



Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PENGEMBANGAN MODEL PREDIKSI SENYAWA ANTIMALARIA MENGGUNAKAN PENDEKATAN MACHINE LEARNING

FATHA ARIYA PRASETYA



**PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**

@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Pengembangan Model Prediksi Senyawa Antimalaria menggunakan pendekatan *Machine Learning*” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2025

Fatha Ariya Prasetya
G6401211078



ABSTRAK

FATHA ARIYA PRASETYA. Pengembangan Model Prediksi Senyawa Antimalaria Menggunakan Pendekatan *Machine Learning*. Dibimbing oleh WISNU ANANTA KUSUMA dan ARYO TEDJO.

Malaria merupakan salah satu penyakit paling mematikan di kawasan tropis dengan tingkat mortalitas yang tinggi akibat infeksi parasit *Plasmodium falciparum*. Tantangan utama dalam pengendalian malaria adalah munculnya resistensi terhadap obat-obatan antimalaria yang ada. Studi ini bertujuan untuk mengembangkan metode penapisan maya (*virtual screening*) senyawa herbal antimalaria berbasis *machine learning* untuk mengidentifikasi kandidat senyawa potensial. Studi ini melakukan perbandingan terhadap tiga algoritma *machine learning*, yaitu Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), dan Light Gradient Boosting Machine (LGBM). Model model tersebut dilatih menggunakan kombinasi deskriptor molekuler, yakni PubChem Fingerprint dan ECFP. Untuk meningkatkan kinerja prediktif, dilakukan pemilihan fitur menggunakan metode Lasso. Hasil terbaik diperoleh dari model XGB dengan penerapan seleksi Lasso dengan nilai R^2 sebesar 0.5599, RMSE sebesar 0.3694, dan MAE sebesar 0.2800. Nilai-nilai tersebut mengindikasikan potensi penggunaannya dalam menyaring senyawa-senyawa herbal untuk tahap awal penemuan obat antimalaria.

Kata kunci: LGBM, malaria, *Plasmodium falciparum*, SVM, XGB.

ABSTRACT

FATHA ARIYA PRASETYA. Developing a Machine Learning Based Predictive Model for Antimalarial Compounds. Supervised by WISNU ANANTA KUSUMA and ARYO TEDJO.

Malaria is one of the deadliest diseases in tropical regions, with a high mortality rate caused by infection with the *Plasmodium falciparum* parasite. The main challenge in malaria control is the emergence of resistance to existing antimalarial drugs. This study aims to develop a virtual screening method for antimalarial herbal compounds using machine learning to identify potential candidate compounds. The study compares Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM) as machine learning algorithms. These models were trained using a combination of molecular descriptors, specifically PubChem Fingerprint and ECFP. To improve predictive performance, feature *selection* was carried out using the Lasso method. The best results were obtained from the XGB model with Lasso *selection*, achieving an R^2 of 0.5599, RMSE of 0.3694, and MAE of 0.2800. These values indicate its potential use in screening herbal compounds during the early stages of antimalarial drug discovery.

Keywords: LGBM, malaria, *Plasmodium falciparum*, SVM, XGB.



@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PENGEMBANGAN MODEL PREDIKSI SENYAWA ANTIMALARIA MENGGUNAKAN PENDEKATAN MACHINE LEARNING

FATHA ARIYA PRASETYA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Ilmu Komputer

**PROGRAM SARJANA ILMU KOMPUTER
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



IPB University

Tim Pengaji pada Ujian Skripsi:

1 Medria Kusuma Dewi H., S.Komp., Ph.D.

@Hak cipta milik IPB University



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengujip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Skripsi : Pengembangan Model Prediksi Senyawa Antimalaria
Menggunakan Pendekatan *Machine Learning*
Nama : Fatha Ariya Prasetya
NIM : G6401211078

Disetujui oleh



Pembimbing 1:

Prof. Dr. Eng. Wisnu Ananta Kusuma S.T., M.T.
19711110 200501 1 000



Pembimbing 2:

Aryo Tedjo, S.Si., M.Si.
19750202 200812 1 001



Diketahui oleh

Ketua Program Sarjana Ilmu Komputer:
Dr. Sony Hartono Wijaya, S.Kom., M.Kom.
19810809 200812 1 002



Tanggal Ujian:
7 Agustus 2025

Tanggal Lulus:



Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Judul yang dipilih dalam penelitian yang dilaksanakan sejak bulan Desember 2024 sampai bulan Juni 2025 ini ialah “Pengembangan Model Prediksi Senyawa Antimalaria Menggunakan Pendekatan *Machine Learning*”.

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada para pembimbing, Prof. Dr. Eng. Wisnu Ananta Kusuma, S.T., M.T. dan Aryo Tedjo, S.Si., M.Si., atas bimbingan, saran, serta dukungan yang luar biasa. Ucapan terima kasih juga disampaikan kepada pembimbing akademik, moderator seminar, dan penguji luar komisi pembimbing yang telah memberikan kontribusi penting dalam penyelesaian tugas akhir ini. Penulis juga mengucapkan terima kasih kepada ayah, ibu, dan seluruh keluarga yang telah memberikan dukungan, doa, serta kasih sayang yang tiada henti. Terima kasih juga untuk Feby, Fadhil, Shilla, Diba, Najla, Thoriq, Lava, Nuril, Ayyas, serta teman-teman lainnya yang tidak dapat disebutkan satu per satu atas dukungan dan kebersamaannya.

Penulis juga menyadari bahwa dalam proses penyusunan tugas akhir ini, masih banyak kekurangan dan keterbatasan baik dalam hal penyusunan materi maupun penyampaian informasi. Oleh karena itu, penulis memohon maaf atas segala kekurangan atau kekeliruan yang mungkin ada dan berharap karya ilmiah ini bermanfaat bagi pihak yang membutuhkan serta bagi kemajuan ilmu pengetahuan.

Bogor, Agustus 2025

Fatha Ariya Prasetya



DAFTAR TABEL	xii
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xii
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
1.5 Ruang Lingkup	4
II TINJAUAN PUSTAKA	5
2.1 Senyawa Herbal dan Malaria	5
2.2 Half-maximal Inhibitory Concentration (IC50)	5
2.3 Deskriptor Molekuler	6
2.4 Extended Conectivity Fingerprint (ECFP)	6
2.5 Pubchem Fingerprint	8
2.6 Lasso	8
2.7 Support Vector Machine (SVM)	9
2.8 Light Gradient Boosting Machine (LGBM)	10
2.9 Extreme Gradient Boosting (XGB)	11
III METODE	13
3.1 Data Penelitian	13
3.2 Peralatan Penelitian	13
3.3 Metode	13
3.3.1 Pra-proses Data	13
3.3.2 Perhitungan Deskriptor	14
3.3.3 Seleksi Fitur	16
3.3.4 Pembagian Data	17
3.3.5 Pembangunan Model	17
3.3.6 Evaluasi dan Perbandingan	18
IV HASIL DAN PEMBAHASAN	21
4.1 Pengumpulan Data	21
4.2 Pembersihan Data	23
4.3 Transformasi Data	25
4.4 Ekstraksi Fitur	27
4.5 Seleksi Fitur (<i>Feature Selection</i>)	27
4.6 Pembagian Data	29
4.7 <i>Hyperparameter tuning</i>	29
4.8 Pelatihan Model dan Evaluasi	30
4.9 Prediksi Senyawa Herbal	38
V SIMPULAN DAN SARAN	39
5.1 Simpulan	39

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



5.2 Saran	39
DAFTAR PUSTAKA	40
LAMPIRAN	44
RIWAYAT HIDUP	47

©Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

- Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- Pengutipan tidak merugikan kepentingan yang wajar IPB University.



DAFTAR TABEL

1	Klasifikasi aktivitas nilai IC50 (Indrayanto <i>et al.</i> 2021)	5
2	Representasi bit pada deskriptor Pubchem Fingerprint	8
3	Contoh hasil ekstraksi senyawa menggunakan deskriptor Pubchem Fingerprint	15
4	Contoh data hasil ekstraksi deskriptor ECFP	15
5	Kombinasi parameter untuk hyperparameter tuning SVM	17
6	Kombinasi parameter untuk hyperparameter tuning XGB	17
7	Kombinasi parameter untuk hyperparameter tuning LGBM	17
8	Contoh data senyawa yang telah diuji dengan parasit Plasmodium falciparum	22
9	Jumlah data sebelum dan setelah dibersihkan	23
10	Lima senyawa dengan nilai IC50 terendah	25
11	Contoh data hasil transformasi	26
12	Contoh hasil ekstraksi deskriptor Pubchem Fingerprint	27
13	Contoh hasil deskriptor ECFP	27
14	Hasil pemilihan nilai alpha terbaik berdasarkan proses hyperparameter tuning	28
15	Perbandingan jumlah fitur awal dan setelah seleksi	28
16	Daftar 10 fitur dengan skor importance tertinggi pada deskriptor ECFP	28
17	Daftar 10 fitur dengan skor importance tertinggi pada deskriptor Pubchem Fingerprint	29
18	Perbandingan jumlah data latih dan data uji	29
19	Hasil hyperparameter tuning pada SVM	30
20	Hasil hyperparameter tuning pada XGB	30
21	Hasil hyperparameter tuning pada LGBM	30
22	Metrik dengan Confidence Interval pada algoritma XGB dengan deskriptor ECFP	31
23	Metrik dengan Confidence Interval pada algoritma XGB dengan deskriptor Pubchem Fingerprint	31
24	Metrik dengan Confidence Interval pada algoritma SVM dengan deskriptor Pubchem Fingerprint	32
25	Metrik dengan Confidence Interval pada algoritma SVM dengan deskriptor ECFP	32
26	Metrik dengan Confidence Interval pada algoritma LGBM dengan deskriptor Pubchem Fingerprint	33
27	Metrik dengan Confidence Interval pada algoritma LGBM dengan deskriptor ECFP	33
28	Hasil evaluasi model ECFP	34
29	Hasil evaluasi model Pubchem Fingerprint	35
30	Contoh data aktual dan prediksi model terbaik	36
31	Hasil evaluasi pengelompokan data dengan deskriptor ECFP	37
32	Hasil evaluasi pengelompokan data dengan deskriptor Pubchem Fingerprint	37
33	Hasil uji dengan sampel senyawa herbal yang diperoleh melalui basis data Pubchem	38



1	Diagram skematik proses vektorisasi ECFP melalui empat teknik <i>substructure-pooling</i> (Dablander <i>et al.</i> 2024)	7
2	Diagram skematik model LGBM : (A) Struktur pohon, (B) contoh algoritma konsep pertumbuhan pohon berbasis daun (<i>leaf wise tree</i>), (C) algoritma Gradient Boosting Decision Tree (Gan <i>et al.</i> 2021)	10
3	Metode penelitian	13
4	Diagram proses ekstraksi deskriptor PubChem Fingerprint	14
5	Diagram proses ekstraksi deskriptor ECFP	15
6	Diagram proses seleksi fitur	16
7	Distribusi tipe standar dalam dataset	21
8	Distribusi standar unit dalam dataset	22
9	Boxplot sebaran nilai IC ₅₀	24
10	Boxplot sebaran nilai IC ₅₀ setelah proses pembersihan outlier	24
11	Histogram distribusi nilai IC ₅₀	25
12	Histogram distribusi nilai IC ₅₀ setelah transfromasi	26
13	Histogram perbandingan model dengan deskriptor ECFP	35
14	Histogram perbandingan model dengan deskriptor Pubchem Fingerprint	36

DAFTAR LAMPIRAN

1	Daftar 30 fitur terbaik pada deskriptor Pubchem Fingerprint	45
2	Daftar 30 fitur terbaik pada deskriptor ECFP	45

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

I PENDAHULUAN

1.1 Latar Belakang

Malaria adalah penyakit paling umum di Afrika dan beberapa negara di Asia dengan jumlah kasus endemis tertinggi. Tingkat kematian malaria secara global berkisar dari 0,3–2,2% dan sebanyak 11–30% kasus dengan gejala yang parah malaria terdapat di daerah dengan iklim tropis. Berbagai studi menunjukkan bahwa prevalensi infeksi parasit malaria telah meningkat sejak tahun 2015 (Talapko *et al.* 2019). Terdapat dua jenis makhluk yang memiliki andil dalam penyebaran malaria, yaitu nyamuk *Anopheles* dan *Plasmodium* (parasit malaria) (Fitriany dan Sabiq 2018). Parasit malaria yang dapat menginfeksi manusia diketahui memiliki enam spesies, yaitu *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale wallickeri*, *Plasmodium ovale curtisi*, *Plasmodium malariae*, dan *Plasmodium knowlesi* (Milner 2018). Kasus kematian tertinggi dari penyakit malaria disebabkan oleh infeksi protozoa intraseluler *Plasmodium falciparum* (Danishuddin *et al.* 2019).

Plasmodium falciparum merupakan virus yang memiliki siklus hidup yang kompleks, melibatkan transmisi wajib melalui vektor nyamuk, dan replikasi aseksual dalam eritrosit dari inang manusia (Abdi *et al.* 2023). Infeksi *Plasmodium falciparum* menyebabkan kerusakan sel darah merah dan produksi yang kurang, sehingga menyebabkan penurunan kadar hemoglobin (Andronescu *et al.* 2023). Menurut penjelasan oleh Vianetha Prima Snak *et al.* (2023) bahwa malaria dengan komplikasi berat dapat menyebabkan asidosis metabolik, anemia berat, hipoglikemia, gagal ginjal akut, atau edema paru akut. Pasien yang dicurigai menderita malaria berat dapat segera diobati berdasarkan hasil Tes Diagnostik Cepat (RDT) malaria. Parasit malaria yang terus berevolusi sehingga menimbulkan resistensi terhadap obat antimalaria menyebabkan munculnya kembali malaria secara global dan menjadi ancaman besar terhadap pengendalian malaria (Shibeshi *et al.* 2020). Namun, proses pengembangan obat tradisional mulai dari identifikasi target hingga persetujuan regulatori memiliki beberapa batasan, yaitu memakan waktu lama (biasanya 10–15 tahun), biaya yang tinggi (hampir melebihi 1 miliar USD), serta memiliki tingkat kegagalan yang tinggi (Villar-delfino *et al.* 2025). Oleh karena itu, diperlukan suatu alat untuk membantu identifikasi senyawa potensial yang dapat digunakan sebagai obat antimalaria.

Machine learning merupakan penerapan komputer dan algoritma matematika yang menggunakan pembelajaran dari data untuk menghasilkan prediksi di masa depan. Proses pembelajaran ini melibatkan dua tahap utama, yaitu latihan (*training*) dan pengujian (*testing*) yang bertujuan untuk memperoleh kecerdasan komputer (Roihan *et al.* 2020). Penerapan *machine learning* pada bidang farmasi terus dikembangkan dan telah memfasilitasi penelitian pada berbagai bidang terkait. Penggunaan teknologi ini bertujuan untuk mengurangi ketergantungan terhadap pengujian makhluk hidup seperti hewan (Elbadawi *et al.* 2021).

Penapisan maya (*virtual screening*) adalah salah satu metode untuk mengidentifikasi senyawa yang memiliki peluang tinggi dalam mengikat protein target tertentu (Masliyana *et al.* 2018). Oliveira *et al.* (2023) juga menjelaskan bahwa *virtual screening* merupakan sebuah teknik *in silico* yang digunakan pada



proses penemuan obat. Struktur molekul yang ada pada basis data besar akan dievaluasi menggunakan metode komputasi yang diharapkan mampu mengidentifikasi molekul yang lebih rentan berikatan dengan target molekuler, seperti protein atau reseptor enzim. Model prediktif berbasis *machine learning* dapat digunakan dalam metode ini sebagai perluasan *chemical libraries*, penemuan deskriptor molekuler baru, dan pencarian kemiripan struktur (Oliveira *et al.* 2023).

Ukuran yang paling banyak digunakan dan informatif sebagai penggambaran seberapa efektif suatu obat untuk menghasilkan respons terapeutik pada target biologisnya (efikasi) adalah IC₅₀ (Aykul dan Hackert 2016). Peningkatan efisiensi efikasi dalam desain dan pengembangan obat dapat menggunakan konsep *ligand efficiency* (LE), *binding efficiency index* (BEI), *ligand lipophilicity* (LLE), dan *ligand-efficiency-dependent lipophilicity* (LELP). Secara berturut-turut, parameter-parameter ini dievaluasi dengan memperhatikan ukuran molekuler (jumlah atom non-hidrogen), berat molekuler, lipofilisitas, dan kombinasi ukuran molekuler dan lipofilisitas. Secara umum, nilai yang lebih tinggi dari indeks-indeks ini menunjukkan peningkatan efisiensi (Bammingher *et al.* 2023).

Mswahili *et al.* (2021) melakukan penelitian terkait klasifikasi menggunakan *machine learning* pada data senyawa untuk kategori aktif dan non-aktif. Data yang digunakan merupakan senyawa yang telah diuji secara langsung terhadap parasit *Plasmodium*. Hasil akhir menunjukkan bahwa model XGB mengungguli model lainnya seperti Artificial Neural Network (ANN) dan Logistic Regression dengan nilai akurasi rata – rata 83%. Kore *et al.* (2025) juga melakukan penelitian serupa untuk data senyawa yang telah diuji pada parasit penyebab malaria dengan menggunakan algoritma Random Forest dan 9 deskriptor molekuler. Deskriptor Avalon Molecular Fingerprints (Avalon MFPs) memiliki hasil terbaik dengan nilai akurasi sebesar 97.3%, *precision* sebesar 93.5%, sensitivitas sebesar 88.4%, dan AUROC sebesar 97.3% untuk data uji.

