampuan molekul menghambat AMPK. Molekul-molekul ini juga diklasifikasikan sebagai aktif atau tidak aktif berdasarkan aktivitasnya, yang direpresentasikan dalam kolom class numeric. Dataset ini juga memiliki 889 kolom fingerprint molekuler, yang menggambarkan fitur kimia spesifik dalam bentuk biner (0 atau 1) yang terlihat pada **Tabel 1**. Data ini berguna untuk penelitian dan analisis di bidang biologi, khususnya dalam hal identifikasi target protein dan pengembangan obat.

Data Preparation

Pada tahapan data preparation dilakukan pembersihan data dengan menghapus baris yang memiliki nilai kosong di kolom standard value dan canonical smiles, sehingga hanya data yang lengkap yang digunakan. Setelah itu, data yang memiliki duplikasi di kolom canonical smiles dihapus agar setiap molekul hanya muncul satu kali. Selanjutnya, hanya tiga kolom penting yang dipilih, yaitu molecule chembl id, canonical smiles, dan standard value, agar data lebih sederhana.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan proses untuk menganalisis dan menyajikan data dengan tujuan mendapatkan pemahaman yang lebih dalam tentang informasi yang terkandung dalam data tersebut. Peran utama Exploratory Data Analysis (EDA) adalah untuk mengeksplorasi data secara menyeluruh dan menggunakan grafik untuk mendukung dan

memperkuat analisis yang dilakukan[11].

Fingerprint Molekuler

Fingerprint molekuler atau lebih dikenal sebagai DNA fingerprinting adalah teknik analisis yang digunakan untuk mengidentifikasi individu berdasarkan pola unik dari DNA mereka. Metode ini memanfaatkan variasi dalam urutan DNA untuk membedakan satu individu dari yang lain. Setiap individu memiliki sekuen DNA yang unik, yang terdiri dari urutan basa nitrogen (adenin, timin, sitosin, dan guanin). Variasi dalam jumlah dan urutan basa ini menciptakan pola yang dapat digunakan sebagai "sidik jari" genetik.

Tanimoto Similarity

Tanimoto similarity adalah ukuran yang digunakan untuk menghitung kesamaan antara dua fingerprint molekuler. Tanimoto merupakan metode yang umum digunakan dalam kimia komputasi dan bioinformatika untuk membandingkan struktur molekul berdasarkan representasi bit dari fitur-fitur yang ada dalam molekul tersebut. Dalam penelitian ini tanimoto similarity digunakan untuk membandingkan dua molekul berdasarkan fingerprint mereka yang dapat membantu dalam penemuan obat.

Random Forest

Random Forest adalah sebuah metode yang dapat meningkatkan akurasi dengan membangkitkan atribut secara acak untuk setiap node. Random Forest terdiri dari sejumlah pohon keputusan (decision trees) yang bekerja secara kolektif untuk mengklasifikasikan data ke dalam kategori tertentu. Setiap pohon keputusan dibangun dengan menetapkan node akar dan berakhir pada sejumlah node daun untuk menghasilkan keputusan akhir[12].

Evaluasi Model

Evaluasi model memiliki peran krusial dalam menilai kinerja model. Dalam penelitian ini, proses evaluasi yang digunakan mengandalkan metrik utama, yaitu akurasi, presisi, recall, dan skor F1. Selain itu, evaluasi model juga dilengkapi dengan analisis menggunakan confusion matrix guna mengidentifikasi pola distribusi kesalahan klasifikasi antar kelas, dapat dilihat pada **Gambar 2**.

Akurasi adalah metrik yang digunakan untuk menunjukkan tingkat ketepatan suatu model prediksi dalam memprediksi kejadian yang benar.

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

Dimana:

TP : Molekul yang benar-benar aktif sesuai prediksi model.

TN : Molekul yang benar-benar tidak aktif, sesuai dengan prediksi model.

FP : Molekul yang tidak aktif tetapi salah diprediksi oleh model sebagai aktif.

FN Molekul yang aktif tetapi salah diprediksi oleh model sebagai tidak aktif.

Presisi mengukur seberapa akurat model dalam memprediksi kelas positif dari seluruh prediksi positif yang dibuat.

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

Recall mengevaluasi kemampuan model dalam secara akurat mengidentifikasi semua sampel yang termasuk dalam kelas positif.

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

F1-score adalah metrik yang menggabungkan presisi dan recall untuk memberikan keseimbangan antara keduanya.

$$F1\ Score = \frac{2 \times (Presisi \times Recall)}{Presisi + Recall}$$

HASIL DAN PEMBAHASAN

Pada penelitian ini dengan menggunakan *Random Forest* untuk mengklasifikasi dan analisis aktivasi *AMPK* (*AMP-Activated Protein Kinase*) berdasarkan data senyawa terkait dengan metabolisme energi dan

penyakit metabolik. Dataset yang digunakan pada penelitian ini adalah hasil dari screening bioactivity *AMPK* yang mencakup beberapa fitur yaitu *MW,LogP,NumHDonors, NumHAcceptors, pIC50* dan *Class Activity* sebagai kolom target yang berisikan status aktivasi senyawa *AMPK Active* dan *Inactive*.

Penelitian ini dimulai dengan analisis eksplorasi data untuk memahami terkait informasi dan distribusi pada dataset, dilihat pada **Gambar 3** Menunjukan pola sebaran jumlah senyawa dalam masing-masing kategori pada bioactivity class *Active* dan *Inactive* berdasarkan nilai *pIC50*. Kategori class tersebut didasarkan pada ambang batas tertentu, untuk pIC50 ≥ 6,5 untuk kelas *Active* dan pIC50 < 6,5 untuk kelas *Inactive*. Pada **Gambar 3** tersebut dengan jumlah senyawa pada kelas *Active* mencerminkan efektivitas screening senyawa potensial untuk mengaktivasi *AMPK*, sedangkan senyawa *Inactive* memberikan informasi penting tentang pola kimia yang kurang relevan, yang dapat digunakan dalam mendesain senyawa baru di masa depan.

Pada **Gambar 4** menunjukkan scatter plot hubungan antara massa molekul (*MW*) dan logaritma partisi (*LogP*), dengan molekul aktif (biru) dan tidak aktif (oranye). Molekul *Active* umumnya terkonsentrasi di wilayah *MW* > 400 dan *LogP* 4–6, sedangkan molekul *Inactive* tersebar lebih luas, termasuk di *MW* < 400. Ukuran simbol merepresentasikan nilai pIC50, di mana molekul dengan aktivitas biologis lebih tinggi cenderung memiliki *MW* dan *LogP* yang lebih besar. Hasil tersebut menunjukkan korelasi antara *MW, LogP*, dan aktivitas biologis molekul. Molekul dengan *MW* dan *LogP* tinggi cenderung lebih *Active*, memberikan panduan penting untuk desain obat, khususnya dalam optimasi kelarutan dan permeabilitas.

