Jurnal Algoritma
E-ISSN: 2302-7339 | P-ISSN: 1412-3622
INSTITUT TEKNOLOGI GARUT

https://jurnal.itg.ac.id/index.php/algoritma
DOI: 10.33364/algoritma/v.21-1.1598

Investigasi Model *Machine Learning* Regresi Pada Senyawa Obat Sebagai *Inhibitor* Korosi

Muhammad Reesa Rosyid[1], Lubna Mawaddah[2], Muhamad Akrom[3*]
[1,2]Universitas Dian Nuswantoro, Indonesia
[3]Research Center for Materials Informatics Universitas Dian Nuswantoro, Indonesia

***email:** m.akrom@dsn.dinus.ac.id

Info Artikel
Dikirim: 12 Januari 2024
Diterima: 31 Mei 2024
Diterbitkan: 31 Mei 2024

Kata kunci:
Hyperparameter Tuning;
Inhibitor Korosi;
Machine Learning;
Model Prediksi;
Senyawa Obat.

ABSTRAK

Korosi merupakan tantangan signifikan bagi daya tahan material, yang seringkali menyebabkan kerugian ekonomi yang besar. Penelitian ini memanfaatkan teknik Machine Learning (ML) untuk memprediksi efektivitas senyawa obat sebagai inhibitor korosi. Kami menggunakan lima algoritma ML yang unggul: Regresi Linear, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, dan XGBoost. Model-model ini dilatih dan dievaluasi menggunakan dataset yang terdiri dari 14 fitur molekuler dengan efisiensi inhibisi korosi (IE%) sebagai variabel target. Hasil pelatihan model awal mengidentifikasi Random Forest dan XGBoost sebagai model yang berkinerja terbaik berdasarkan metrik Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan R-squared (R^2). Penyetelan hiperparameter lebih lanjut menggunakan GridSearchCV menunjukkan bahwa XGBoost, setelah penyetelan, secara signifikan mengungguli model lainnya, mencapai kesalahan terendah dan nilai R^2 tertinggi, menunjukkan akurasi prediktif yang superior untuk kasus ini. Temuan ini menegaskan potensi ML, khususnya XGBoost, dalam meningkatkan pemodelan prediktif inhibitor korosi, sehingga memberikan wawasan berharga bagi bidang ilmu korosi.

1. PENDAHULUAN

Pelemahan logam sering terjadi melalui korosi, yang muncul dari interaksi elektrokimia antara permukaan logam dan kondisi korosif di sekitarnya [1], [2], [3]. Seiring waktu, korosi secara bertahap mengurangi masa pakai material, seringkali lebih pendek dari yang diperkirakan [4]. Kerugian global akibat korosi, yang diperkirakan mencapai US$2,5 triliun pada tahun 2013, mewakili 3,4% dari PDB global, namun penghematan biaya yang signifikan, berkisar antara 15 hingga 35%, dapat dicapai dengan menerapkan praktik pengendalian korosi yang telah terbukti [5], [6], [7]. Penggunaan inhibitor korosi merupakan pendekatan paling efisien untuk mencegah korosi logam [8], [9], [10]. Karena adanya heteroatom seperti nitrogen (N), fosfor (P), sulfur (S), arsenik (As), atau oksigen (O) dalam struktur molekulnya, serta elektron bebas atau π dalam cincin aromatik atau ikatan ganda, senyawa organik sering kali berfungsi sebagai inhibitor korosi yang efektif dengan membentuk lapisan pelindung pada permukaan logam melalui interaksi kuat dengan atom logam dan molekul organik, sehingga dinilai sebagai inhibitor potensial di antara banyak zat farmasi karena kesamaan strukturalnya [11], [12], [13].