Penelitian yang dilakukan oleh Mswahili *et al.* (2021) dan Kore *et al.* (2025) berfokus pada klasifikasi senyawa berdasarkan kategori nilai IC₅₀, bukan pada prediksi nilai IC₅₀ secara kuantitatif. Padahal, nilai IC₅₀ memiliki signifikansi farmakologis yang tinggi karena merepresentasikan konsentrasi senyawa yang diperlukan untuk menghambat 50% aktivitas parasit (Aykul dan Hackert 2016). Parameter ini berperan penting dalam penentuan dosis awal yang optimal pada tahap pra-klinis pengembangan obat. Prediksi nilai IC₅₀ yang akurat dapat memberikan landasan ilmiah untuk meminimalkan risiko toksisitas sekaligus memaksimalkan efektivitas terapeutik (Badwan *et al.* 2023). Berdasarkan celah penelitian (*research gap*) tersebut, studi ini diinisiasi untuk mengembangkan model prediktif yang mampu mengestimasi nilai IC₅₀ secara langsung, sehingga dapat menyediakan dasar kuantitatif yang lebih kuat bagi penentuan dosis dalam proses penemuan obat antimalaria.

Pembuatan model *machine learning* pada penelitian ini akan membandingkan algoritma Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), dan LGBM dengan kombinasi deskriptor PubChem Fingerprint (PubChem FP) dan ECFP. Mswahili *et al.* (2021) melakukan penelitian mengenai perbandingan algoritma *machine learning* guna memprediksi nilai IC₅₀ pada senyawa untuk obat antimalaria. Penelitian tersebut menunjukkan bahwa algoritma XGB memiliki kemampuan paling baik dibandingkan algoritma lainnya seperti Artificial Neural Network (ANN) dan Logistic Regression (LR) dengan nilai recall sebesar 81% dan



skor F1 sebesar 83%. Studi literatur yang dilakukan oleh Naveed dan Husnain (2025) mengenai sistem rekomendasi obat menunjukkan bahwa algoritma Light Gradient Boosting Machine (LGBM) menghasilkan nilai Mean Absolute Percentage Error (MAPE) terendah dibandingkan dengan algoritma lainnya, yakni sebesar 0.0993. Temuan ini menjadi salah satu pertimbangan utama dalam pemilihan LGBM sebagai algoritma dasar dalam pengembangan model pada penelitian ini. Pemilihan algoritma Support Vector Machine (SVM) didasari penelitian oleh (Xue *et al.* 2024) yang menunjukkan bahwa algoritma tersebut mampu menghasilkan nilai evaluasi Mean Absolute Error (MAE) sebesar 0.0928. Nilai ini mencerminkan tingkat kesalahan prediksi yang sangat rendah, sehingga mengindikasikan kinerja model yang baik dalam melakukan estimasi.

Ketiga algoritma ini akan dikombinasikan dengan beberapa deskriptor molekuler, yaitu PubChem Fingerprint dan ECFP. Pemilihan deskriptor Pubchem FP berdasarkan penelitian oleh Kumar *et al.* (2024) yang menunjukkan bahwa model dengan deskriptor ini memiliki nilai Mean Average Error (MAE) yang rendah sebesar 0.1645. Kemudian, penggunaan deskriptor ECFP didasari penelitian oleh Díaz-Eufracio dan Medina-Franco (2022) yang menjelaskan bahwa nilai akurasi rata-rata untuk ketujuh model dengan deskriptor ini tergolong tinggi, sebesar 95.14%. Hal ini menunjukkan deskriptor ini cukup baik dalam membangun model prediksi karena hubungan struktur dan aktivitas yang lebih konsisten. Oleh karena itu, perbandingan ketiga algoritma dengan dua deskriptor ini diharapkan dapat memberikan hasil yang optimal dan mendukung tujuan penelitian secara komprehensif.

1.2 Rumusan Masalah

Permasalahan yang dihadapi adalah kebutuhan akan pengembangan senyawa-senyawa yang mampu menekan pertumbuhan parasit malaria dalam tubuh manusia. Namun, tantangan utama yang dihadapi adalah biaya yang tinggi, ketergantungan pada metode pengujian yang melibatkan organisme hidup, serta resistensi parasit malaria terhadap obat-obatan yang sudah ada akibat evolusi. Oleh karena itu, penelitian ini akan dilakukan pembuatan model *machine learning* menggunakan algoritma LGBM, XGB, dan SVM dengan deskriptor PubChem Fingerprint dan ECFP untuk memprediksi senyawa-senyawa yang berpotensi sebagai obat antimalaria.

1.3 Tujuan

Berdasarkan rumusan masalah tersebut, tujuan yang ingin dicapai pada penelitian ini adalah memprediksi senyawa yang berpotensi sebagai antimalaria dilihat dari nilai IC₅₀ menggunakan model *machine learning*, mengevaluasi hasil prediksi senyawa menggunakan data latih dan data uji untuk optimasi akurasi model, serta menentukan model terbaik dari perbandingan algoritma *machine learning* dan deskriptor molekuler.

1.4 Manfaat

Beberapa manfaat yang diharapkan dari penelitian ini yaitu memberikan pengetahuan terkait teknik pencarian senyawa menggunakan pendekatan *machine*



4

learning serta informasi dan rekomendasi untuk bidang kesehatan terkait identifikasi senyawa yang paling berpotensi sebagai obat anti malaria dengan mengukur nilai IC₅₀.

1.5 Ruang Lingkup

Penelitian yang ideal bagi kasus ini sebaiknya dilengkapi dengan validasi *in silico*, seperti penambatan molekuler dan simulasi dinamika molekuler. Namun, penelitian ini hanya berfokus pada pemodelan tanpa melibatkan validasi *in silico* tersebut.

II TINJAUAN PUSTAKA

2.1 Senyawa Herbal dan Malaria

Peningkatan resistensi antimalaria, terutama di Asia Tenggara, akibat penggunaan antimalaria yang tidak terkendali (Ribeiro *et al.* 2023). Ada dua strategi utama untuk pengelolaan dan pengendalian infeksi malaria, yaitu mengendalikan vektor dan pengelolaan kasus terinfeksi. Pengelolaan kasus terinfeksi pada dasaranya bergantung pada obat atau kombinasi antimalaria. Menurut Pandey *et al.* (2023), kulit kayu Cinchona merupakan antimalaria efektif pertama yang digunakan pada abad ke-17. Bahan aktif pada tumbuhan ini, yaitu Kuinina, menjadi pengobatan referensi untuk demam intermiten dan menjadi obat penting untuk malaria.

2.2 Half-maximal Inhibitory Concentration (IC₅₀)

Half-maximal Inhibitory Concentration (IC₅₀) merupakan parameter yang menunjukkan konsentrasi agen farmakologis tertentu yang diperlukan untuk menghambat suatu aktivitas biologis hingga mencapai 50% dari tingkat aktivitas awalnya (Aykul dan Hackert 2016). Penggunaan IC₅₀ dalam uji viabilitas sel memberikan ukuran kuantitatif yang memungkinkan peneliti membandingkan efektivitas berbagai senyawa serta menentukan pilihan yang tepat dalam pengembangan senyawa atau agen terapeutik (Sánchez-Díez *et al.* 2025).

Menurut Indrayanto *et al.* (2021), senyawa dengan nilai IC₅₀ < 1 μM digolongkan memiliki aktivitas sangat baik atau poten yang menunjukkan bahwa konsentrasi sangat rendah sudah mampu menghasilkan efek biologis yang signifikan terhadap target uji. Nilai IC₅₀ antara 1–20 μM dikategorikan sebagai aktivitas baik atau memiliki sitotoksitas sangat kuat, menandakan efektivitas yang tinggi namun memerlukan konsentrasi sedikit lebih besar dibanding kategori poten. Senyawa dengan nilai IC₅₀ antara 20–100 μM tergolong memiliki aktivitas sedang, yang berarti efek biologis masih terlihat namun potensi relatif lebih rendah dibanding dua kategori sebelumnya. Pada rentang IC₅₀ 100–200 μM, senyawa dikategorikan memiliki aktivitas rendah, yang menunjukkan bahwa diperlukan konsentrasi lebih tinggi untuk mencapai efek yang diinginkan. Senyawa dengan nilai IC₅₀ > 200 μM dianggap tidak aktif, karena konsentrasi yang dibutuhkan untuk menghasilkan efek biologis terlalu besar dan umumnya tidak praktis atau tidak relevan secara farmakologis.

Tabel 1 Klasifikasi aktivitas nilai IC₅₀ (Indrayanto *et al.* 2021)

Konsentrasi IC ₅₀	Kriteria
< 1 μM	Aktivitas sangat baik atau poten
1 – 20 μM	Aktivitas baik atau sitotoksitas sangat kuat
20 – 100 μM	Aktivitas sedang
100 – 200 μM	Aktivitas rendah
> 200 μM	Tidak aktif



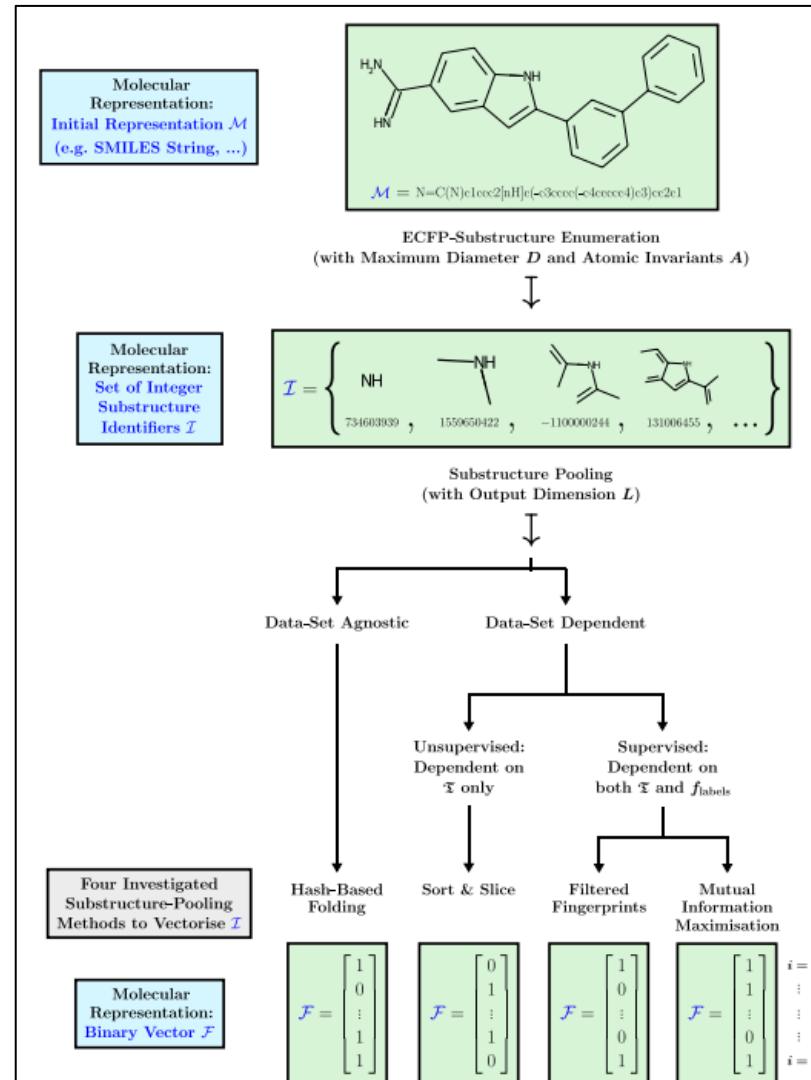
2.3 Deskriptor Molekuler

Deskriptor molekuler merupakan hasil akhir dari prosedur logis dan matematis yang mentransformasikan informasi kimia yang terkandung dalam representasi simbolis suatu molekul menjadi angka yang berguna atau hasil dari beberapa eksperimen standar (Moriwaki *et al.* 2018). *Molecular fingerprint* merupakan profil properti suatu molekul yang berbentuk vektor bit dengan elemen vektornya yang menunjukkan keberadaan atau frekuensi properti tertentu. Deskriptor molekuler dan *molecular fingerprint* memerlukan peran penting dalam analisis *Quantitative structure-activity relationship* (QSAR) dan *Structure-activity relationships* (SAR), penyaringan molekul virtual, pencarian senyawa berbasis kemiripan, identifikasi target biomolekuler, dan analisis jaringan interaksi protein-ligand (Dong *et al.* 2015).

Menurut Dong *et al.* (2015) juga, perhitungan deskriptor molekuler dan *fingerprint* dapat menggunakan bantuan alat seperti BlueDesc, PaDEL, Mold2, ChemAxon, Chemopy, dan sebagainya. Beberapa *software* untuk desain obat generik seperti MOE, SYBYL-X, dan Discovery Studio juga menyediakan fungsionalitas pada perhitungan deskriptor. Namun, alat-alat tersebut hanya mencakup subset deskriptor molekuler dan *molecular fingerprint* sehingga perlu dikombinasikan dengan beberapa alat untuk mendapatkan hasil yang komprehensif.

2.4 Extended Connectivity Fingerprint (ECFP)

Extended Connectivity Fingerprints (ECFPs) merupakan jenis *fingerprint* molekul berbasis atom yang berbentuk melingkar (*atom-centric circular fingerprints*). Setiap atom memiliki lingkungan atom (*atom environment*) yang menjadi komponen fitur dalam ECFP. Masing-masing fitur direpresentasikan oleh sebuah nilai dan himpunan nilai tersebut membentuk struktur ECFP secara keseluruhan (Asahara dan Miyao 2022).



Gambar 1 Diagram skematik proses vektorisasi ECFP melalui empat teknik *substructure-pooling* (Dablander *et al.* 2024)

Menurut Dablander *et al.* (2024), proses vektorisasi Extended Connectivity Fingerprints (ECFP) dimulai dari representasi awal molekul (M), misalnya dalam bentuk SMILES. Representasi ini kemudian diubah menjadi sekumpulan substruktur melalui proses enumerasi ECFP dengan mempertimbangkan diameter ikatan maksimum dan invarian atom. Setiap substruktur yang dihasilkan direpresentasikan sebagai integer *identifier* unik (I) menggunakan proses *hashing*. Selanjutnya, himpunan *identifier* ini diproses melalui tahap *substructure pooling* untuk menghasilkan representasi vektor berdimensi tetap (L). Tahap ini dapat dilakukan dengan pendekatan dataset agnostic seperti *hash-based folding* yang tidak bergantung pada data pelatihan atau pendekatan dataset *dependent* yang mempertimbangkan karakteristik data. Pendekatan dataset *dependent* terbagi menjadi metode *unsupervised* yang hanya bergantung pada I seperti *sort & slice* dan metode *supervised* yang bergantung pada I dan label data (f_{labels}) seperti *filtered fingerprints* serta *mutual information maximization*. Hasil akhir dari proses ini adalah representasi molekul dalam bentuk *binary vector* (F) yang siap digunakan untuk analisis lanjutan atau pemodelan *machine learning*.



2.5 Pubchem Fingerprint

PubChem fingerprint adalah jenis *molecular fingerprint* biner yang merepresentasikan keberadaan atau ketiadaan sub-struktur kimia tertentu pada molekul. Sidik jari ini mengkodekan informasi fragmen molekul dengan dimensi 881 bit (Fernández-De Gortari *et al.* 2017). Setiap bit mengindikasikan fitur struktural spesifik yang terdeteksi, seperti cincin aromatik, gugus fungsi, atau pola atom.

Tabel 2 Representasi bit pada deskriptor Pubchem Fingerprint

Posisi bit	Representasi bit
0	≥ 4 atom hidrogen
1	≥ 8 atom hidrogen
2	≥ 16 atom hidrogen
...	...
877	Memiliki pola NC1C(Br)CCC1
878	Memiliki pola ClC1C(Cl)CCC1
879	Memiliki pola ClC1C(Br)CCC1
880	Memiliki pola BrC1C(Br)CCC1

2.6 Lasso

LASSO (Least Absolute Shrinkage and *Selection Operator*) merupakan teknik yang diinisiasi oleh Tibshirani pada tahun 1996. Metode ini digunakan untuk estimasi parameter regresi dan seleksi variabel secara bersama-sama (Muthukrishnan dan Rohini 2017). Lasso juga termasuk dalam keluarga regresi *penalized least squares* (regresi dengan pinalti), yang dimana menggunakan pinalti L1. Berdasarkan publikasi oleh Muthukrishnan dan Rohini (2017), estimasi koefisien LASSO dapat didefinisikan dengan meminimasi fungsi (1) berikut.

$$\hat{\beta}^{lasso} = \arg \min \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \dots \dots \dots (1)$$

Variabel y_i mewakili nilai target untuk sampel ke-i, x_{ij} mewakili nilai fitur ke-j untuk sampel ke-i, β_j mewakili koefisien regresi yang ingin dipelajari, dan λ mewakili parameter pinalti yang mengontrol seberapa kuat penyusutan koefisien. Persoalan ini juga bisa dituliskan sebagai formula (2)

$$\min \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \dots \dots \dots (2)$$

dengan syarat:

$$\sum_{j=1}^p |\beta_j| \leq t$$

LASSO bekerja dengan menambahkan penalti berupa jumlah nilai absolut dari koefisien regresi (L1 penalty). Semakin besar nilai λ , semakin besar penalti yang dikenakan terhadap besarnya koefisien. Hal ini mendorong beberapa koefisien β_j untuk menjadi tepat sama dengan nol. Koefisien yang bernilai nol menunjukkan



bawa fitur tersebut tidak dipakai dalam model, sehingga secara otomatis LASSO melakukan seleksi variabel.