Pada **Gambar 5** menunjukkan distribusi nilai *pIC50* berdasarkan kelas bioaktivitas. Molekul *Active* memiliki nilai *pIC50* yang lebih tinggi, dengan rentang antara 5,5 hingga 9 dan median mendekati 7,5. Sebaliknya, molekul tidak aktif memiliki nilai *pIC50* yang lebih rendah, yaitu dalam rentang 4 hingga 5,5 dengan median sekitar 5. Hasil tersebut mengindikasikan bahwa molekul aktif secara signifikan memiliki potensi bioaktivitas yang lebih besar dibandingkan molekul tidak aktif. Adanya perbedaan tersebut menandakan pentingnya nilai *pIC50* sebagai indikator aktivitas biologis molekul.

Pada **Gambar 6** menampilkan distribusi massa molekul (*MW*) berdasarkan kelas bioaktivitas. Molekul *Active* memiliki *MW* yang lebih tinggi, dengan rentang antara 400 hingga 600 dan median mendekati 500. Sebaliknya, molekul *Inactive* menunjukkan MW yang lebih rendah, dengan rentang 200 hingga 500 dan median sekitar 400. Distribusi tersebut menunjukkan bahwa massa molekul yang lebih besar cenderung berhubungan dengan bioaktivitas yang lebih tinggi. Pada hasil *MW* ini dapat digunakan sebagai desain obat, parameter MW dapat digunakan sebagai panduan memprediksi potensi aktivitas molekul.

Pada **Gambar 7** menunjukkan distribusi *LogP* (koefisien partisi lipofilik) untuk molekul dalam kategori *Active* dan *Inactive* . Median *LogP* kelas *Active* sedikit lebih tinggi, dengan distribusi yang lebih terpusat dibandingkan kelas *Inactive*, yang menunjukkan variasi lebih besar dan adanya outlier. Nilai *LogP* yang lebih tinggi berkaitan dengan sifat lipofilik molekul, memengaruhi penetrasi membran dan bioaktivitas. Perbedaan tersebut menunjukkan bahwa molekul *Active* memiliki nilai *LogP* yang lebih konsisten, sementara molekul *Inactive* lebih bervariasi.

Pada **Gambar 8** menampilkan distribusi *NumHDonors* (jumlah donor hidrogen) berdasarkan dua kelas bioaktivitas, yaitu *Active* dan *Inactive*. Dari visualisasi tersebut, terlihat bahwa kelas bioaktivitas *Inactive* memiliki nilai median jumlah donor hidrogen yang lebih tinggi dibandingkan kelas

bioaktivitas *Active*. Distribusi untuk kelas *Inactive* juga menunjukkan variasi yang lebih besar, seperti terlihat dari rentang interkuartil (IQR) dan keberadaan outlier. Sebaliknya, kelas *Active* memiliki nilai yang cenderung lebih terpusat dengan beberapa outlier pada rentang rendah. Hal tersebut menandakan bahwa jumlah donor hidrogen cenderung lebih tinggi pada molekul dengan bioaktivitas *Inactive*, namun perlu diperhatikan bahwa distribusi nilai untuk kelas ini lebih bervariasi.

Pada **Gambar 9** menunjukkan distribusi *NumHAcceptors* (jumlah akseptor hidrogen) untuk kedua kelas bioaktivitas. Kelas bioaktivitas *Active* memiliki median jumlah akseptor hidrogen yang sedikit lebih tinggi dibandingkan kelas tidak aktif, namun distribusinya lebih sempit, yang tercermin dari rentang IQR yang lebih kecil. Sementara itu, kelas *Inactive* menunjukkan variasi yang lebih besar dan memiliki beberapa outlier pada nilai yang lebih tinggi. Perbedaan dalam distribusi ini mengindikasikan bahwa molekul dengan bioaktivitas *Active* cenderung memiliki jumlah akseptor hidrogen yang lebih konsisten dibandingkan molekul dengan bioaktivitas *Inactive.*

Pada **Gambar 10** menunjukkan representasi struktur molekul kimia dengan nomor identifikasi unik, seperti 8, 15, 33, dan seterusnya. Setiap molekul divisualisasikan dengan struktur yang beragam, terdiri atas atom-atom dengan pewarnaan tertentu, seperti atom karbon (hitam), oksigen (merah), nitrogen (biru), dan sulfur (kuning). Struktur molekul yang ditampilkan juga mencerminkan keberagaman jenis ikatan, termasuk ikatan tunggal, rangkap dua, dan percabangan, serta keberadaan gugus fungsi, seperti karbonil dan atom hetero. Variasi ini menunjukkan bahwa dataset yang digunakan memiliki cakupan yang luas dan berpotensi untuk digunakan dalam analisis kimia komputasi. Selain itu, representasi "Fingerprint" pada gambar tersebut menunjukkan karakteristik unik setiap molekul dalam bentuk digital. Fingerprint ini biasanya digunakan untuk analisis lebih lanjut, seperti pencocokan molekul, pengelompokan struktur kimia, atau prediksi sifat fisik, kimia, maupun biologis. Data fingerprint dapat diolah menjadi vektor numerik yang sangat relevan untuk penerapan pembelajaran mesin, terutama dalam konteks prediksi aktivitas biologis atau toksisitas molekul. Keanekaragaman struktur molekul dalam dataset ini menunjukkan potensinya dalam mendukung penelitian, misalnya dalam desain obat berbasis struktur atau eksplorasi senyawa baru dengan aktivitas tertentu. Hasil ini mengindikasikan bahwa dataset molekul yang ditampilkan memiliki nilai yang signifikan dalam pengembangan ilmu kimia dan farmasi, terutama dalam penerapan metode kimia komputasi modern. Pembahasan ini juga menegaskan bahwa penggunaan teknik fingerprint dalam analisis molekul dapat mempermudah proses klasifikasi, pencarian senyawa mirip, atau pengelompokan berdasarkan sifat tertentu. Dengan demikian, dataset ini dapat berkontribusi pada percepatan proses penelitian dan pengembangan di bidang kimia farmasi maupun bidang terkait lainnya.

Pada **Gambar 11** ditampilkan struktur kimia senyawa resveratrol. Resveratrol adalah senyawa polifenol yang dikenal karena memiliki berbagai manfaat farmakologis, termasuk aktivitas antioksidan, antiinflamasi, dan antikanker. Struktur kimianya menunjukkan adanya gugus hidroksil (-OH) pada cincin aromatik, yang berperan penting dalam aktivitas antioksidan melalui kemampuan untuk mendonorkan elektron dan menangkal radikal bebas. Senyawa resveratrol berkaitan dengan AMPK karena kemampuannya untuk mengatur metabolisme energi dan meningkatkan sensitivitas insulin. Resveratrol dapat meningkatkan aktivitas AMPK melalui mekanisme yang melibatkan peningkatan AMP / ATP ratio dalam sel yang biasanya menghambat kompleks enzim tertentu, yang dapat menjadikan resveratrol sebagai kandidat potensial untuk pengobatan penyakit metabolik.

Gambar 12 menunjukkan hasil analisis Tanimoto Similarity untuk menemukan senyawa yang memiliki kemiripan struktur dengan resveratrol. Kemiripan diukur berdasarkan koefisien Tanimoto, dengan nilai tertinggi mencapai 0.220 dan nilai terendah 0.134. Meskipun nilai kemiripan ini tergolong rendah, senyawa-senyawa yang diidentifikasi memiliki beberapa fitur struktural yang sama, seperti cincin aromatik dan gugus hidroksil. Hasil ini menunjukkan bahwa tidak ada senyawa yang sangat mirip dengan resveratrol dalam dataset, namun senyawa dengan nilai Tanimoto lebih tinggi dapat menjadi kandidat awal untuk studi lanjutan terkait sifat bioaktivitasnya.

Gambar 13 menggambarkan hasil pohon keputusan pertama dalam algoritma Random Forest, yang digunakan untuk mengklasifikasikan bioaktivitas molekul. Pada setiap simpul, fitur molekular seperti Polar Surface Area atau LogP digunakan sebagai kriteria pemisahan, dengan nilai ambang tertentu. Hasil klasifikasi menunjukkan bagaimana fitur-fitur tersebut mempengaruhi pemisahan molekul ke dalam kategori aktivitas tertentu. Struktur pohon ini memberikan wawasan tentang parameter utama yang berkontribusi terhadap aktivitas molekul, seperti luas permukaan polar dan sifat lipofilik, yang sering dikaitkan dengan kemampuan interaksi molekul dengan target biologis.

Gambar 14 menunjukkan hasil dari 10 fitur penting yang memiliki kontribusi dalam prediksi menggunakan model random forest, dimana fitur MW (Molecular Weight) memiliki kontribusi terbesar (0.036099) dalam memengaruhi prediksi model, diikuti oleh fitur-fitur PubchemFP568 (0.030960), PubchemFP194 (0.030567), dan PubchemFP195 (0.026850), yang merepresentasikan pola struktur kimia tertentu. Fitur **pIC50** (0.025820), yang sering dikaitkan dengan efektivitas molekul dalam aktivitas biologis, memiliki kontribusi yang sedikit lebih rendah. Ini menunjukkan bahwa berat molekul merupakan faktor utama yang digunakan model dalam menentukan hasil prediksi, sedangkan fitur struktur kimia dan aktivitas biologis tetap relevan meskipun kontribusinya lebih kecil.

Evaluasi model pada **Tabel 2** menunjukan bahwa model Random Forest bekerja sangat baik untuk data ini, dengan tingkat akurasi tinggi dan performa yang solid pada kelas mayoritas ("Inactive"). Namun, performa yang lebih rendah pada kelas minoritas ("Active") tercermin dari recall yang hanya 0.57 . Hal ini menunjukkan bahwa model cenderung kurang sensitif dalam mendeteksi kelas "Active," yang dapat menyebabkan beberapa data penting terlewatkan. Ketidakseimbangan kelas dalam data (141 untuk "Inactive" vs. 33 untuk "Active") kemungkinan menjadi faktor yang mempengaruhi performa ini.

Hasil evaluasi model menggunakan matriks kebingungan (Confusion Matrix) menunjukkan kinerja prediksi model terhadap dua kelas, yaitu "Active" dan "Inactive." Pada **Gambar 15**, model berhasil mengklasifikasikan 27 sampel sebagai "Inactive" secara benar (true negatives) dan 4 sampel sebagai "Active" secara benar (true positives). Namun, terdapat kesalahan klasifikasi pada 3 sampel yang diprediksi sebagai "Inactive" padahal sebenarnya "Active" (false negatives) dan 1 sampel yang diprediksi sebagai "Active" padahal sebenarnya "Inactive" (false positives). Secara keseluruhan, model menunjukkan performa yang baik dengan tingkat kesalahan yang minimal.

Kurva ROC pada **Gambar 16** memperlihatkan kinerja diskriminasi model yang diukur melalui area under the curve (AUC). Dengan nilai AUC sebesar 0,95, model memiliki kemampuan yang sangat baik dalam membedakan antara kelas "Active" dan "Inactive." Nilai AUC yang mendekati 1 mengindikasikan tingkat prediksi yang tinggi, di mana model mampu meminimalkan tingkat false positive rate

sambil mempertahankan true positive rate yang tinggi.

KESIMPULAN

Penelitian ini menunjukkan bahwa fitur molekuler, seperti massa molekul *MW*, logaritma partisi *LogP*, jumlah donor hidrogen (*NumHDonors*), dan jumlah akseptor hidrogen (*NumHAcceptors*), memiliki hubungan yang erat dengan bioaktivitas molekul. Pola regulasi AMPK diidentifikasi melalui distribusi fitur, di mana molekul *Active* cenderung memiliki *MW* lebih dari 400, *LogP* 4–6, dan nilai *pIC50* lebih dari 6,5. Analisis *Tanimoto similarity* terhadap senyawa yang memiliki kemiripan dengan *Resveratrol* menunjukkan bahwa dataset hanya mencakup senyawa dengan nilai kemiripan tertinggi 0,220 dan terendah 0,134. Algoritma Random Forest menunjukkan kemampuan yang sangat baik dalam mengklasifikasikan data bioaktivitas dengan akurasi tinggi (AUC 0,95) dimana fitur MW (Molecular Weight) memiliki kontribusi terbesar (0.036099) dalam memengaruhi prediksi model, diikuti oleh fitur-fitur PubchemFP568 (0.030960), PubchemFP194 (0.030567), dan PubchemFP195 (0.026850).