2.7 Support Vector Machine (SVM)

Support Vector Machines (SVM) adalah algoritma pembelajaran terawasi yang digunakan untuk klasifikasi dan regresi. SVM bekerja dengan menggambar batas antara kelompok data melalui pesawat keputusan, yang memisahkan objek dengan keanggotaan kelas berbeda. SVM membangun *hyperplane* dalam ruang multidimensi untuk memisahkan data dengan label kelas yang berbeda, serta dapat menangani variabel kontinu dan kategorikal (Wang *et al.* 2018).

Menurut Babatunde, Surajudeen A Olaniyan dan Owolabi (2025), data pelatihan untuk SVM dapat direpresentasikan oleh pasangan (x_i, y_i) , di mana $i = 1, \dots, m$, dengan x adalah input berdimensi l , yaitu $x \in R^l$ dan $y \in R$ untuk m data pelatihan dan satu deskriptor input. Artikel ini juga menjelaskan formulasi mengenai SVM dengan fokus pada regresi sesuai dengan fungsi (3).

$$f(x) = \langle \omega, x \rangle + b \quad \dots\dots\dots(3)$$

Dimana $f(x)$ merupakan output dari SVM linear, $\langle \omega, x \rangle$ merupakan produk titik antara ω dan x , dan b adalah konstanta sebagai faktor pengondisian. Fungsi kerugian ϵ -insentif yang diajukan oleh Vapnik untuk meminimalkan risiko empiris didefinisikan sebagai persamaan (4) berikut.

$$\epsilon(y, x) = \begin{cases} 0, & \text{jika } |f(x) - y| \leq \epsilon \\ |f(x) - y| - \epsilon, & \text{lainnya} \end{cases} \quad \dots\dots\dots(4)$$

Risiko empiris yang harus diminimalkan dapat dituliskan sebagai persamaan (5).

$$emp(f) = \frac{1}{m} \sum_{i=1}^m L_\epsilon(y_i, f(x_i)) \quad \dots\dots\dots(5)$$

Dimana $emp(f)$ mendekati risiko empiris seiring dengan bertambahnya jumlah nilai m . Masalah optimisasi dapat didefinisikan pada formula (6) dan solusinya menghasilkan parameter yang optimal.

$$\min \left(\frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m (\epsilon_i^+ + \epsilon_i^-) \right) \quad \dots\dots\dots(6)$$

Dengan kendala :

$$\begin{aligned} y_i - \langle \omega, x_i \rangle - b &\leq \epsilon + \epsilon_i^+ \\ \langle \omega, x_i \rangle + b - y_i &\leq \epsilon + \epsilon_i^- \\ \epsilon_i^-, \epsilon_i^+ &\geq 0 \end{aligned}$$

C adalah faktor regulasi yang mengontrol keseimbangan antara kompleksitas model dan penalti yang diberikan untuk kesalahan yang melampaui zona yang ditentukan oleh fungsi kerugian insensitif.

Untuk data non-linear, kernel seperti RBF digunakan untuk menghitung kedekatan antara dua titik data x dan x' (Ungureanu 2025). Kernel digunakan untuk

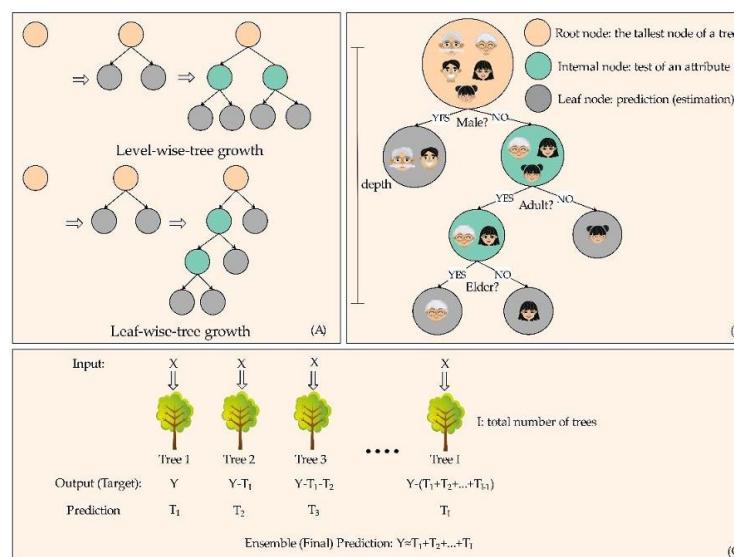
menggantikan perhitungan langsung dalam ruang fitur asli dengan menghitung kedekatan antara data yang dipetakan ke ruang fitur yang lebih tinggi. Menurut Ungureanu (2025), kernel RBF dirumuskan sesuai dengan persamaan (7) berikut.

Variabel γ mewakili gamma yang mengatur kecepatan penurunan nilai kernel seiring dengan peningkatan jarak antara dua titik data. Kemudian fungsi regresi dengan kernel tersebut dirumuskan sesuai dengan persamaan (8)

dengan α_i adalah koefisien yang diperoleh setelah melatih model SVM dan berfungsi untuk mewakili kontribusi masing-masing titik data x_i terhadap prediksi model $f(x)$.

2.8 Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine (LGBM) merupakan model pembelajaran *ensemble* yang berbasis pada algoritma pohon keputusan (*decision tree*). Model ini merupakan implementasi yang lebih efisien dibandingkan model Gradient Boosting Decision Tree (GDBT) karena LGBM mengadopsi algoritma GOSS (Gradient-based One-side Sampling) untuk mengurangi jumlah fitur dan algoritma EFB (Exclusive Feature Bundling) untuk meningkatkan algoritma histogram dalam pemrosesan fitur (Li *et al.* 2023).



Gambar 2 Diagram skematis model LGBM : **(A)** Struktur pohon, **(B)** contoh algoritma konsep pertumbuhan pohon berbasis daun (*leaf-wise-tree*), **(C)** algoritma Gradient Boosting Decision Tree (Gan *et al.* 2021)

Menurut William dan Altamimi (2024), fungsi prediksi $F(x)$ dibangun dalam bentuk bertahap (*stage-wise*) yang berarti untuk setiap langkah m , model menggunakan pohon keputusan lemah $fm(x)$ yang di akumulasikan untuk meningkatkan prediksi yang lebih baik sesuai dengan persamaan (9) berikut.



Dimana $fm(x)$ merupakan pohon keputusan lemah (*weak learner*) pada langkah ke – m dan M merupakan jumlah total pohon dalam model. Pada model ini, tujuan objektif untuk tugas regresi adalah mengurangi fungsi kerugian. Salah satu fungsi kerugian yang sering digunakan adalah Mean Square Error (MSE), dimana mengukut selisih antara nilai yang diprediksi oleh model dan nilai sebenarnya. Fungsi ini dirumuskan pada persamaan (10) berikut.

Nilai y_i mewakili data nilai sebenarnya dan $F(x_i)$ merupakan nilai prediksi untuk data ke- i .

Pada model LGBM, proses pembelajaran dilakukan dengan menggunakan algoritma Gradient Descent. William dan Altamimi (2024) juga menjelaskan bahwa pada setiap iterasi, model menambahkan pohon keputusan baru $fm(x)$ untuk mengurangi kesalahan residu (selisih antara nilai prediksi dengan nilai sebenarnya). Pembaruan model dilakukan dengan persamaan (11) berikut.

Pada formula ini, γ_m merupakan laju pembelajaran (*learning rate*), $F_{m-1}(x)$ adalah model yang dihasilkan pada iterasi sebelumnya, dan $f_m(x)$ adalah pohon keputusan yang baru ditambahkan. Pembaruan untuk pohon keputusan yang baru adalah dengan meminimalkan gradien dari fungsi kerugian, yang dihitung berdasarkan turunan parsial dari fungsi kerugian terhadap prediksi model.

Fungsi (12) diatas menunjukkan bahwa residi (kesalahan) pada data ke - i dihitung dengan turunan dari fungsi kerugian, dan pohon keputusan baru $f_m(x)$ difitkan untuk meminimalkan gradien dari fungsi kerugian.

Pada pustaka scikit-learn dari Python, parameter *number of leaves*, *maximum depth*, dan *n estimator* secara berturut turut mewakili jumlah daun, kedalaman maksimum pohon, dan jumlah pohon. Kemudian, parameter *subsample* dan *colsample bytree* berfungsi untuk mengontrol proporsi data pelatihan dan proporsi fitur yang akan digunakan untuk membangun setiap pohon.

2.9 Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting (XGB) adalah algoritma Gradient Boosting yang diadaptasi untuk meningkatkan komputasi, skalabilitas, dan performa generalisasi dibandingkan dengan metode Gradient Boosting tradisional. XGB dirancang untuk lebih efisien dalam hal pengolahan data besar dan lebih baik dalam menangani masalah *overfitting* (Shahri *et al.* 2021). Pada algoritma ini, semua data kategorikal harus diubah menjadi bentuk numerik, karena XGB hanya bekerja dengan tipe data

tersebut. Salah satu cara mengubah data kategorikal menjadi numerik adalah dengan menggunakan One-Hot Encoding. Shahri *et al.* (2021) menjelaskan bahwa model dibangun menggunakan fungsi umum yang menggambarkan estimasi prediksi pada setiap langkah t , sesuai pada formula (12) berikut.

Model ini bertujuan untuk mengurangi overfitting yang dicapai dengan menambahkan komponen regulasi dalam fungsi objektif untuk mengevaluasi model. Fungsi objektif dalam XGB menggabungkan dua komponen utama, yaitu fungsi kerugian yang mengukur kesalahan prediksi dan term regulasi yang membantu mengurangi kompleksitas model. Fungsi kerugian $L(\theta)$ mengukur perbedaan antara prediksi \hat{y}_t dan nilai target y_i . Sedangkan term regulasi $\Omega(\theta)$ memberikan penalti untuk kompleksitas model yang terlalu besar. Fungsi objektif XGB secara umum dapat dituliskan sebagai persamaan (13) berikut.

Dengan fungsi term regulasi untuk jumlah daun T dan bobot daun w sebagai berikut.

$$\Omega(\theta) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \dots \quad (14)$$

Pembaruan bobot daun pada XGB dilakukan dengan menggunakan gradien pertama untuk meminimalkan kesalahan dan memperbaiki kualitas model. Bobot daun w_j dihitung dengan gradien pertama dari fungsi kerugian untuk setiap daun pohon, dan gradien kedua digunakan untuk memperbaiki pembaruan ini. Fungsi *loss* pada setiap iterasi t dihitung sebagai persamaan 15 berikut.

Dimana g_i dan h_i merupakan gradien pertama dan kedua dari fungsi kerugian, γ merupakan pinlati untuk jumlah daun pohon, T adalah jumlah daun pohon, dan λ adalah pinalti untuk bobot pohon. Setelah bobot daun diperbarui, XGB menggunakan teknik pruning pohon untuk mengurangi kompleksitas model. Fungsi kerugian yang menghitung kontribusi dari setiap daun dihitung dengan formula (16) berikut.

Dengan I_j adalah set sampel yang ada di daun j

Pada pustaka scikit-learn dari Python, parameter *maximum depth* dan *n_estimators* secara berturut turut mewakili kedalaman maksimum pohon serta jumlah pohon. Kemudian, parameter *subsample* berfungsi untuk mengontrol proporsi data pelatihan yang akan digunakan untuk membangun setiap pohon.

III METODE

3.1 Data Penelitian

Data penelitian dikumpulkan dari basis data ChEMBL pada tanggal 9 Oktober 2024. Data yang diunduh mencakup informasi ID senyawa, representasi struktur dalam bentuk SMILES, serta hasil uji biologis berupa nilai IC₅₀. Proses pengambilan data dilakukan dengan menggunakan kata kunci *Plasmodium falciparum*, dengan filter pada tipe standar (*standard type*) IC₅₀ dan satuan unit (*standard unit*) nanomolar (nM). Pengumpulan data ini memanfaatkan *library chembl_webresource_client* yang disediakan oleh ChEMBL untuk mengakses basis data secara langsung. Seluruh data yang digunakan pada penelitian ini dapat diperoleh di <https://github.com/TropBRC-BioinfoLab/antimalaria-ic50-prediction>.

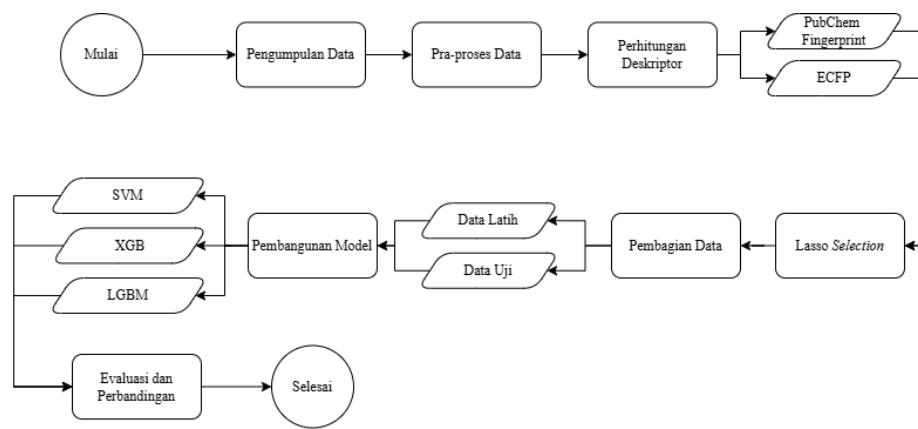
3.2 Peralatan Penelitian

Penelitian ini dilakukan dengan menggunakan perangkat lunak dan perangkat keras sebagai berikut:

1. Perangkat keras berupa laptop dengan Processor Intel® Core™ i5-10300H CPU @ 2.50 GHz, penyimpanan Hard Disk sebesar 512GB, dan memori RAM sebesar 16GB.
2. *Operating System (OS)* Windows 11 64-bit, *code editor* Visual Studio Code, platform pemodelan dan eksperimen *Kaggle Notebook*, serta bahasa pemrograman Python sebagai perangkat lunak.
3. Data penelitian yang terdapat pada repositori <https://github.com/TropBRC-BioinfoLab/antimalaria-ic50-prediction>.

3.3 Metode

Penelitian akan diawali dengan pengumpulan data dan diakhiri dengan evaluasi serta perbandingan. Seluruh alur penelitian tertera pada diagram berikut.



Gambar 3 Metode penelitian

3.3.1 Pra-proses Data

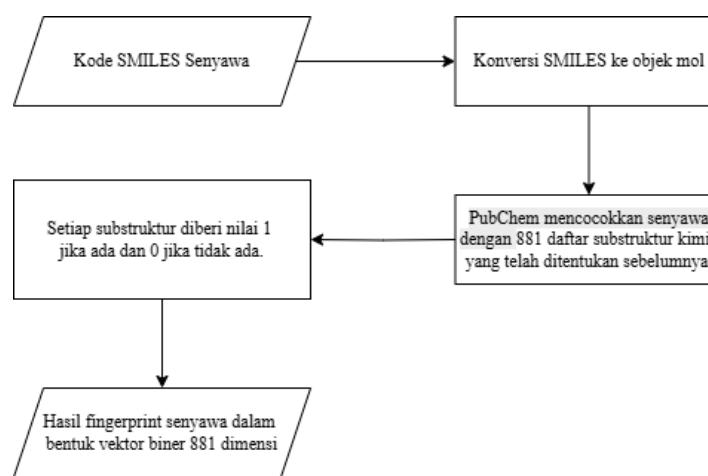
Pada tahap pra-pemrosesan data, langkah pertama yang dilakukan adalah menghapus data duplikat serta menghilangkan entri dengan nilai IC₅₀ atau kode

SMILES yang kosong. Selanjutnya, nilai IC_{50} dikonversi ke bentuk pIC_{50} menggunakan rumus $pIC_{50} = - \log_{10} (IC_{50}$ dalam molar). Nilai IC_{50} dapat memiliki rentang yang sangat luas, misalnya dari 0.001 nM hingga lebih dari 100 μM . Rentang yang ekstrem ini berpotensi menimbulkan *skewness* pada distribusi data yang dapat memengaruhi kinerja model machine learning. Distribusi yang tidak seimbang dapat menyebabkan model cenderung bias terhadap nilai-nilai tertentu dan mengurangi kemampuan generalisasi. Untuk mengatasi hal tersebut, dilakukan transformasi ke pIC_{50} yang mengubah skala menjadi lebih ringkas, yaitu sekitar 0 hingga 12. Setelah konversi, dilakukan pembersihan data outlier untuk meningkatkan kualitas dataset sebelum tahap ekstraksi deskriptor dan pemodelan *machine learning*.