UCAPAN TERIMA KASIH

Segala puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas segala limpahan rahmat dan karunia-Nya sehingga penulisan tugas besar ini dapat diselesaikan dengan baik. Penulis menyampaikan rasa hormat dan terima kasih yang sebesar-besarnya kepada:

1. **Tirta Setiawan, S.Pd., M.Si.**, selaku dosen pembimbing sekaligus pengampu mata kuliah Bioinformatika, atas bimbingan, arahan, dan dukungan yang telah diberikan selama proses pengerjaan tugas besar ini. Dedikasi dan ilmunya telah memberikan inspirasi dan wawasan yang sangat berharga bagi penulis.
2. Rekan-rekan, yaitu **M. Gilang Martiansya, M Ghozi Alvin Karim, Lia Alyani, Nadilla Andhara Putri, Anisa Dini Amalia, M. Faqih**, atas kerja sama, semangat, dan dukungan yang telah diberikan selama pengerjaan tugas ini. Kontribusi dan kolaborasi kalian sangat berarti dalam menyelesaikan proyek ini dengan hasil yang maksimal.

Semoga tugas besar ini dapat memberikan manfaat dan menjadi kontribusi positif bagi perkembangan ilmu pengetahuan di bidang bioinformatika.

DAFTAR RUJUKAN

[1] D. G. Hardie, F. A. Ross, and S. A. Hawley, "AMPK: A nutrient and energy sensor that maintains energy homeostasis," Nat. Rev. Mol. Cell Biol., vol. 13, no. 4, pp. 251–262, 2012.

[2] C. Canto and J. Auwerx, "AMPK as a target for metabolic therapeutic," Cell Metab., vol. 14, no. 3, pp. 242–255, 2011.

[3] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[4] A. Statnikov, C. F. Aliferis, L. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," Bioinformatics, vol. 21, no. 5, pp. 631–643, 2005.

[5] BMC Bioinformatics, "Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle," 2015.

[6] BMC Bioinformatics, "Gene selection and classification of microarray data using random forest," 2012.

[7] Academic OUP, "Data mining in the life sciences with Random Forest: A walk in the park or lost in the jungle?," Brief. Bioinform., vol. 14, no. 3, pp. 315–326, 2013.

[8] Academic OUP, "Integrating biological knowledge and gene expression data using pathway-guided random forests: A benchmarking study," Bioinformatics, vol. 36, no. 15, pp. 4301–4308, 2020.

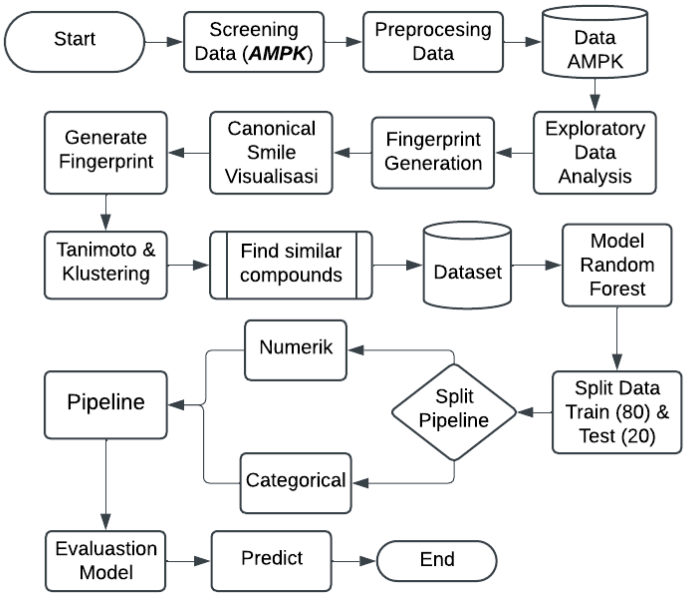
[9] J. J. Smith et al., "The role of AMPK in metabolic regulation of adipose tissue," J. Mol. Biol., vol. 430, no. 12, pp. 1949–1961, 2018.

[10] Janitra, "Implementasi Algoritme Random Forest pada Klasifikasi Dataset Credit Approval," 2022.

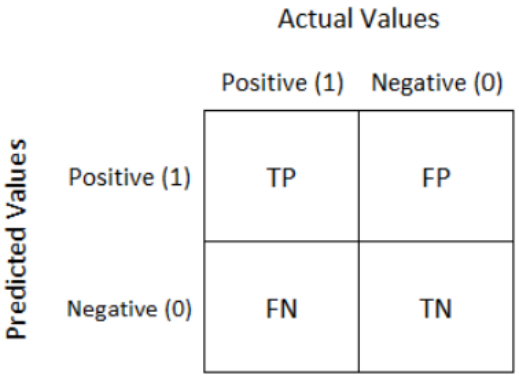
[11] Radhi, M., Sitompul, D. R. H., Sinurat, S. H., & Indra, E. (2021, Februari). ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS(EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK. *JUSIKOM PRIMA (Jurnal Sistem InformasidanIlmu Komputer Prima)*, *4*.

[12] Amaliah, S., Nusrang, M., & Aswi. (n.d.). Penerapan Metode Random ForestUntuk Klasifikasi Varian Minuman Kopi Di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and ResearchISSN 2684-7590 (Online)*, *4*, 121-127.

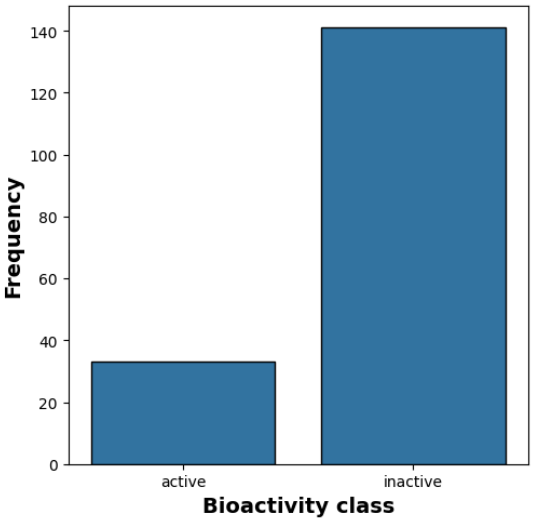
LAMPIRAN GAMBAR



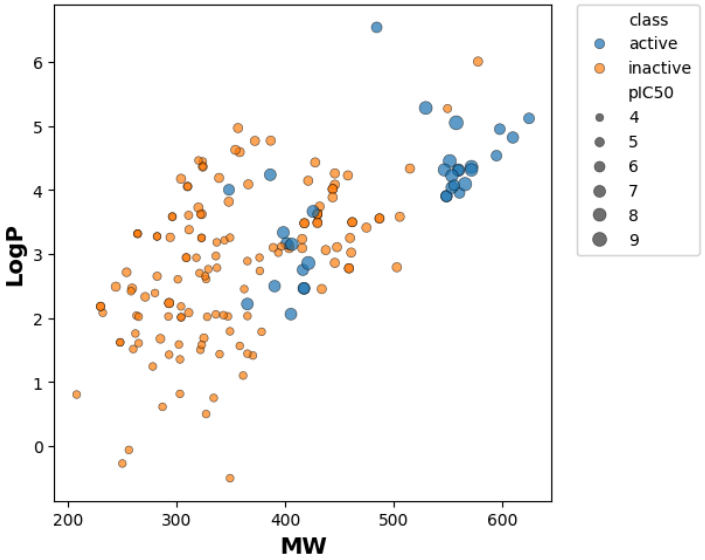
Gambar 1. Flowchart Penelitian



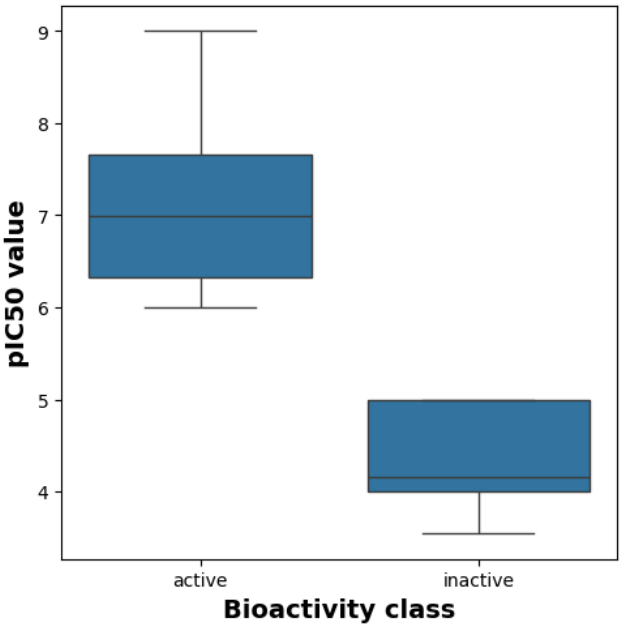
Gambar 2. *Confusion Matriks*



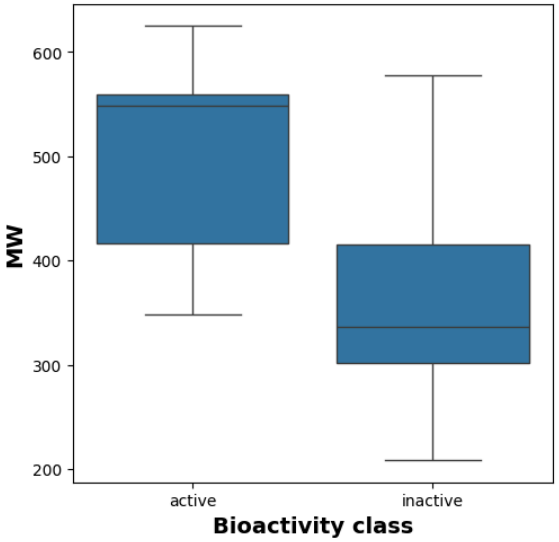
Gambar 3. Distribusi AMPK Bioactivity Class



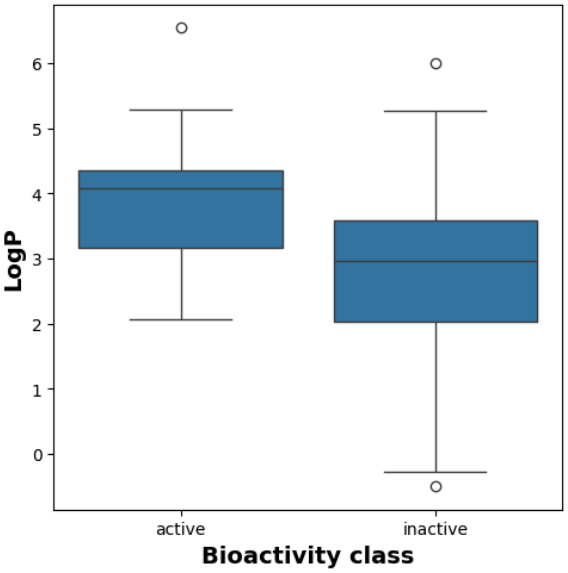
Gambar 4. Distribusi Sebaran Scatter Plot MW vs LogP