3.3.2 Perhitungan Deskriptor

Perhitungan dua deskriptor utama meliputi deskriptor PubChem Fingerprint (881 bit) untuk mendeteksi substruktur kimia dan deskriptor ECFP (2048 bit) berbasis graf yang mendeskripsikan koneksi atom dalam molekul. Nilai biner mewakili ada dan tidaknya suatu fitur pada senyawa tersebut. Representasi ini digunakan sebagai input dalam analisis *machine learning* untuk membandingkan performa model dalam memprediksi aktivitas senyawa.

Proses ekstraksi senyawa dilakukan dengan memanfaatkan representasi SMILES (*Simplified Molecular Input Line Entry System*) sebagai input awal. SMILES adalah format teksual yang merepresentasikan struktur molekul secara kompak dan dapat diproses secara komputasional. Langkah pertama dimulai dengan konversi kode SMILES menjadi objek molekul digital menggunakan pustaka kimia komputasi (*library*) yang disediakan oleh RDKit (<https://www.rdkit.org/docs/source/rdkit.Chem.rdmolfiles.html#rdkit.Chem.rdmolfiles.MolFromSmiles>). Objek molekul ini kemudian dianalisis untuk mencocokkan keberadaan berbagai substruktur kimia.



Gambar 4 Diagram proses ekstraksi deskriptor PubChem Fingerprint

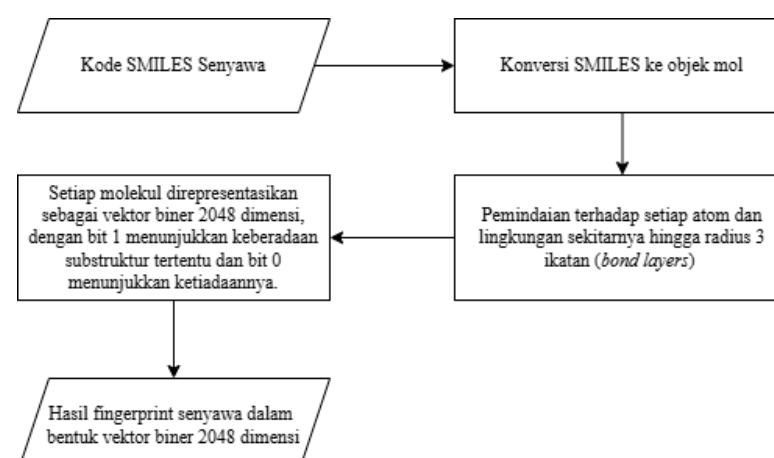
PubChem Substructure Fingerprint merupakan metode yang menggunakan 881 substruktur kimia yang telah didefinisikan sebelumnya. Sistem secara otomatis mencocokkan objek molekul dengan daftar substruktur

tersebut. Untuk setiap substruktur, akan diberikan nilai 1 jika substruktur tersebut terdapat dalam senyawa dan 0 jika tidak ada.

Tabel 3 Contoh hasil ekstraksi senyawa menggunakan deskriptor Pubchem Fingerprint

Senyawa	bit0	bit1	bit2	...	bit880
Senyawa 1	0	1	0	...	1
Senyawa 2	1	0	1	...	0
...
Senyawa - n	1	0	1	...	1

Tabel di atas menunjukkan hasil representasi Pubchem Fingerprint dari sejumlah senyawa kimia dalam bentuk biner. Setiap baris pada tabel merepresentasikan satu senyawa yang telah dikonversi dari format SMILES ke bentuk digital dan dianalisis menggunakan metode PubChem Fingerprint. Sementara itu, setiap kolom mulai dari bit0 hingga bit880 menunjukkan status keberadaan substruktur kimia tertentu yang telah ditentukan sebelumnya dalam skema *fingerprint* PubChem.



Gambar 5 Diagram proses ekstraksi deskriptor ECFP

Pada deskriptor ECFP, objek molekul selanjutnya dipindai untuk mendeteksi setiap atom dan lingkungan kimianya hingga radius tiga ikatan (bond layers) guna menangkap informasi struktural lokal dari senyawa tersebut. Hasil pemindaian ini digunakan untuk menghasilkan representasi vektor biner berdimensi 2048, di mana setiap bit merepresentasikan keberadaan atau ketidakhadiran suatu pola atau fragmen struktur tertentu dalam molekul.

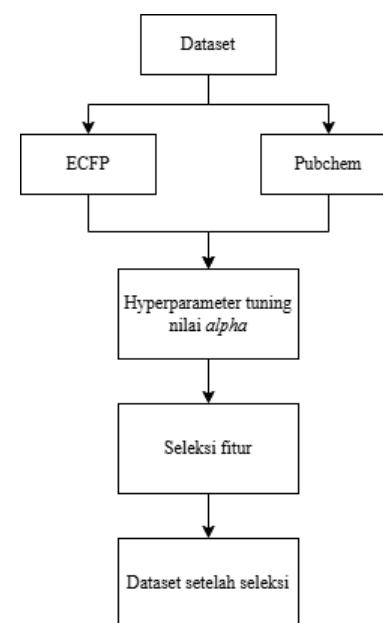
Tabel 4 Contoh data hasil ekstraksi deskriptor ECFP

Senyawa	bit0	bit1	bit2	...	Bit2047
Senyawa 1	0	1	0	...	1
Senyawa 2	1	0	1	...	0
...

Senyawa - n	1	0	1	...	1
-------------	---	---	---	-----	---

Tabel di atas menunjukkan hasil representasi Extended-Connectivity Fingerprint (ECFP) dari sejumlah senyawa kimia dalam bentuk vektor biner. Setiap baris pada tabel merepresentasikan satu senyawa yang telah dikonversi dari format SMILES ke bentuk digital, kemudian dianalisis menggunakan metode ECFP. Setiap kolom mulai dari bit0 hingga bit2047, menunjukkan keberadaan atau ketidakhadiran pola-pola atom dan lingkungan kimianya yang terdeteksi berdasarkan radius 3 ikatan.

3.3.3 Seleksi Fitur



Gambar 6 Diagram proses seleksi fitur

Langkah selanjutnya setelah ekstraksi deskriptor adalah melakukan *hyperparameter tuning* terhadap nilai *alpha* pada metode Lasso (*Least Absolute Shrinkage and Selection Operator*). Parameter *alpha* mengendalikan kekuatan regularisasi sehingga pemilihan nilai yang tepat menjadi penting untuk menyeimbangkan antara kompleksitas model dan kemampuan generalisasi. *Hyperparameter tuning* dilakukan untuk menemukan nilai *alpha* yang memberikan kinerja optimal dalam mengurangi kesalahan prediksi pada data validasi.

Setelah nilai *alpha* optimal diperoleh, dilakukan tahap seleksi fitur dengan menerapkan model Lasso menggunakan nilai *alpha* tersebut. Proses ini mengeliminasi fitur-fitur yang koefisiennya mengecil hingga nol, sehingga hanya fitur yang relevan dan signifikan yang dipertahankan. Hasil akhirnya adalah dataset dengan jumlah fitur yang lebih sedikit dibandingkan dataset awal, di mana fitur-fitur tersebut dianggap paling berpengaruh dalam memprediksi aktivitas senyawa. Dataset setelah seleksi ini kemudian digunakan sebagai masukan untuk tahap pemodelan.



3.3.4 Pembagian Data

Pada tahap pembagian data, data yang telah diproses dibagi menjadi dua set untuk mendukung pelatihan dan evaluasi model. Data latih digunakan untuk membangun model dan mencakup 70% dari total data. Data uji digunakan untuk mengevaluasi performa model pada data yang tidak terlihat sebelumnya, mencakup 30% dari total data. Pembagian ini bertujuan memastikan model dapat belajar secara efektif sekaligus mengukur kemampuan generalisasi terhadap data baru.

3.3.5 Pembangunan Model

Model dibangun menggunakan algoritma XGB, LGBM dan SVM dengan bantuan pustaka scikit-learn untuk memprediksi aktivitas antimalaria senyawa berdasarkan deskriptor molekuler. *Hyperparameter tuning* dilakukan untuk mengoptimalkan parameter seperti jumlah pohon keputusan, kedalaman pohon, *learning rate*, dan ukuran *batch*. Proses ini bertujuan untuk menemukan kombinasi parameter terbaik yang dapat meningkatkan akurasi dan kemampuan model dalam melakukan prediksi nilai IC₅₀. Rentang nilai parameter yang digunakan untuk masing-masing algoritma, yaitu LGBM, XGB, dan Support Vector Machine (SVM) ditampilkan pada tabel berikut.

Tabel 5 Kombinasi parameter untuk *hyperparameter tuning* SVM

Parameter	Kombinasi Nilai
<i>Regularization coefficient</i> (C)	0.1, 1, 10, 100
Epsilon (ϵ)	0.01, 0.1, 0.2
Gamma	<i>scale</i> , 0.01, 0.1, 1
Kernel	<i>rbf</i>

Tabel 6 Kombinasi parameter untuk *hyperparameter tuning* XGB

Parameter	Kombinasi Nilai
<i>Learning rate</i> (γ)	0.01, 0.05, 0.1
<i>Max depth</i>	10, 15, 20
<i>Sub sample</i>	0.5, 0.7, 0.9
<i>N estimators</i>	200, 500, 1000

Tabel 7 Kombinasi parameter untuk *hyperparameter tuning* LGBM

Parameter	Kombinasi Nilai
<i>Number of Leaves</i>	31, 63, 127
<i>Max depth</i>	-1, 3, 5, 7
<i>Learning rate</i> (γ)	0.01, 0.05, 0.1
<i>N estimators</i>	100, 300, 500

<i>Sub sample</i>	0.6, 0.8, 1.0
<i>Colsample bytree</i>	0.6, 0.8, 1.0

3.3.6 Evaluasi dan Perbandingan

Pada tahap evaluasi dan perbandingan, model yang dibangun dievaluasi berdasarkan metrik kinerja Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan R². Selain itu, digunakan juga metrik tambahan berupa Mean Absolute Percentage Error (MAPE), Median Absolute Error (MedAE), dan Explained Variance Score (EVS).

- Mean Squared Error (MSE) mengukur rata-rata kuadrat selisih antara nilai yang diprediksi (\hat{y}_t) dan nilai aktual (y_t) dari seluruh jumlah data (n). MSE direpresentasikan pada persamaan (17).

2. Mean Absolute Error (MAE) mengukur rata-rata dari nilai mutlak kesalahan prediksi yang memberikan gambaran seberapa jauh prediksi dari nilai sebenarnya. MAE direpresentasikan pada persamaan (18).

3. R-squared (R^2) mengukur proporsi variabilitas dalam data yang dapat dijelaskan oleh model. Semakin tinggi nilai R^2 maka semakin baik model dalam menjelaskan variasi data. R^2 direpresentasikan pada persamaan (19).

4. Mean Absolute Percentage Error (MAPE) mengukur seberapa besar kesalahan prediksi dibandingkan nilai aktual dalam bentuk persentase. MAPE direpresentasikan pada persamaan (20).

5. Median Absolute Error (MedAE) berfungsi untuk menunjukkan nilai tengah dari kesalahan prediksi absolut. MedAE direpresentasikan pada persamaan (21).

6. Explained Variance Score (EVS) mengukur seberapa besar variasi data target yang dapat dijelaskan model secara konsisten tanpa terpengaruh nilai rata-rata. EVS direpresentasikan pada persamaan (22).

Untuk mendapatkan gambaran ketepatan estimasi metrik kinerja dan stabilitas model, dilakukan pula perhitungan Confidence Interval (CI) pada hasil evaluasi menggunakan data latih dengan teknik *cross-validation 5 fold*. Pada setiap *fold*, model dilatih pada sebagian data dan divalidasi pada bagian lain sehingga diperoleh 5 nilai metrik kinerja yang berbeda. Berdasarkan nilai-nilai tersebut, dihitung rata-rata metrik dan interval kepercayaan sebesar 0.95. Interval kepercayaan ini merepresentasikan rentang nilai metrik kinerja yang diharapkan pada populasi data secara umum dengan tingkat keyakinan 95%. Persamaan untuk mencari *lower bound* dan *upper bound* pada CI sebagai berikut.

Dengan \bar{x} adalah nilai rata-rata, t merupakan tingkat kepercayaan, s mewakili simpangan baku, dan n adalah jumlah data. Kemudian, dihitung rata-rata dengan persamaan sebagai berikut.

$$Mean\ CI = \frac{Lower\ Bound + Upper\ Bound}{2}(25)$$

Setelah proses *cross-validation* selesai dan didapatkan gambaran performa beserta CI pada data latih, model kemudian dilatih ulang menggunakan seluruh data latih (tanpa pembagian *fold*) untuk mendapatkan model final yang optimal. Model final ini kemudian dievaluasi ulang menggunakan data uji yang belum pernah dilihat sebelumnya untuk mengukur kemampuan generalisasi model terhadap data baru di luar data latih.

Hasil prediksi akan diklasifikasikan ke dalam lima kategori berdasarkan tinggi rendahnya nilai IC₅₀, yaitu aktivitas sangat baik atau poten, aktivitas baik atau sitotoksitas sangat kuat, aktivitas sedang, aktivitas rendah, dan tidak aktif. Untuk mengevaluasi model dengan data klasifikasi, digunakan empat metrik evaluasi utama, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Ketiga metrik terakhir digunakan menggunakan metode *weight averaging* yang mempertimbangkan proporsi jumlah data pada tiap kelas, sehingga hasil evaluasi lebih representatif ketika distribusi data antar kelas tidak seimbang.

1. Akurasi (*accuracy*) proporsi prediksi yang benar dibandingkan dengan seluruh data uji. Metrik ini memberikan informasi seberapa sering model memberikan prediksi yang tepat.

$$Akurasi = \frac{\text{Jumlah prediksi benar}}{\text{Jumlah seluruh data}} \dots\dots\dots(26)$$

2. *Precision* mengukur tingkat ketepatan model dalam memprediksi suatu kelas positif. Metrik ini menunjukkan berapa banyak data yang benar-benar positif

dari seluruh data yang diprediksi positif oleh model. Pada kasus multi-kelas, *precision* dihitung untuk setiap kelas, kemudian dirata-ratakan menggunakan *weighted averaging*. Bobot (*weight*) pada masing-masing kelas ditentukan oleh jumlah sampel di kelas tersebut. Rumus *precision* untuk kelas ke-*i* adalah:

Sedangkan rumus *weighted precision* adalah

$$P_{weighted} = \frac{\sum_{i=1}^C P_i \times n_i}{N} \quad \dots \quad (28)$$

3. *Recall* mengukur kemampuan model dalam menemukan seluruh data yang sebenarnya positif. Metrik ini fokus pada meminimalkan kesalahan prediksi negatif (False Negative). Seperti pada *precision*, *recall* untuk multi-kelas dihitung per kelas, lalu digabung menggunakan *weighted averaging*. Rumus *recall* untuk kelas ke-*i* adalah:

$$Precision_i = \frac{(True\ Positive)_i}{(True\ Positive)_i + (False\ Negative)_i} \quad \dots \dots \dots (29)$$

Sedangkan rumus *weighted recall* adalah

4. F1 – score merupakan rata-rata harmonis dari precision dan recall. Untuk multi-kelas, F1-score dihitung per kelas kemudian dirata-ratakan menggunakan *weighted averaging* dengan bobot sesuai jumlah sampel per kelas:

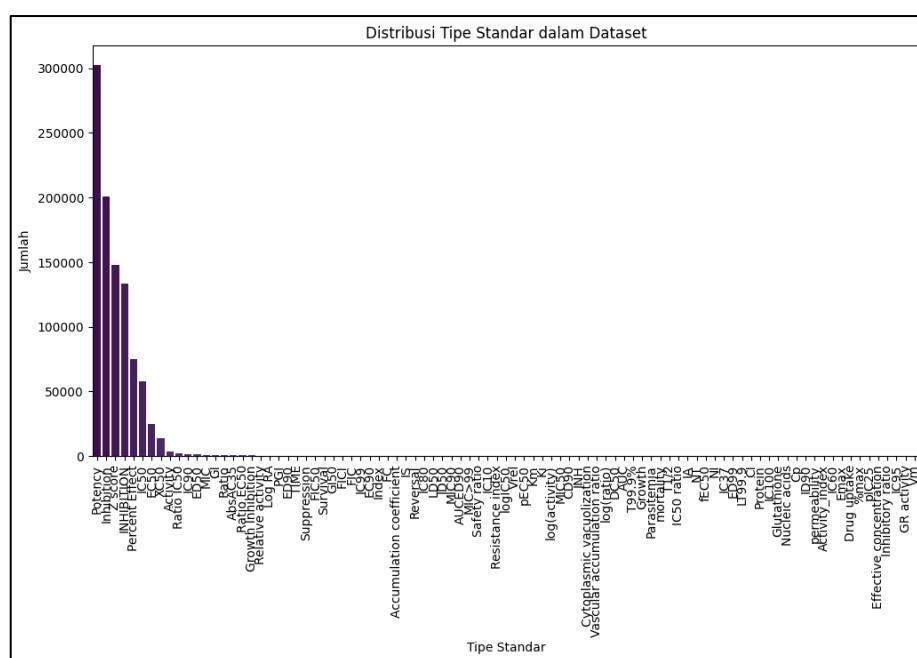
Sedangkan rumus *weighted F1 – score* adalah

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau s

IV HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Dalam penelitian ini, sebanyak 969.219 data senyawa yang telah diujii terhadap parasit dikumpulkan melalui basis data ChEMBL. Distribusi *standard type* pada dataset mewakili jenis pengukuran aktivitas biologis yang digunakan untuk menilai efektivitas suatu senyawa terhadap target biologis tertentu. Setiap entri dalam dataset memiliki satu jenis *standard type* yang menunjukkan parameter uji, seperti potensi, persentase inhibisi, atau konsentrasi efektif. Lima tipe standar dengan jumlah entri terbanyak dalam dataset adalah Potency, Inhibition, IC₅₀, Percent inhibition, dan Effect. Potency memiliki jumlah entri paling tinggi sekitar 302.506 data, diikuti oleh Inhibition sebanyak 334.492 data, Z Score sebanyak 147.592, Percent Effect sebanyak 74.710, dan IC₅₀ sebanyak 57.490 data. Distribusi lengkap dari seluruh standard type dapat dilihat pada grafik berikut.