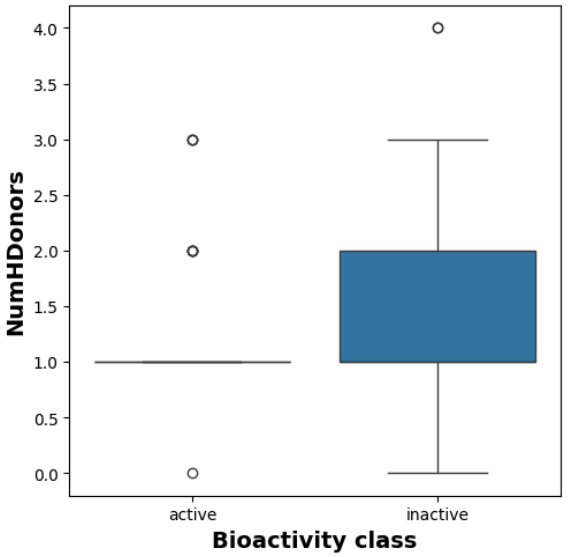
Gambar 5. Distribusi Boxplot pIC50



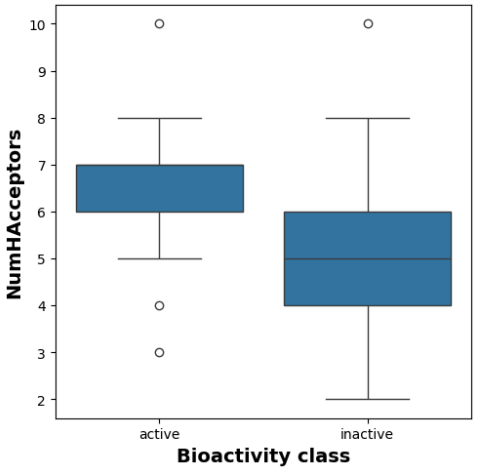
Gambar 6. Distribusi Boxplot MW



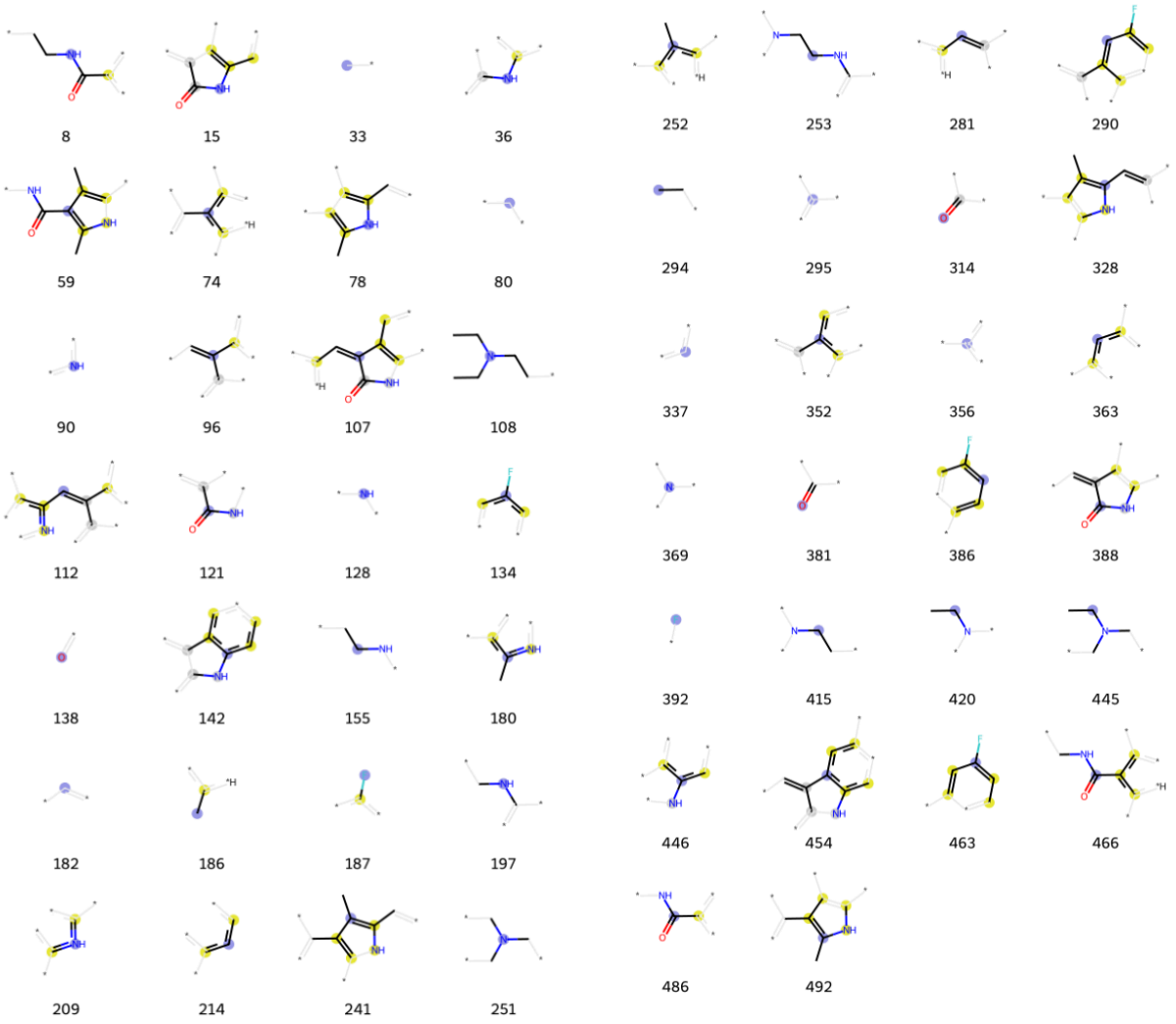
Gambar 7. Distribusi Boxplot LogP



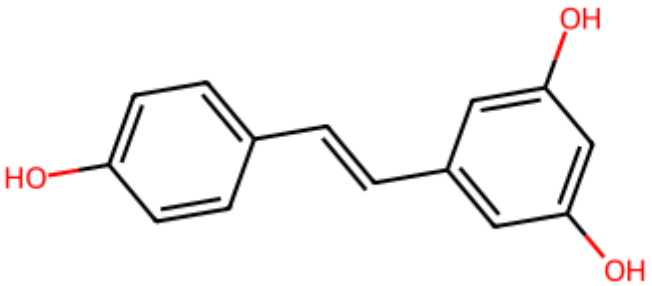
Gambar 8. Distribusi Boxplot NumHDonors



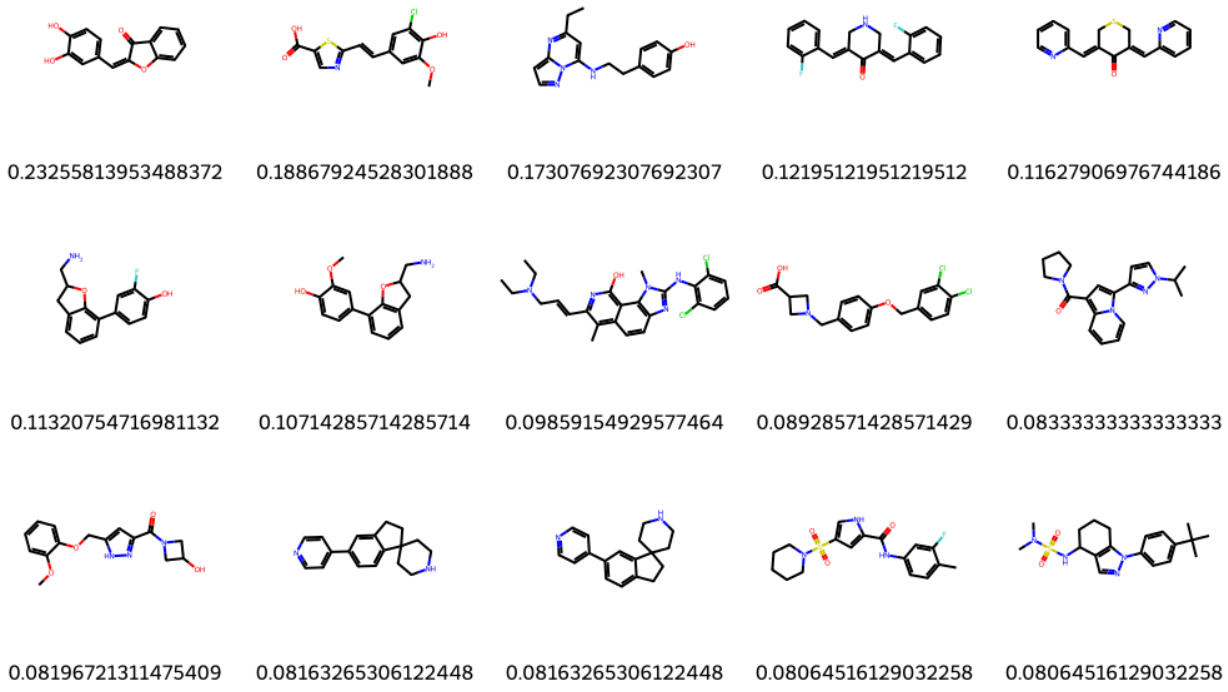
Gambar 9. Distribusi Boxplot NumHAcceptros



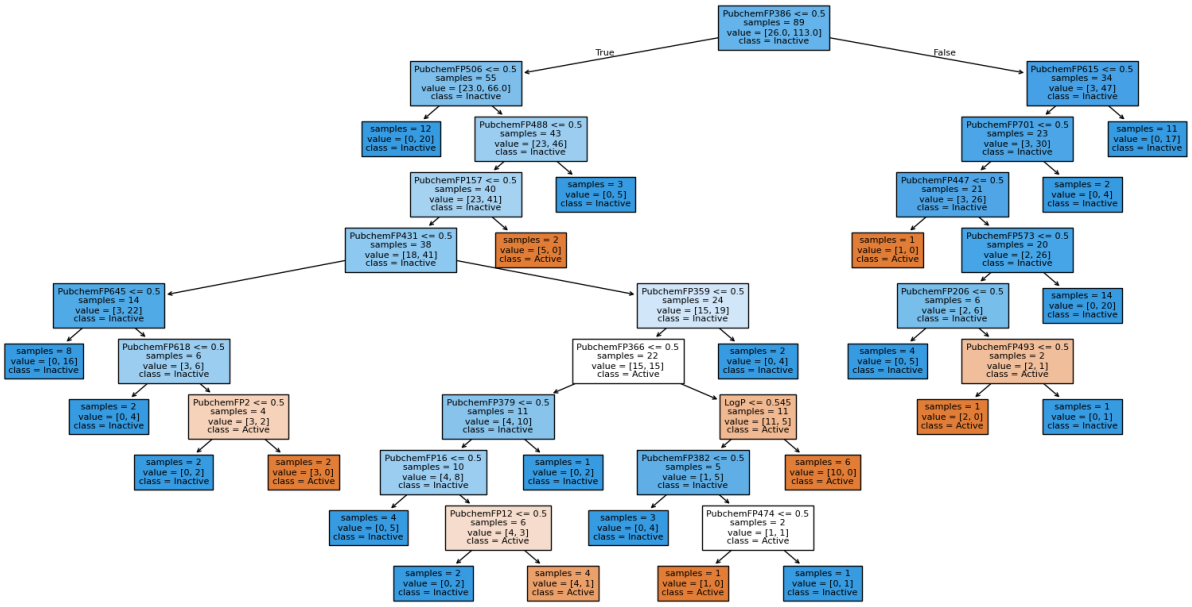
Gambar 10. Fingerprint



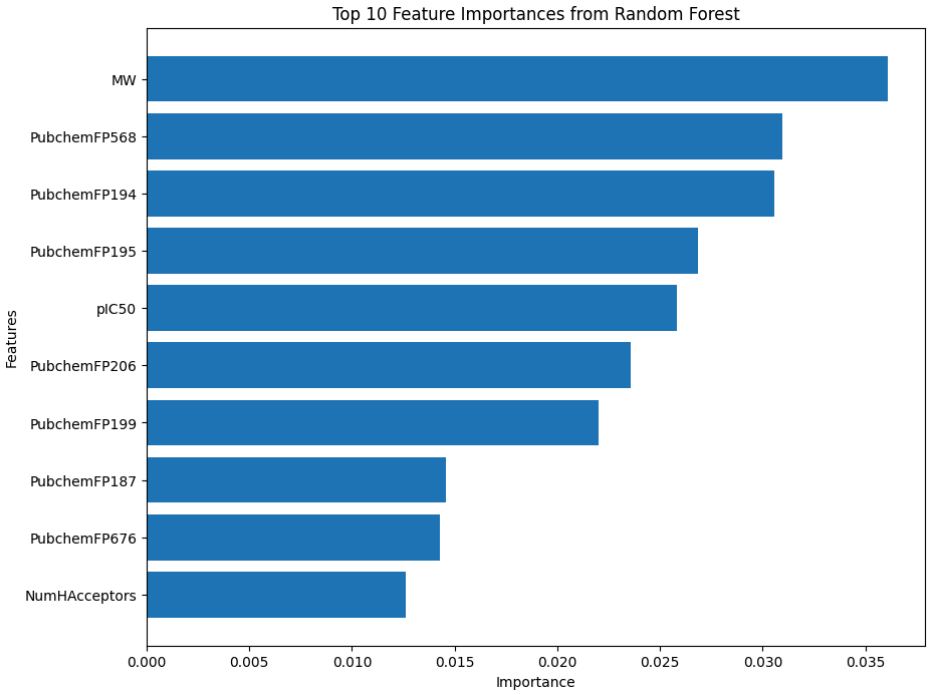
Gambar 11 . Resveratrol



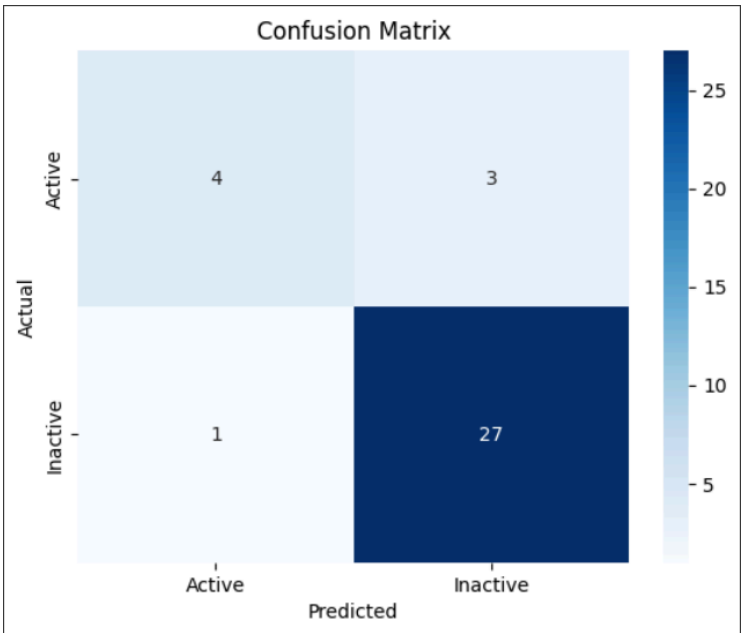
Gambar 12. Hasil Tanimoto senyawa yang mirip dengan Resveratrol