Gambar 7 Distribusi tipe standar dalam dataset

Dalam penelitian ini, tipe standar yang digunakan difokuskan pada IC₅₀ karena merupakan salah satu parameter paling umum dalam mengevaluasi efektivitas senyawa terhadap target biologis. IC₅₀ menunjukkan konsentrasi senyawa yang dibutuhkan untuk menghambat 50% aktivitas biologis, sehingga sangat relevan untuk prediksi potensi senyawa antimalaria. Untuk memastikan konsistensi pengukuran, dilakukan analisis terhadap distribusi satuan standar (*standard unit*) yang digunakan dalam data IC₅₀. Hasil analisis menunjukkan bahwa sebagian besar data menggunakan satuan nM (nanomolar) yang berjumlah sebanyak 50.251, sementara satuan lainnya seperti µg/mL, µM/well, dan ppm hanya digunakan dalam jumlah kecil. Dominasi satuan nM ini menjadi dasar untuk



4.2 Pembersihan Data

Setelah berhasil mengumpulkan 45.314 data senyawa, langkah selanjutnya adalah melakukan pembersihan data yang mencakup penghapusan nilai null pada kolom *canonical SMILES* dan *standard value*, penghapusan nilai 0 pada kolom *standard value*, serta penanganan nilai outlier. Terdapat 82 entri dengan nilai null pada kolom *canonical SMILES* dan 1 entri dengan nilai null pada kolom *standard value*. Selain itu, terdapat 96 entri dengan *standard value* bernilai 0. Senyawa dengan nilai IC₅₀ sebesar 0 perlu dihapus karena tidak memiliki makna biologis yang valid, mengingat nilai ini seharusnya berupa angka kontinu positif lebih besar dari 0. Tabel 6 berikut menjelaskan jumlah data sebelum dan setelah proses pembersihan.

Tabel 9 Jumlah data sebelum dan setelah dibersihkan

Jumlah data awal	Jumlah senyawa dengan nilai null	Jumlah senyawa dengan IC ₅₀ bernilai 0	Jumlah data akhir
45.314	83	96	45.135

Boxplot IC₅₀ pada Gambar 9 memperlihatkan sebaran nilai IC₅₀ dalam dataset dengan cukup jelas. Mayoritas data terkonsentrasi pada rentang nilai yang relatif rendah dan rapat di sekitar median, sebagaimana terlihat pada bagian kotak utama. Namun, terdapat banyak titik outlier yang tersebar jauh di atas *whisker* atas sehingga menunjukkan adanya nilai IC₅₀ yang sangat tinggi dibandingkan mayoritas data. Penyebaran outlier ini sangat mencolok dan menunjukkan bahwa terdapat senyawa-senyawa dengan nilai IC₅₀ yang ekstrem dalam dataset.

Secara biologis, outlier dengan nilai IC₅₀ yang tinggi mencerminkan senyawa dengan potensi lemah karena memerlukan konsentrasi besar untuk menghasilkan efek penghambatan 50%. Meskipun informasi ini relevan secara ilmiah, keberadaan nilai-nilai ekstrem tersebut dapat memengaruhi kinerja model *machine learning* dalam memprediksi potensi senyawa. Nilai yang terlalu jauh dari mayoritas data berpotensi memperlebar rentang distribusi sehingga menyebabkan model kesulitan mempelajari pola umum, meningkatkan risiko bias, dan penurunan akurasi prediksi. Oleh karena itu, penanganan outlier perlu dipertimbangkan dalam tahap prapemrosesan data untuk memastikan model mendapatkan input yang lebih seimbang dan representatif.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

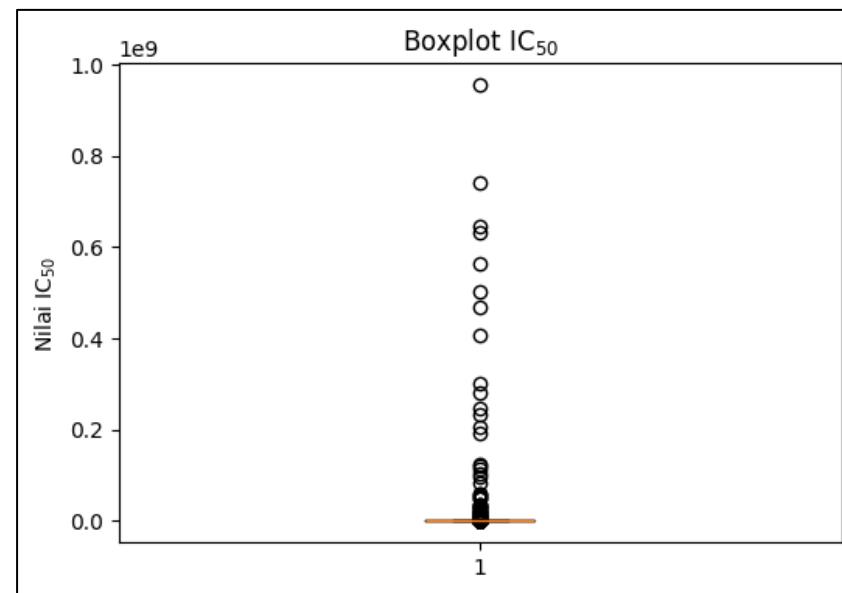
Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

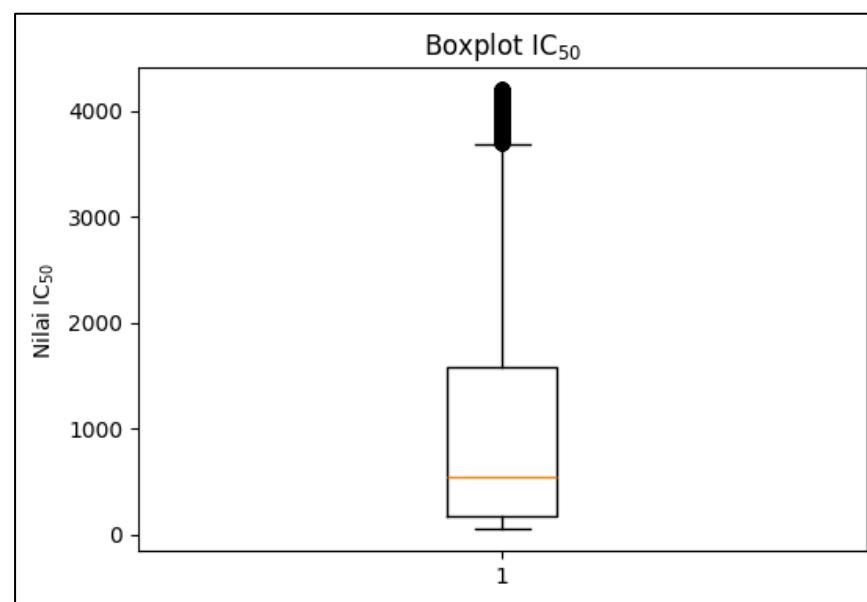
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Gambar 9 Boxplot sebaran nilai IC₅₀

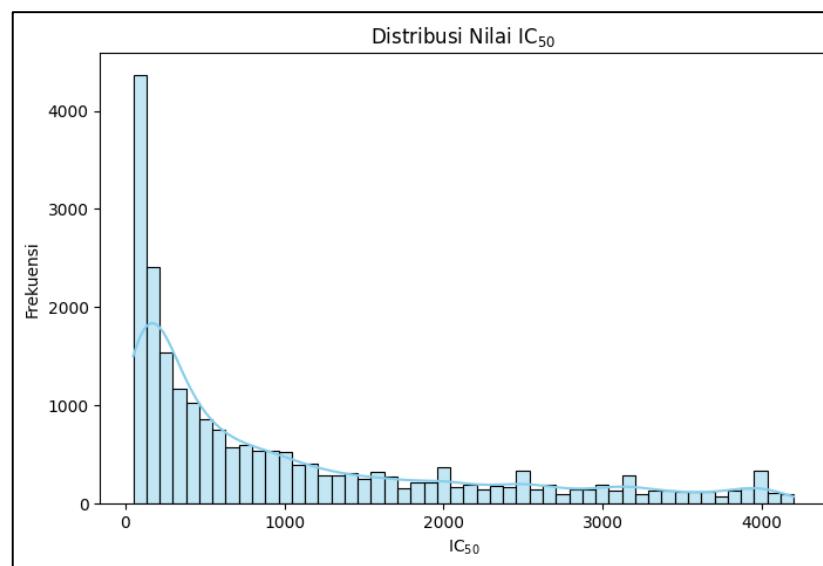
Untuk menangani outlier pada nilai IC₅₀, dilakukan identifikasi menggunakan metode Interquartile Range (IQR). Langkah pertama adalah menghitung kuartil pertama (Q1) dan kuartil ketiga (Q3) dari distribusi nilai IC₅₀. Selanjutnya, nilai IQR diperoleh dari selisih Q3 dan Q1. Berdasarkan nilai IQR tersebut, ditentukan batas bawah dan batas atas dengan rumus $Q1 - 1.5 \times IQR$ dan $Q3 + 1.5 \times IQR$. Nilai-nilai yang berada di luar rentang batas bawah dan atas ini diidentifikasi sebagai outlier. Pada tahap ini, hanya nilai IC₅₀ yang berada di antara batas bawah dan atas yang diambil dan dipertahankan untuk analisis lebih lanjut. Gambar 10 menunjukkan sebaran nilai IC₅₀ setelah dilakukan penanganan.



Gambar 10 Boxplot sebaran nilai IC₅₀ setelah proses pembersihan outlier

4.3 Transformasi Data

Hasil pengukuran nilai skewness dan kurtosis menunjukkan adanya perubahan distribusi yang signifikan setelah dilakukan transformasi. Sebelum transformasi, nilai skewness sebesar 1.25 mengindikasikan distribusi data yang miring ke kanan (*positively skewed*), dengan ekor distribusi yang lebih panjang di sisi nilai yang lebih besar. Nilai kurtosis sebesar 0.46 menunjukkan distribusi yang sedikit lebih lebar atau *platykurtic* dibanding distribusi normal, meski tidak terlalu ekstrem. Gambar 11 berikut merupakan histogram sebaran distribusi nilai IC₅₀.



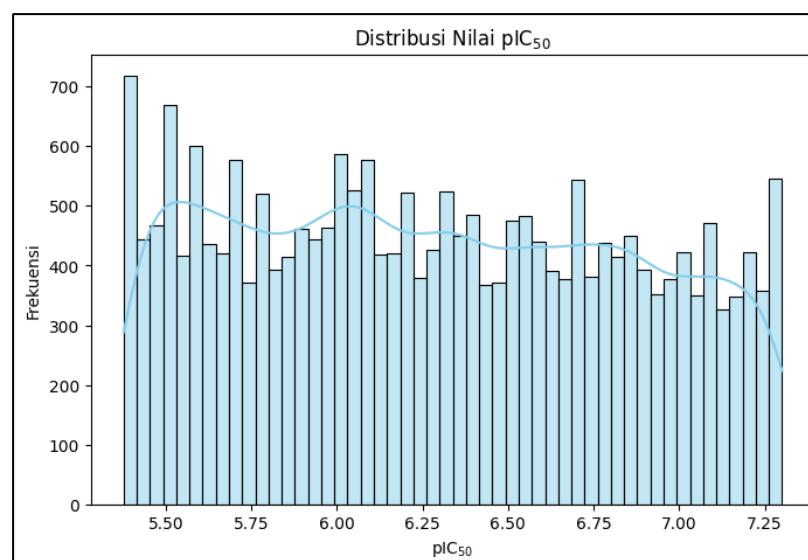
Gambar 11 Histogram distribusi nilai IC₅₀

Tabel 10 Lima senyawa dengan nilai IC50 terendah

No	Canonical SMILES	Nama IUPAC Senyawa	IC ₅₀ (nM)
1	Cc1ccc(S(=O)(=O)NCCCCNc2ccnc3cc(Cl)ccc23)cc1	N-[4-[(7-chloroquinolin-4-yl)amino]butyl]-4-methylbenzenesulfonamide	50
2	CCCN(CCCNc1ccnc2cc(Cl)ccc12)Cc1ccsc1	N-(7-chloroquinolin-4-yl)-N'-propyl-N'-(thiophen-3-ylmethyl)propane-1,3-diamine	50
3	COc1cc2c(cc1CN1CCC(N3CCCCC3)CC1)[nH]c1c3ccc(Cl)cc3ncc21	3-chloro-8-methoxy-9-[(4-piperidin-1-yl)piperidin-1-yl)methyl]-11H-indolo[3,2-c]quinoline	50
4	CCN(CC)Cc1cc(Cl)cc(Nc2nc3cccc3n2Cc2ccc(C(F)(F)F)cc2)c1O	4-chloro-2-(diethylaminomethyl)-6-[[1-[[4-(trifluoromethyl)phenyl]methyl]benzimidazol-2-yl]amino]phenol	50

5	CCN(CC)CC#CCNc1ccnc2cc(Cl)ccc12	N-(7-chloroquinolin-4-yl)-N',N'-diethylbut-2-yne-1,4-diamine	50
---	---------------------------------	--	----

Proses transformasi nilai IC_{50} ke pIC_{50} dilakukan untuk mengubah skala data menjadi bentuk logaritmik yang lebih sesuai untuk analisis regresi. Nilai IC_{50} awal yang dicatat dalam satuan nanomolar (nM) terlebih dahulu dikonversi ke satuan molar dengan membaginya dengan 10^9 . Setelah itu, nilai pIC_{50} dihitung menggunakan rumus $-\log_{10}(IC_{50} \text{ dalam molar})$. Setelah dilakukan transformasi, nilai skewness menurun drastis menjadi 0.10 yang mendekati nol, sehingga menandakan distribusi data menjadi jauh lebih simetris. Selain itu, nilai kurtosis berubah menjadi -1.17, menunjukkan distribusi yang lebih rata atau *platykurtic* dengan puncak yang lebih landai dibanding distribusi normal. Gambar 12 menunjukkan histogram sebaran nilai IC_{50} setelah dilakukan proses transformasi.



Gambar 12 Histogram distribusi nilai IC_{50} setelah transformasi

Tabel 11 Contoh data hasil transformasi

No	ChemBL ID	Canonical SMILES	IC_{50} (nM)	pIC_{50}
1	CHEMBL16300	O=C(NO)c1ccccc1	17600	4.754487
2	CHEMBL287556	O=c1cc(CO)occ1O	1580000	2.801343
3	CHEMBL328154	CC(C)c1cc(Nc2ccnc3cc(Cl)ccc23)cc(CNC(C)(C)C)c1O	1.24	8.906578
4	CHEMBL330342	CC(C)(C)NCc1cc(Nc2ccnc3cc(Cl)ccc23)cc(C(C)(C)C)c1O	8.36	8.077793
5	CHEMBL73090	COc1ccc(C(=O)/C=C/c2ccccc2)c(OC)c1	55600	4.254925

4.4 Ekstraksi Fitur

Senyawa yang telah melalui tahap pra-pemrosesan data akan diekstraksi lebih lanjut untuk mengidentifikasi karakteristik struktural dan kimianya. Ekstraksi deskriptor dilakukan menggunakan berbagai metode, di mana PubChem Fingerprint diekstraksi dengan bantuan library DeepChem yang menghasilkan representasi fitur sebanyak 881 bit. Selain itu, deskriptor Extended Connectivity Fingerprint (ECFP) diekstraksi menggunakan *library* RDKit yang menghasilkan 2048 bit fitur.

Tabel 12 Contoh hasil ekstraksi deskriptor Pubchem Fingerprint

<i>Molecule ChemBL ID</i>	Pub 0	Pub 1	Pub 2	Pub 3	Pub 4	...	Pub 880	pIC ₅₀
CHEMBL77052	1	1	1	1	0	...	0	8.3716
CHEMBL16300	1	0	0	0	0	...	0	4.7544
CHEMBL307153	1	1	1	1	0	...	0	8.0199
CHEMBL339049	1	1	1	0	0	...	0	7.5459
CHEMBL316098	1	1	1	0	0	...	0	8.8326

Tabel 13 Contoh hasil deskriptor ECFP

<i>Molecule ChemBL ID</i>	ecfp 0	ecfp 1	ecfp 2	ecfp 3	ecfp 4	...	ecfp 2047	pIC ₅₀
CHEMBL77052	0	0	0	0	0	...	0	8.3716
CHEMBL16300	0	0	0	0	0	...	0	4.7544
CHEMBL307153	0	0	0	0	0	...	0	8.0199
CHEMBL339049	0	0	0	0	0	...	0	7.5459
CHEMBL316098	0	0	0	0	0	...	0	8.8326

Setelah proses ekstraksi menggunakan ketiga deskriptor, ditemukan bahwa beberapa senyawa memiliki fitur dengan nilai *null*. Untuk menghindari potensi gangguan dalam pembangunan model, senyawa-senyawa tersebut dihapus dari dataset. Setelah tahap ini, jumlah senyawa akhir yang berhasil diekstraksi menggunakan deskriptor PubChem Fingerprint adalah 22.577 senyawa dan deskriptor ECFP sebanyak 22.635 senyawa.

4.5 Seleksi Fitur (*Feature Selection*)

Setelah melalui tahapan awal berupa ekstraksi deskriptor molekuler dan pembersihan data, tahap selanjutnya adalah proses *hyperparameter tuning* nilai *alpha* pada regresi Lasso yang bertujuan untuk menentukan tingkat regularisasi optimal dalam seleksi fitur. Model dilatih menggunakan fungsi LassoCV dari pustaka scikit-learn dengan skema *5 fold cross-validation*. Parameter *max_iter* diatur sebesar 10.000 untuk memastikan proses konvergen, dan *random_state* ditetapkan untuk menjaga replikasi hasil. Selama proses ini, beberapa nilai *alpha* dicoba secara otomatis dan performanya diukur menggunakan metrik Mean Squared Error (MSE). Tabel 11 merupakan lima nilai *alpha* terbaik berdasarkan hasil proses ini.

Tabel 14 Hasil pemilihan nilai *alpha* terbaik berdasarkan proses *hyperparameter tuning*

No	MSE Pubchem FP	MSE ECFP	Alpha Pubchem FP	Alpha ECFP
1	0.247063	0.253150	0.000089	0.000277
2	0.247064	0.253151	0.000083	0.000258
3	0.247066	0.253155	0.000096	0.000297
4	0.247068	0.253157	0.000078	0.000241
5	0.247073	0.253166	0.000072	0.000225

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tabel 15 Perbandingan jumlah fitur awal dan setelah seleksi

Deskriptor	Jumlah fitur awal	Jumlah fitur setelah seleksi
PubchemFP	881	448
ECFP	2048	1271

Pada analisis ini, Tabel 10 dan Tabel 11 menyajikan daftar 10 fitur dengan skor *importance* tertinggi pada deskriptor ECFP dan PubChem Fingerprint. Pada Tabel 10, fitur yang memiliki skor *importance* tertinggi dimiliki oleh ecfp38. Namun, tabel ini tidak memiliki kolom representasi yang spesifik karena setiap bit yang terbentuk mewakili fitur yang berbeda-beda. Hal ini akibat dari proses ekstraksi yang bersifat dinamis. Pada PubChem Fingerprint, fitur Pub750 yang merepresentasikan ada tidaknya substruktur Nc1cc(Cl)ccc1 memiliki skor *importance* tertinggi. Makna fitur pada deskriptor ini dapat dijelaskan karena telah terdefinisi dengan jelas pada situs web PubChem.

Tabel 16 Daftar 10 fitur dengan skor *importance* tertinggi pada deskriptor ECFP

Fitur	Skor Importance
ecfp38	0.025974
ecfp1097	0.023231
ecfp103	0.020107
ecfp1733	0.019576
ecfp722	0.015856
ecfp1911	0.011919
ecfp650	0.011033



ecfp1917	0.010369
ecfp1152	0.010040
ecfp251	0.009842

Tabel 17 Daftar 10 fitur dengan skor importance tertinggi pada deskriptor Pubchem Fingerprint

Fitur	Representasi Fitur pada Senyawa	Skor Importance
Pub750	Nc1cc(Cl)ccc1	0.117291
Pub601	N-C:C:C-N	0.058639
Pub309	O-O	0.038260
Pub514	O-N-C-C	0.036241
Pub16	≥ 4 N	0.034295
Pub19	≥ 2 O	0.026026
Pub248	≥ 1 cincin karbon jenuh atau aromatik berukuran 10 atom	0.025317
Pub636	C-N-C-N-C	0.024224
Pub643	[#1]-C-C-N-[#1]	0.023755
Pub696	C-C-C-C-C-C-C-C	0.023639

4.6 Pembagian Data

Tahapan selanjutnya adalah pembagian data menjadi 70% untuk latih dan 30% untuk uji secara acak dengan parameter random_state = 42 untuk memastikan hasil yang konsisten. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi performa pada data yang belum pernah dilihat. Jumlah data pada masing-masing subset ditampilkan pada tabel 15 berikut.

Tabel 18 Perbandingan jumlah data latih dan data uji

Deskriptor	Jumlah data latih	Jumlah data uji
PubchemFP	15803	6774
ECFP	15844	6791

4.7 Hyperparameter tuning

Tahapan selanjutnya adalah *hyperparameter tuning*, yaitu proses pencarian kombinasi parameter terbaik untuk meningkatkan performa model. Hasil dari proses tuning ini dijelaskan pada tabel berikut.

Tabel 5 Hasil *hyperparameter tuning* pada SVM

Parameter	Pubchem Fp	ECFP	Lasso – Pubchem Fp	Lasso - ECFP
<i>Regularization coefficient</i>	0.1	1	10	10
Epsilon	0.1	0.1	0.01	0.2
Gamma	<i>scale</i>	<i>scale</i>	<i>scale</i>	<i>scale</i>
Kernel	rbf	rbf	rbf	rbf

Tabel 6 Hasil *hyperparameter tuning* pada XGB

Parameter	Pubchem Fp	ECFP	Lasso – Pubchem Fp	Lasso - ECFP
<i>Learning rate</i>	0.01	0.01	0.01	0.01
<i>Max depth</i>	20	20	20	20
<i>Sub sample</i>	0.7	0.5	0.5	0.9
<i>N estimators</i>	200	500	500	1000

Tabel 7 Hasil *hyperparameter tuning* pada LGBM

Parameter	Pubchem Fp	ECFP	Lasso – Pubchem Fp	Lasso - ECFP
<i>Number of Leaves</i>	127	31	127	127
<i>Max depth</i>	5	5	-1	-1
<i>Learning rate</i>	0.01	0.01	0.1	0.1
<i>N estimators</i>	300	300	500	500
<i>Sub sample</i>	0.6	0.6	0.6	0.6
<i>Colsample bytree</i>	1	0.6	0.8	0.8

4.8 Pelatihan Model dan Evaluasi

Untuk mengetahui kinerja dari masing-masing model, dilakukan evaluasi terhadap tiga algoritma yaitu Support Vector Machine (SVM), XGB (XGB), dan LightGBM (LGBM). Evaluasi dilakukan pada dua skenario, yaitu tanpa seleksi fitur (*Non-lasso selection*) dan dengan seleksi fitur menggunakan Lasso (*Lasso selection*). Data latih akan dievaluasi menggunakan *cross validation* dengan *5 fold* untuk memastikan kestabilan dan kemampuan model dalam menggeneralisasi pada data yang tidak terlihat sebelumnya. Pada setiap iterasi *cross validation*, dataset pelatihan dibagi menjadi lima bagian (*fold*), di mana model dilatih pada empat *fold* dan diuji pada *fold* yang tersisa. Proses ini diulang sebanyak lima kali sehingga setiap bagian data akan digunakan sebagai data uji satu kali. Hasil evaluasi dari setiap *fold* akan dihitung dan dirata-ratakan untuk memberikan gambaran yang lebih stabil tentang kinerja model.

Tabel 8 Metrik dengan Confidence Interval pada algoritma XGB dengan deskriptor ECFP

Metrik	Non-lasso selection			Lasso selection		
	<i>Mean</i>	95% Confidence Interval		<i>Mean</i>	95% Confidence Interval	
		<i>lower</i>	<i>upper</i>		<i>lower</i>	<i>upper</i>
MSE	0.1475	0.1426	0.1525	0.1474	0.1410	0.1538
RMSE	0.3841	0.3777	0.3905	0.3839	0.3756	0.3923
MAE	0.3003	0.2952	0.3054	0.2884	0.2813	0.2954
R ²	0.5289	0.5143	0.5434	0.5293	0.5105	0.5480
MAPE	0.0481	0.0471	0.0491	0.0462	0.0449	0.0474
MedAE	0.2464	0.2399	0.2529	0.2187	0.2099	0.2275
EVS	0.5294	0.5147	0.5442	0.5299	0.5110	0.5488

Tabel 9 Metrik dengan Confidence Interval pada algoritma XGB dengan deskriptor Pubchem Fingerprint

Metrik	Non-lasso selection			Lasso selection		
	<i>Mean</i>	95% Confidence Interval		<i>Mean</i>	95% Confidence Interval	
		<i>lower</i>	<i>upper</i>		<i>lower</i>	<i>upper</i>
MSE	0.1504	0.1463	0.1545	0.1451	0.1390	0.1511
RMSE	0.3878	0.3826	0.3931	0.3808	0.3728	0.3888
MAE	0.3065	0.3038	0.3092	0.2900	0.2854	0.2945
R ²	0.5184	0.4946	0.5423	0.5355	0.5035	0.5657
MAPE	0.0491	0.0486	0.0495	0.0464	0.0456	0.0472
MedAE	0.2552	0.2501	0.2604	0.2250	0.2206	0.2294
EVS	0.5186	0.4950	0.5423	0.5358	0.5059	0.5657

Pada hasil evaluasi kinerja model XGB dengan dua deskriptor molekuler yang berbeda, yaitu ECFP dan Pubchem Fingerprint, terlihat bahwa kedua deskriptor memberikan performa yang cukup mirip meskipun ada perbedaan kecil pada beberapa metrik. Untuk ECFP, model yang menggunakan Non-lasso *selection* menghasilkan MSE sekitar 0.1475 dengan CI 95% antara 0.1426 dan 0.1525, sedangkan Pubchem Fingerprint dengan Non-lasso *selection* menghasilkan MSE sedikit lebih tinggi, sekitar 0.1504 dengan CI 95% antara 0.1463 dan 0.1545. Secara keseluruhan, nilai-nilai MSE, RMSE, dan MAE pada ECFP sedikit lebih rendah dibandingkan dengan Pubchem Fingerprint, meskipun perbedaan tersebut relatif kecil dan tidak signifikan.

Namun, nilai R² dan EVS pada kedua deskriptor cukup mirip, dengan ECFP memberikan sedikit keuntungan pada R² yang lebih tinggi (sekitar 0.5289) dibandingkan dengan Pubchem Fingerprint (sekitar 0.5186). Metrik lain seperti MAPE dan MedAE menunjukkan performa yang sebanding di kedua deskriptor, dengan sedikit perbedaan yang menunjukkan bahwa keduanya cukup efektif dalam memprediksi aktivitas antimalaria senyawa, meskipun ECFP sedikit lebih unggul dalam beberapa metrik.

Tabel 10 Metrik dengan Confidence Interval pada algoritma SVM dengan deskriptor Pubchem Fingerprint

Metrik	Non-lasso selection			Lasso selection		
	Mean	95% Confidence Interval		Mean	95% Confidence Interval	
		lower	upper		lower	upper
MSE	0.2241	0.2217	0.2266	0.1598	0.1551	0.1645
RMSE	0.4734	0.4708	0.4761	0.3997	0.3938	0.4056
MAE	0.3884	0.3854	0.3913	0.3019	0.2975	0.3064
R ²	0.2826	0.2594	0.3058	0.4884	0.4618	0.5150
MAPE	0.0619	0.0615	0.0623	0.0483	0.0476	0.0491
MedAE	0.3475	0.3403	0.3548	0.2283	0.2212	0.2353
EVS	0.2840	0.2600	0.3081	0.4887	0.4625	0.5150

Tabel 11 Metrik dengan Confidence Interval pada algoritma SVM dengan deskriptor ECFP

Metrik	Non-lasso selection			Lasso selection		
	Mean	95% Confidence Interval		Mean	95% Confidence Interval	
		lower	upper		lower	upper
MSE	0.1485	0.1438	0.1532	0.1496	0.1434	0.1558
RMSE	0.3853	0.3792	0.3914	0.3867	0.3787	0.3948
MAE	0.2940	0.2880	0.3000	0.2993	0.2911	0.3076
R ²	0.5259	0.5130	0.5388	0.5224	0.5051	0.5398
MAPE	0.0470	0.0459	0.0481	0.0480	0.0465	0.0494
MedAE	0.2262	0.2196	0.2328	0.2310	0.2225	0.2394
EVS	0.5262	0.5130	0.5394	0.5231	0.5055	0.5407

Hasil evaluasi model Support Vector Machine (SVM) dengan dua deskriptor molekuler, yaitu Pubchem Fingerprint dan ECFP, menunjukkan bahwa penggunaan Lasso *selection* memberikan perbaikan signifikan pada beberapa metrik terutama untuk deskriptor Pubchem. Pada skenario Lasso *selection*, nilai MSE turun dari 0.2241 menjadi 0.1598 dan RMSE turun dari 0.4734 menjadi 0.3997. Selain itu, MAE juga mengalami penurunan dari 0.3884 menjadi 0.3019 serta R² meningkat dari 0.2826 menjadi 0.4884. Hal ini menunjukkan bahwa seleksi fitur dengan Lasso membantu model dalam menjelaskan variansi data dengan lebih baik. Metrik EVS juga meningkat dari 0.2840 menjadi 0.4887 pada Lasso *selection* yang menandakan peningkatan kinerja model. Confidence Interval (CI) untuk Non-lasso *selection* menunjukkan rentang yang relatif sempit, misalnya untuk MSE berada di antara 0.2217 dan 0.2266 yang mengindikasikan estimasi yang cukup stabil. Di sisi lain, CI untuk Lasso *selection* sedikit lebih lebar dengan MSE antara 0.1551 dan 0.1645, namun masih menunjukkan kestabilan estimasi yang baik.

Pada deskriptor ECFP, perbedaan antara skenario Non-lasso *selection* dan Lasso *selection* tidak begitu signifikan. Nilai MSE hanya sedikit berubah dari



0.1485 menjadi 0.1496 dan RMSE sedikit meningkat dari 0.3853 menjadi 0.3867. MAE sedikit turun dari 0.2944 menjadi 0.2940 serta R^2 dan EVS menunjukkan peningkatan kecil yang masing-masing dari 0.5250 menjadi 0.5296, dan dari 0.5262 menjadi 0.5231. Confidence Interval untuk Non-lasso *selection* relatif sempit dengan MSE berada di antara 0.1438 dan 0.1532 yang menunjukkan ketebalan hasil yang baik. CI untuk Lasso *selection* sedikit lebih lebar dengan MSE antara 0.1434 dan 0.1558, namun masih menunjukkan estimasi yang relatif stabil.

Tabel 12 Metrik dengan Confidence Interval pada algoritma LGBM dengan deskriptor Pubchem Fingerprint

Metrik	Non-lasso <i>selection</i>			Lasso <i>selection</i>		
	<i>Mean</i>	95% Confidence Interval		<i>Mean</i>	95% Confidence Interval	
		<i>lower</i>	<i>upper</i>		<i>lower</i>	<i>upper</i>
MSE	0.2301	0.2262	0.2341	0.1499	0.1423	0.1574
RMSE	0.4797	0.4756	0.4838	0.3870	0.3773	0.3967
MAE	0.4031	0.3986	0.4076	0.2919	0.2852	0.2987
R^2	0.2636	0.2504	0.2767	0.5201	0.4852	0.5551
MAPE	0.0646	0.0639	0.0652	0.0467	0.0455	0.0478
MedAE	0.3799	0.3742	0.3856	0.2220	0.2169	0.2270
EVS	0.2637	0.2506	0.2768	0.5204	0.4856	0.5553

Tabel 13 Metrik dengan Confidence Interval pada algoritma LGBM dengan deskriptor ECFP

Metrik	Non-lasso <i>selection</i>			Lasso <i>selection</i>		
	<i>Mean</i>	95% Confidence Interval		<i>Mean</i>	95% Confidence Interval	
		<i>lower</i>	<i>upper</i>		<i>lower</i>	<i>upper</i>
MSE	0.2313	0.2270	0.2356	0.1496	0.1434	0.1558
RMSE	0.4809	0.4765	0.4854	0.3867	0.3787	0.3948
MAE	0.4049	0.4004	0.4094	0.2993	0.2911	0.3076
R^2	0.2615	0.2520	0.2710	0.5224	0.5051	0.5398
MAPE	0.0649	0.0640	0.0658	0.0480	0.0465	0.0494
MedAE	0.3809	0.3676	0.3942	0.2310	0.2225	0.2394
EVS	0.2617	0.2520	0.2715	0.5231	0.5055	0.5407

Hasil evaluasi model LightGBM (LGBM) menunjukkan bahwa penggunaan Lasso *selection* memberikan peningkatan yang signifikan pada beberapa metrik, terutama pada deskriptor Pubchem Fingerprint. Pada skenario Lasso *selection*, MSE menurun dari 0.2301 menjadi 0.1499, RMSE turun dari 0.4797 menjadi 0.3870, dan MAE berkurang dari 0.4031 menjadi 0.2919. Selain itu, R^2 juga meningkat dari 0.2636 menjadi 0.5201 yang menunjukkan bahwa seleksi fitur dengan Lasso membantu model dalam menjelaskan variansi data lebih baik. Confidence interval pada Non-lasso *selection* lebih sempit sehingga menunjukkan



estimasi yang stabil. Sedangkan pada Lasso *selection*, rentang CI sedikit lebih lebar namun tetap menunjukkan kestabilan yang baik.