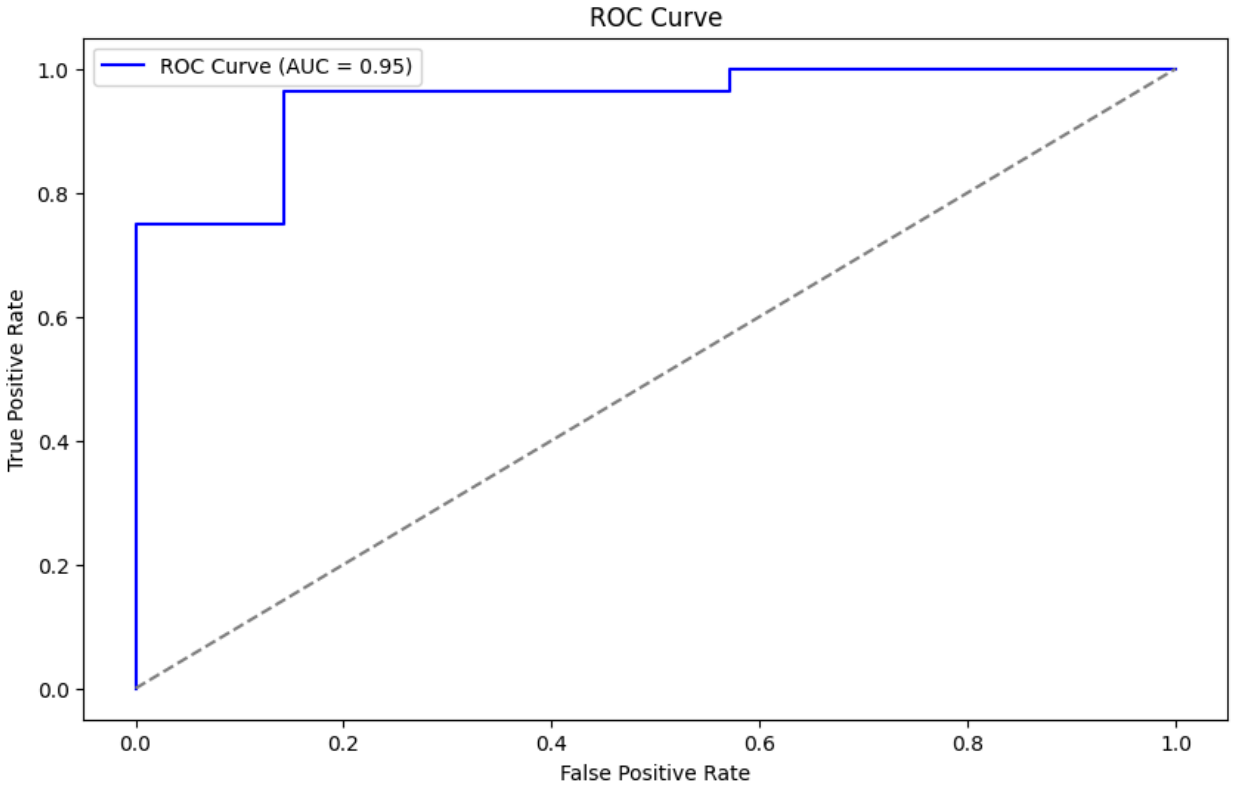
Gambar 13. Hasil Pohon Keputusan Pertama Pada RandomForest



Gambar 14. Fitur Penting yang memiliki kontribusi hasil prediksi model



Gambar 15. Hasil Evaluasi Confussion Matrix Model RandomForest



Gambar 16. Hasil Kurva ROC Model RandomForest

LAMPIRAN TABEL

Tabel 1. 5 Dataset teratas

No	molecul_chembl id	Canonical Smile	MW	LogP	NumHDonors	NumHAcceptors	pIC50	Class
1	CHEMBL535	CCN(CC)CCNC(=O)c1c(C)[nH]c(/C=C2\C(=O)Nc3ccc(F...	398.48	3.33	3	3	7.20	active

2	CHEMBL2819577	CCN(CC)C/C=C/c1nc(O)c2c(ccc3nc(Nc4c(Cl)cccc4Cl...	484.43	6.54	2	6	6.00	active
3	CHEMBL4378885	COc1cc2ncnc(Nc3ccc(Cl)c3F)c2cc1CN1CCCC1	386.85	4.77	1	5	5.00	inactive
4	CHEMBL2078155	COCCN1CC(C(N)=O)(N(C)Cc2cc3c(Nc4ccc(Cl)c4F)nc...	502.97	2.79	2	8	5.00	inactive
5	CHEMBL4256061	COc1cc2ncnc(Nc3ccc(Cl)c3F)c2cc1CN(C)C1(C(N)=O...	486.97	3.55	2	7	5.00	inactive

Tabel 2. 5 Dataset teratas

No	MW	LogP	NumHDonors	NumHAcceptors	pIC50	Class_Numeric	PubchemFP0	…	PubchemFP877
1	398.48	3.33	3	3	7.20	0	1	…	0
2	484.43	6.54	2	6	6.00	0	1	…	0
3	386.85	4.77	1	5	5.00	1	1	…	0
4	502.97	2.79	2	8	5.00	1	1	…	0
5	486.97	3.55	2	7	5.00	1	1	…	0

Tabel 3. ***Classification Report***

Parameter	Presisi	Recall	F1-Score	Support
Active	0.80	0.57	0.67	7
Inactive	0.90	0.96	0.93	28
Accuracy			0.89	35
Macro avg	0.85	0.77	0.80	35
Weighted avg	0.88	0.89	0.88	35

Lampiran Project

---

1. Code

Screening: Klik di sini untuk membuka code screening

Model: Klik di sini untuk membuka code Model

---

2. PPT

Presentasi: Klik di sini untuk membuka PPT

---

3. Poster

Poster Proyek: Klik di sini untuk membuka Poster

4. Video

Video Proyek : Klik di sini untuk membuka Video

5. Github

Repository Github: Klik di sini untuk membuka Repo Github