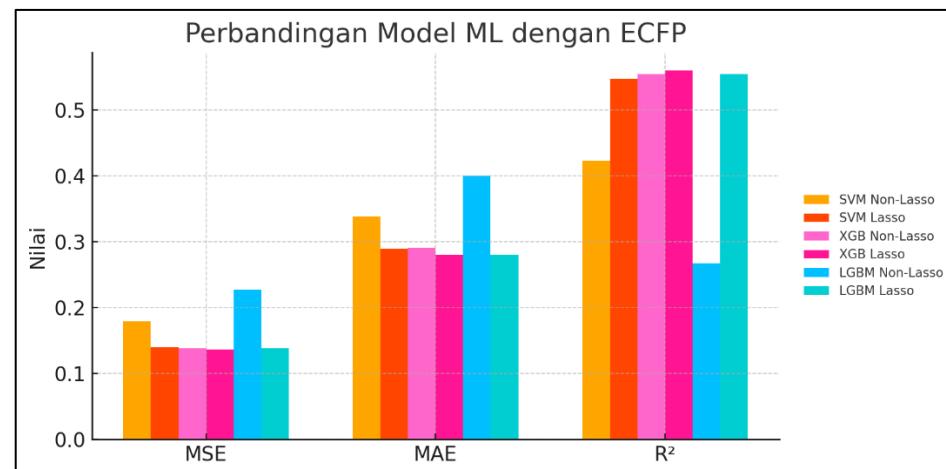
Pada deskriptor ECFP, hasil yang serupa terlihat meskipun perbedaannya tidak sebesar pada Pubchem Fingerprint. MSE turun dari 0.2313 menjadi 0.1496, RMSE menurun dari 0.4809 menjadi 0.3867, dan MAE turun dari 0.4031 menjadi 0.2915. R^2 meningkat dari 0.2615 menjadi 0.5224 meskipun tidak sebesar pada deskriptor Pubchem Fingerprint. Confidence Interval untuk Non-lasso *selection* lebih lebar dibandingkan dengan Lasso *selection*, tetapi tetap menunjukkan kestabilan yang baik. Secara keseluruhan, meskipun Lasso *selection* memberikan peningkatan pada kedua deskriptor, dampaknya lebih terlihat jelas pada Pubchem Fingerprint dibandingkan ECFP.

Setelah *cross validation* pada data latih, model dilatih ulang menggunakan seluruh dataset pelatihan untuk memaksimalkan pemanfaatan data. Model kemudian diuji dengan data *testing* yang terpisah untuk mengevaluasi kinerjanya pada data yang belum pernah dilihat sebelumnya guna memastikan kemampuan generalisasi model.

Tabel 14 Hasil evaluasi model ECFP

Metrik	SVM		XGB		LGBM	
	Non-lasso <i>selection</i>	Lasso <i>selection</i>	Non-lasso <i>selection</i>	Lasso <i>selection</i>	Non- lasso <i>selection</i>	Lasso <i>selection</i>
MSE	0.1790	0.1404	0.1381	0.1365	0.2272	0.1381
RMSE	0.4231	0.3746	0.3716	0.3694	0.4766	0.3716
MAE	0.3387	0.2892	0.2914	0.2800	0.4000	0.2803
R^2	0.4227	0.5474	0.5547	0.5599	0.2674	0.5548
MAPE	0.0539	0.0462	0.0466	0.0447	0.0640	0.0447
MedAE	0.2900	0.2225	0.2402	0.2174	0.3774	0.2152
EVS	0.4243	0.5474	0.5547	0.5599	0.2676	0.5548

Hasil evaluasi menunjukkan bahwa XGB (XGB) dengan Lasso *selection* memberikan performa terbaik, ditandai dengan nilai MSE dan MAE terendah (0.1365 dan 0.2800) serta nilai R^2 dan EVS tertinggi (0.5599). Secara umum, penggunaan Lasso *selection* terbukti mampu meningkatkan akurasi model dengan mengurangi fitur yang tidak relevan atau redundant. Peningkatan performa paling signifikan terlihat pada LGBM, di mana MSE turun dari 0.2272 menjadi 0.1381, dan R^2 meningkat dari 0.2674 menjadi 0.5548 setelah dilakukan seleksi fitur. Model SVM juga mengalami peningkatan serupa dengan MSE menurun dari 0.1790 menjadi 0.1404, dan R^2 meningkat dari 0.4227 menjadi 0.5474.

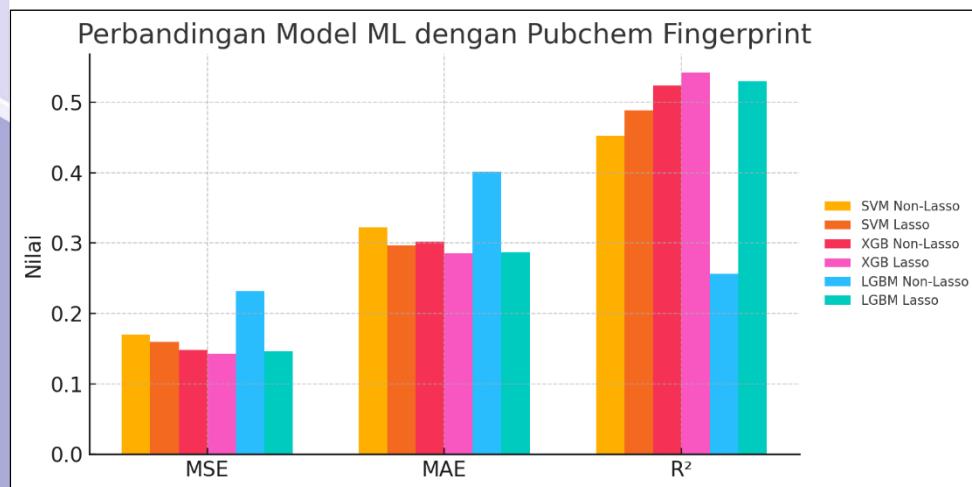


Gambar 13 Histogram perbandingan model dengan deskriptor ECFP

Tabel 15 Hasil evaluasi model Pubchem Fingerprint

Metrik	SVM		XGB		LGBM	
	Non-lasso selection	Lasso selection	Non-lasso selection	Lasso selection	Non-lasso selection	Lasso selection
MSE	0.1704	0.1592	0.1481	0.1426	0.2315	0.1463
RMSE	0.4128	0.3990	0.3849	0.3776	0.4812	0.3825
MAE	0.3221	0.2966	0.3020	0.2854	0.4017	0.2874
R ²	0.4523	0.4883	0.5238	0.5417	0.2558	0.5297
MAPE	0.0515	0.0475	0.0484	0.0457	0.0644	0.0460
MedAE	0.2604	0.2213	0.2502	0.2140	0.3738	0.2159
EVS	0.4524	0.4884	0.5240	0.5418	0.2558	0.5298

Representasi Pubchem Fingerprint yang terdiri dari 881 fitur menghasilkan performa terbaik pada model XGB dengan Lasso *selection*. Model ini mencatat nilai MSE sebesar 0.1426, MAE sebesar 0.2854, dan R² sebesar 0.5417 yang menunjukkan tingkat akurasi prediksi yang tinggi. Penerapan Lasso *selection* secara konsisten meningkatkan performa seluruh model. Sebagai contoh pada model LGBM, nilai R² meningkat dari 0.2558 menjadi 0.5298 setelah dilakukan seleksi fitur dan nilai MSE menurun dari 0.2315 menjadi 0.1463. Peningkatan serupa juga terjadi pada model SVM, baik dalam penurunan error maupun kenaikan nilai R². Meskipun jumlah fitur pada Pubchem lebih sedikit dibandingkan ECFP, seleksi fitur tetap memberikan dampak signifikan. Hal ini menunjukkan bahwa dalam representasi Pubchem terdapat sejumlah fitur yang tidak relevan dan dapat dieliminasi oleh metode Lasso.



Gambar 14 Histogram perbandingan model dengan deskriptor Pubchem Fingerprint

Nilai evaluasi yang diperoleh pada model terbaik melebihi *benchmark* yang umum ditemukan pada perangkat lunak prediksi kimia seperti DataWarrior ($R^2 \approx 0.5251$). Selain itu, penelitian oleh Simeon dan Jongkon (2019) menyatakan bahwa model QSAR yang dapat diterima harus memenuhi syarat $R^2 > 0.5$. Oleh karena itu, model yang dihasilkan dapat dianggap memiliki akurasi prediktif yang kompetitif untuk studi QSAR. Untuk mengilustrasikan performa model secara lebih konkret, Tabel 30 menyajikan lima contoh senyawa beserta nilai IC₅₀ aktual dan hasil prediksi model. Pemilihan senyawa dilakukan secara acak dari data uji.

Tabel 30 Contoh data aktual dan prediksi model terbaik

No	SMILES	Nilai IC ₅₀ Aktual	Nilai IC ₅₀ Prediksi
1	C=Cc1cccc(CN2C[C@ @H](CN(C)CC)[C@ @H])(CO)C2)c1	5,978137	5,978009
2	Cc1ccc2c(c1)CN(C(=O)C1CCCCC1)C(CN(C)C)CO2.Cl	6,055409	6,045037
3	O=C1c2cccc2SC(c2ccc(Cl)cc2)N1CCCCNc1cc nc2cc(Cl)ccc12	5,895513	5,945778
4	Clc1ccc2c(NCCN(CCC34CC5CC(CC(C5)C3)C 4)CCC34CC5CC(CC(C5)C3)C4)ccnc2c1	6,996927	6,814641
5	Cc1cc(Nc2cccc(F)c2)nc(NCc2cccc2)n1	6,264561	6,067737

Sebagai bagian dari evaluasi tambahan, data senyawa dikelompokkan berdasarkan nilai IC₅₀ serta akurasi, *precision*, *recall*, dan F1-score dihitung dengan membandingkan pengelompokan nilai IC₅₀ antara data uji aktual dan data uji prediksi, sesuai dengan metodologi yang digunakan dalam penelitian oleh Indrayanto *et al.* (2021).

Tabel 31 Hasil evaluasi pengelompokan data dengan deskriptor ECFP

Metrik	SVM		XGB		LGBM	
	Non-lasso selection	Lasso selection	Non-lasso selection	Lasso selection	Non-lasso selection	Lasso selection
Accuracy	0.8078	0.8066	0.8115	0.8149	0.6634	0.8149
Precision	0.8048	0.8037	0.8100	0.8121	0.7437	0.8121
Recall	0.8078	0.8066	0.8115	0.8149	0.6634	0.8149
F1-score	0.8034	0.8019	0.8052	0.8113	0.5468	0.8118

Tabel 31 menunjukkan hasil evaluasi pengelompokan data menggunakan deskriptor ECFP pada model SVM, XGB, dan LGBM dengan dua metode seleksi fitur, yaitu non-Lasso dan Lasso. Secara umum, penggunaan Lasso *selection* meningkatkan kinerja model. Pada model SVM, terjadi penurunan performa model dimana *accuracy* non-Lasso adalah 0.8078 dan Lasso adalah 0.8066. Kemudian, model XGB menunjukkan *accuracy* non-Lasso 0.8115 dan Lasso 0.8149. Selain itu, model LGBM memiliki *accuracy* non-Lasso 0.6634 dan Lasso 0.8149. Peningkatan serupa juga terlihat pada metrik lainnya untuk model XGB dan LGBM, seperti *precision*, *recall*, serta F1-score, yang menunjukkan bahwa Lasso *selection* berkontribusi pada peningkatan kinerja model dalam pengelompokan data.

Tabel 32 Hasil evaluasi pengelompokan data dengan deskriptor Pubchem Fingerprint

Metrik	SVM		XGB		LGBM	
	Non-lasso selection	Lasso selection	Non-lasso selection	Lasso selection	Non-lasso selection	Lasso selection
Accuracy	0.7136	0.8103	0.8006	0.8153	0.6686	0.8125
Precision	0.7092	0.8081	0.8072	0.8149	0.7295	0.8104
Recall	0.7136	0.8103	0.8006	0.8153	0.6686	0.8125
F1-score	0.6869	0.8060	0.7886	0.8094	0.5703	0.8082

Tabel 32 menunjukkan hasil evaluasi pengelompokan data menggunakan deskriptor Pubchem Fingerprint pada model SVM, XGB, dan LGBM dengan dua metode seleksi fitur, yaitu non-Lasso dan Lasso. Berdasarkan hasil evaluasi, penggunaan Lasso *selection* memberikan peningkatan pada seluruh metrik (*accuracy*, *precision*, *recall*, dan F1-score) untuk ketiga model. Pada model SVM, *accuracy* meningkat dari 0.7136 (non-Lasso) menjadi 0.8103 (Lasso). Kemudian pada XGB, *accuracy* meningkat dari 0.8006 (non-Lasso) menjadi 0.8153 (Lasso). Selain itu pada LGBM, *accuracy* meningkat dari 0.6686 (non-Lasso) menjadi 0.8125 (Lasso). Peningkatan serupa juga terlihat pada metrik *precision*, *recall*, dan F1-score, yang menunjukkan bahwa Lasso *selection* berkontribusi dalam meningkatkan kinerja model pada seluruh metrik di ketiga model (SVM, XGB, dan LGBM) dalam pengelompokan data dengan deskriptor Pubchem Fingerprint.



4.9 Prediksi Senyawa Herbal

Tabel 33 Hasil uji dengan sampel senyawa herbal yang diperoleh melalui basis data Pubchem

@Hak cipta milik IPB University	CID	Nama Senyawa	Sumber	Prediksi pIC ₅₀	Kategori
	969516	Curcumin	Akar kunyit (<i>Curcuma longa</i>)	5.873	Aktivitas baik atau sitotoksitas sangat kuat
	5281792	Rosmarinic Acid	Rosemary (<i>Rosmarinus officinalis</i>) dan basil (<i>Ocimum basilicum</i>)	5.935	Aktivitas baik atau sitotoksitas sangat kuat
	2353	Berberine	<i>Berberis vulgaris</i> dan <i>Coptis chinensis</i>	6.064	Aktivitas sangat baik atau poten
	445154	Resveratrol	Anggur merah (<i>Vitis vinifera</i>) dan tanaman <i>Polygonum cuspidatum</i> <i>Mangifera indica</i> (mangga)	5.911	Aktivitas baik atau sitotoksitas sangat kuat
	5280804	Isoquercetin	dan <i>Camellia sinensis</i> (teh hijau)	5.438	Aktivitas baik atau sitotoksitas sangat kuat
	45485025	Hordatine A	Barley (<i>Hordeum vulgare</i>)	5.842	Aktivitas baik atau sitotoksitas sangat kuat
	23915	Schisandrin	<i>Schisandra chinensis</i>	5.807	Aktivitas baik atau sitotoksitas sangat kuat
	10228	Osthole	<i>Cnidium monnieri</i>	5.835	Aktivitas baik atau sitotoksitas sangat kuat

Tabel 13 menunjukkan hasil uji beberapa senyawa herbal yang diperoleh melalui basis data PubChem. Senyawa-senyawa seperti curcumin, rosmarinic acid, dan berberine menunjukkan nilai pIC₅₀ yang bervariasi, mulai dari 5.437886 hingga 6.0641026. Hal ini mengindikasikan potensi aktivitas yang baik hingga sangat kuat dalam menghambat aktivitas protein parasit malaria. Semua senyawa yang diuji menunjukkan hasil yang menjanjikan dalam hal penghambatan pertumbuhan parasit malaria.

V SIMPULAN DAN SARAN

5.1 Simpulan

Penapisan senyawa (*compound screening*) merupakan tahap krusial dalam penemuan dan pengembangan obat, di mana ribuan hingga jutaan kandidat senyawa dievaluasi untuk mengidentifikasi molekul dengan potensi aktivitas biologis yang tinggi. Dalam penelitian ini, proses penapisan difokuskan pada pengembangan model prediksi senyawa antimalaria dengan mengestimasi nilai IC_{50} sebagai indikator potensi inhibisi terhadap parasit *Plasmodium*. Pendekatan *in silico* dengan pemodelan *machine learning* digunakan untuk mempercepat proses penapisan serta mengurangi ketergantungan pada uji laboratorium yang memakan waktu dan biaya tinggi. Representasi molekul menggunakan deskriptor ECFP dan PubChem Fingerprint dimanfaatkan untuk mengekstraksi informasi struktural yang relevan yang kemudian diolah oleh berbagai algoritma pembelajaran mesin untuk memprediksi nilai IC_{50} secara kuantitatif.

Berdasarkan hasil evaluasi model *machine learning* terhadap dua jenis representasi molekul yaitu ECFP dan Pubchem Fingerprint, dapat disimpulkan bahwa model XGB dengan seleksi fitur menggunakan Lasso konsisten memberikan performa prediktif terbaik. Pada representasi ECFP, model ini mencapai nilai MSE sebesar 0.1365, MAE sebesar 0.2800, serta nilai R^2 dan EVS sebesar 0.5599 sehingga menunjukkan kemampuan prediktif yang baik dalam memodelkan hubungan antara fitur molekul dan target. Sementara itu, pada representasi Pubchem Fingerprint, kombinasi yang sama menghasilkan nilai MSE sebesar 0.1426, MAE sebesar 0.2854, serta R^2 dan EVS sebesar 0.5417 dan 0.5418 yang secara konsisten menegaskan keunggulan kombinasi XGB dan Lasso. Secara umum, penggunaan teknik Lasso *selection* terbukti meningkatkan performa seluruh model yang diuji karena mampu mengurangi *noise* dari fitur yang tidak relevan dan membantu model belajar lebih efektif. Di sisi lain, model LGBM tanpa seleksi fitur menunjukkan performa terburuk dengan nilai MSE sebesar 0.2272, MAE sebesar 0.4000, serta R^2 hanya 0.2674 pada representasi ECFP. Pada deskriptor Pubchem Fingerprint, model LGBM juga menjadi yang terburuk dengan MSE sebesar 0.2315, MAE sebesar 0.4017, dan R^2 sebesar 0.2558. Temuan ini menunjukkan bahwa pemilihan model yang tepat, strategi seleksi fitur yang efektif, dan representasi fitur yang sesuai merupakan faktor krusial dalam membangun model prediktif yang andal dalam domain *cheminformatics*.

5.2 Saran

Sebagai tindak lanjut dari penelitian ini, disarankan untuk mengeksplorasi metode representasi molekul lainnya, seperti Graph Neural Network (GNN), Molecular Embedding (seperti Mol2Vec), atau deskriptor fisikokimia guna membandingkan efektivitasnya terhadap model prediktif. Selanjutnya, penggunaan metode seleksi fitur alternatif seperti Recursive Feature Elimination (RFE) dapat dipertimbangkan untuk membandingkan efektivitas seleksi fitur terhadap performa akhir model.



DAFTAR PUSTAKA

- Abdi AI, Achcar F, Sollelis L, Silva-Filho JL, Mwikali K, Muthui M, Mwangi S, Kimingi HW, Orindi B, Andisi Kivisi C, *et al.* 2023. Plasmodium falciparum adapts its investment into replication versus transmission according to the host environment. *Elife*. 12:1–20. doi:10.7554/elife.85140.
- Alibrahim H, Ludwig SA. 2021. Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. 2021 IEEE Congr Evol Comput CEC 2021 - Proc., siap terbit.
- Andronescu LR, Buchwald AG, Sharma A, Bauleni A, Mawindo P, Liang Y, Gutman JR, Mathanga DP, Chinkhumba J, Laufer MK. 2023. Plasmodium falciparum infection and disease in infancy associated with increased risk of malaria and anaemia in childhood. *Malar J*. 22(1):1–10. doi:10.1186/s12936-023-04646-8.
- Asahara R, Miyao T. 2022. Extended Connectivity Fingerprints as a Chemical Reaction Representation for Enantioselective Organophosphorus-Catalyzed Asymmetric Reaction Prediction. *ACS Omega*. 7(30):26952–26964. doi:10.1021/acsomega.2c03812.
- Aykul S, Hackert EM. 2016. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Anal Biochem*. 508:97–103. doi:10.1016/j.ab.2016.06.025.
- Babatunde, Surajudeen A Olaniyan OM, Owolabi TO. 2025. Application of real-coded genetic algorithm to the optimization of support vector regression in the modeling and characterization of hydrocarbon reservoirs. *Niger J Appl Sci Innov Technol*. 1(1):122–138.
- Badwan BA, Liaropoulos G, Kyrodimos E, Skaltsas D, Tsirigos A. 2023. Machine learning approaches to predict drug efficacy and toxicity in oncology. *Cell Reports Methods*. 3(2):100413. doi:10.1016/j.crmeth.2023.100413.
- Bammingher K, Pichler V, Vraka C, Nehring T, Pallitsch K, Lieder B, Hacker M, Wadsak W. 2023. On the road towards small-molecule programmed cell death 1 ligand 1 positron emission tomography tracers: a ligand-based drug design approach. *Pharmaceuticals*. 16(7). doi:10.3390/ph16071051.
- Dablander M, Hanser T, Lambotte R, Morris GM. 2024. Sort & Slice: a simple and superior alternative to hash-based *folding* for extended-connectivity fingerprints. *J Cheminform*. 16(1). doi:10.1186/s13321-024-00932-y.
- Danishuddin, Madhukar G, Malik MZ, Subbarao N. 2019. Development and rigorous validation of antimalarial predictive models using machine learning approaches. *SAR QSAR Environ Res*. 30(8):543–560. doi:10.1080/1062936X.2019.1635526.
- Díaz-Eufracio BI, Medina-Franco JL. 2022. Machine Learning Models to Predict

Protein–Protein Interaction Inhibitors. *Molecules.* 27(22). doi:10.3390/molecules27227986.

Dong J, Cao DS, Miao HY, Liu S, Deng BC, Yun YH, Wang NN, Lu AP, Zeng W Bin, Chen AF. 2015. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J Cheminform.* 7(1):1–10. doi:10.1186/s13321-015-0109-z.

Elbadawi M, Gaisford S, Basit AW. 2021. Advanced machine-learning techniques in drug discovery. *Drug Discov Today.* 26(3):769–777. doi:10.1016/j.drudis.2020.12.003.

Fernández-De Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. 2017. Database fingerprint (DFP): an approach to represent molecular databases. *J Cheminform.* 9(1):1–9. doi:10.1186/s13321-017-0195-1.

Fitriany J, Sabiq A. 2018. Malaria. *J Averrouus.* 4(2):83. doi:doi.org/10.29103/averrous.v4i2.1039.

Gan M, Pan S, Chen Y, Cheng C, Pan H, Zhu X. 2021. Application of the machine learning lightgbm model to the prediction of the water levels of the lower columbia river. *J Mar Sci Eng.* 9(5). doi:10.3390/jmse9050496.

Indrayanto G, Putra GS, Suhud F. 2021. *Validation of in-vitro bioassay methods: Application in herbal drug research.* Ed ke-1. Volume ke-46. Elsevier Inc.

Kore M, Acharya D, Sharma L, Vembar SS, Sundriyal S. 2025. Development and experimental validation of a machine learning model for the prediction of new antimalarials. *BMC Chem.* 19(1). doi:10.1186/s13065-025-01395-4.

Kumar S, Bhowmik R, Oh JM, Abdelgawad MA, Ghoneim MM, Hamed R, Serwi A, Kim H, Mathew B. 2024. Machine learning driven web - based app platform for the discovery of monoamine oxidase B inhibitors. *Sci Rep.*, siap terbit.

Li H, Tan Q, Deng J, Dong B, Li B, Guo J, Zhang S, Bai W. 2023. A Comprehensive Prediction Method for Pore Pressure in Abnormally High-Pressure Blocks Based on Machine Learning. *Processes.* 11(9). doi:10.3390/pr11092603.

Masliyana A, Mas DA, Putra J. 2018. Penapisan maya metabolit sekunder tanaman obat indonesia terhadap xantin oksidase virtual screening secondary metabolites of indonesian drug plants againts xantin oksidase. *Indones Nat Res Pharm J.* 3(1):2502–8421. <https://pubchem.ncbi.nlm.nih.gov/>.

Milner DA. 2018. Malaria pathogenesis. *Cold Spring Harb Perspect Med.* 8(1):1–12. doi:10.1101/cshperspect.a025569.

Moriwaki H, Tian YS, Kawashita N, Takagi T. 2018. Mordred: A molecular descriptor calculator. *J Cheminform.* 10(1):1–14. doi:10.1186/s13321-018-0258-y.

Mswahili ME, Martin GL, Woo J, Choi GJ, Jeong YS. 2021. Antimalarial drug



predictions using molecular descriptors and machine learning against plasmodium falciparum. *Biomolecules*. 11(12):1–15. doi:10.3390/biom11121750.

Muthukrishnan R, Rohini R. 2017. LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE Int Conf Adv Comput Appl ICACA 2016.*, siap terbit.

Naveed S, Husnain M. 2025. A drug recommendation system based on response prediction: Integrating gene expression and K-mer fragmentation of drug SMILES using LightGBM. *Intell Med.* 11 December 2024:100206. doi:10.1016/j.ibmed.2025.100206.

Oliveira T, Silva M, Maia E, Silva A, Taranto A. 2023. Virtual screening algorithms in drug discovery: a review focused on machine and deep learning methods. *Drugs Drug Candidates*. 2(2):311–334. doi:10.3390/ddc2020017.

Pandey SK, Anand U, Siddiqui WA, Tripathi R. 2023. Drug Development Strategies for Malaria: With the Hope for New Antimalarial Drug Discovery. *Adv Med.* 2023:1–10. doi:10.1155/2023/5060665.

Ribeiro G de JG, Rei Yan SL, Palmisano G, Wrenger C. 2023. Plant Extracts as a Source of Natural Products with Potential Antimalarial Effects: An Update from 2018 to 2022. *Pharmaceutics*. 15(6):1–19. doi:10.3390/pharmaceutics15061638.

Roihan A, Sunarya PA, Rafika AS. 2020. Pemanfaatan machine learning dalam berbagai bidang : Review paper. *Indones J Comput Inf Technol.* 5 April:75–82. doi:10.31294/ijcit.v5i1.7951.

Sánchez-Díez M, Romero-Jiménez P, Alegría-Aravena N, Gavira-O'Neill CE, Vicente-García E, Quiroz-Troncoso J, González-Martos R, Ramírez-Castillejo C, Pastor JM. 2025. Assessment of Cell Viability in Drug Therapy: IC50 and Other New Time-Independent Indices for Evaluating Chemotherapy Efficacy. *Pharmaceutics*. 17(2). doi:10.3390/pharmaceutics17020247.

Shahri NHNB, Lai SBS, Mohamad MB, Rahman HABA, Rambli A Bin. 2021. Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data. *Math Stat.* 9(3):379–385. doi:10.13189/ms.2021.090320.

Shibeshi MA, Kifle ZD, Atnafie SA. 2020. Antimalarial drug resistance and novel targets for antimalarial drug discovery. *Infect Drug Resist.* 13:4047–4060. doi:10.2147/IDR.S279433.

Simeon S, Jongkon N. 2019. Construction of quantitative structure activity relationship (QSAR) models to predict potency of structurally diverse janus kinase 2 inhibitors. *Molecules*. 24(23). doi:10.3390/molecules24234393.

Talapko J, Skrlec I, Alebic T, Jukic M, Vcev A. 2019. Malaria: the past and the present. *Clin Sci.* 179(7):1–17. doi:10.3390/microorganisms7060179.



- Ungureanu N. 2025. Mechanical and Thermal Behavior of Hemp-Reinforced Starch / Agar Biocomposites : Insights from Finite Element Simulation and Machine Learning Models. *Polymers (Basel)*. 17:1–22. doi:<https://doi.org/10.3390/polym17070855>.
- Vianetha Prima Snak E, Nyoman Wande I, Nyoman Mahartini N. 2023. Severe falciparum malaria with multiple complications in Sanglah Hospital Denpasar. *Indones J Clin Pathol Med Lab.* 29(2):206–210. doi:<https://doi.org/10.24293/ijcpml.v29i2.1830>.
- Villar-delfino PH, Christo PP, Maria C, Volpe O. 2025. Drug Repurposing and Artificial Intelligence in Multiple Sclerosis : Emerging Strategies for Precision Therapy. *Sclerosis*. 3:1–13. doi:<https://doi.org/10.3390/sclerosis3030028>.
- Wang J, Wang Q, Zhou J, Wang X, Cheng L. 2018. Operation space design of microbial fuel cells combined anaerobic–anoxic–oxic process based on support vector regression inverse model. *Eng Appl Artif Intell.* 72 April:340–349. doi:[10.1016/j.engappai.2018.04.005](https://doi.org/10.1016/j.engappai.2018.04.005).
- William IO, Altamimi EM. 2024. Light Gradient Boosting Machine (LGBM) for Daily Solar Radiation Prediction Using Claude3. *3rd Int Conf Front Acad Res.* June:1–10.
- Xue H, Zhang R, Yan X, Wang R, Zhang P. 2024. Study of PARP inhibitors for breast cancer based on enhanced multiple kernel function SVR with PSO. *Front Pharmacol.* 15 February:1–22. doi:[10.3389/fphar.2024.1257253](https://doi.org/10.3389/fphar.2024.1257253).

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

LAMPIRAN



Lampiran 1 Daftar 30 fitur terbaik pada deskriptor Pubchem Fingerprint

Fitur	Representasi Fitur pada Senyawa	Skor Importance
Pub750	Nc1cc(Cl)ccc1	0.117291
Pub601	N-C:C:C-N	0.058639
Pub309	O-O	0.038260
Pub514	O-N-C-C	0.036241
Pub16	≥ 4 N	0.034295
Pub19	≥ 2 O	0.026026
Pub248	≥ 1 cincin karbon jenuh atau aromatik berukuran 10 atom	0.025317
Pub636	C-N-C-N-C	0.024224
Pub643	[#1]-C-C-N-[#1]	0.023755
Pub696	C-C-C-C-C-C-C	0.023639
Pub697	C-C-C-C-C(C)-C	0.019946
Pub528	[#1]-N-C-[#1]	0.019622
Pub593	N-C-C-C-N	0.015989
Pub569	N-C-C-N	0.014944
Pub13	≥ 32 C	0.014073
Pub3	≥ 32 H	0.013965
Pub143	≥ 1 cincin berukuran 5 atom	0.013701
Pub667	C=C-C-O-[#1]	0.013619
Pub338	C(~C)(~C)(~H)(~N)	0.012955
Pub438	C(-C)(-N)(=N)	0.012655
Pub496	N:N-C-[#1]	0.011795
Pub571	[#1]-C-O-[#1]	0.011348
Pub535	O=C-C-C	0.010563
Pub488	N-C=N-[#1]	0.010540
Pub374	C(~H)(~H)(~H)	0.010111
Pub20	≥ 4 O	0.010096
Pub716	Cc1ccc(N)cc1	0.009956
Pub37	≥ 1 Cl	0.009593
Pub454	N(-C)(=O)	0.009314

Lampiran 2 Daftar 30 fitur terbaik pada deskriptor ECFP

Fitur	Skor Importance
ecfp38	0.025974
ecfp1097	0.023231
ecfp103	0.020107
ecfp1733	0.019576

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

ecfp722	0.015856
ecfp1911	0.011919
ecfp650	0.011033
ecfp1917	0.010369
ecfp1152	0.010040
ecfp251	0.009842
ecfp1162	0.009503
ecfp1143	0.009288
ecfp1062	0.008105
ecfp1453	0.007360
ecfp561	0.006367
ecfp1160	0.006132
ecfp1750	0.005846
ecfp1631	0.005569
ecfp1467	0.005160
ecfp701	0.005061
ecfp784	0.004883
ecfp118	0.004879
ecfp1984	0.004848
ecfp1104	0.004801
ecfp371	0.004791
ecfp1475	0.004720
ecfp1964	0.004674
ecfp388	0.004552
ecfp881	0.004552
ecfp1392	0.004452

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

RIWAYAT HIDUP

Penulis dilahirkan di Kota Metro, Provinsi Lampung pada 23 Februari 2003 sebagai anak ke dua dari pasangan bapak Agus Marianto dan ibu Sulistiawati. Pendidikan sekolah menengah atas (SMA) ditempuh di Sekolah Menengah Atas (SMA) Negeri 1 Metro , dan lulus pada tahun 2021. Pada tahun tersebut, penulis diterima sebagai mahasiswa program sarjana (S-1) di Program Studi Ilmu Komputer, pada Sekolah Sains Data, Matematika, dan Informatika IPB.

Selama menjalani program S-1, penulis aktif berperan sebagai staf kewirausahaan di Himpunan Mahasiswa Ilmu Komputer (HIMALKOM), yang memberikan kesempatan untuk mengembangkan keterampilan manajerial, organisasi, dan kepemimpinan dalam konteks kewirausahaan. Selain itu, penulis juga mengikuti program studi independen Bangkit Academy yang berfokus pada pembelajaran mesin (*machine learning*), yang memperdalam pengetahuan dalam penerapan algoritma dan model prediksi untuk analisis data. Dalam mempersiapkan tugas akhir, penulis juga memperluas pengalaman praktis dengan menjalani magang di tiga lembaga/perusahaan terkemuka, yaitu Center for Transdisciplinary and Sustainability Sciences (CTSS), PT. Asuransi Jasa Indonesia (Jasindo), dan PT. Tunas Ridean Tbk. Pengalaman ini mendorong penulis untuk menerapkan konsep-konsep ilmiah dalam konteks dunia industri, serta memperkuat pemahaman mengenai penerapan metode statistik dan teknik analisis data dalam berbagai sektor.