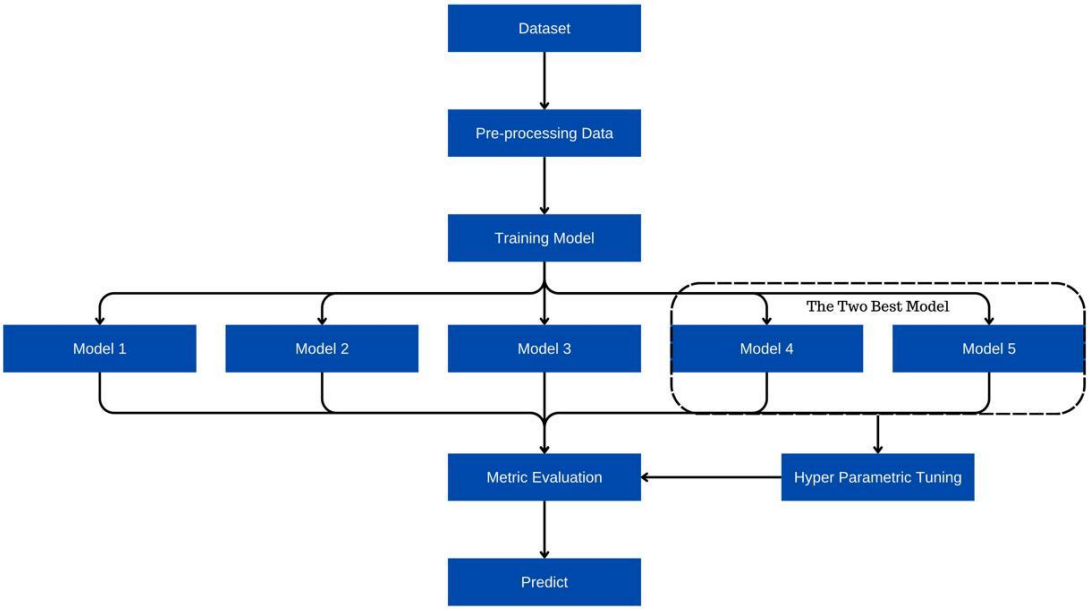
Penelitian tentang inhibitor korosi memerlukan investasi yang besar dalam hal waktu, biaya, dan eksperimen. Kemajuan teknologi dan ilmu data telah mendorong munculnya informatika material, terutama penggunaan machine learning (ML), sebagai fokus utama dalam menjelajahi bahan-bahan baru [14]. Dalam bidang informatika material, upaya untuk menemukan bahan antikorosi, khususnya inhibitor korosi, merupakan area yang kompleks dan berkembang pesat. Penggunaan metodologi hubungan kuantitatif struktur-aktivitas atau struktur-properti (QSAR/QSPR), yang didukung oleh ML, telah menjadi pendekatan yang dapat diandalkan

untuk menguraikan korelasi rumit antara atribut struktural senyawa kimia dan fungsi biologisnya [15]. Penerapan metode QSAR atau QSPR, yang difasilitasi oleh algoritma ML seperti Regresi Linear, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, dan XGBoost, merupakan teknik yang dapat diandalkan untuk mengungkap hubungan kompleks antara karakteristik struktural senyawa kimia dan fungsi biologisnya dalam upaya menemukan bahan antikorosi [16], [17].

Berbagai penelitian telah dilakukan untuk mengembangkan model prediktif efisiensi inhibisi korosi. Perez et al. [1] mengembangkan model QSAR-ARX untuk memprediksi efektivitas inhibisi korosi dari obat kadaluarsa pada permukaan baja, mencapai nilai MSE sebesar 49.47 dan RMSE sebesar 7.03. Quadri et al. [18] menggunakan 20 turunan piridazin untuk mengembangkan model prediktif antikorosi dengan model ANN menghasilkan MSE sebesar 111.59 dan RMSE sebesar 10.56. Liu et al. [19] mengembangkan model SVM untuk mengevaluasi turunan benzimidazol sebagai inhibitor korosi, mencapai nilai R^2 sebesar 0.9589 dan RMSE sebesar 4.45. Pham et al. [20] mengembangkan model prediksi efisiensi inhibisi korosi senyawa organik pada baja karbon menggunakan algoritma gradient boosting decision tree (GB) dan teknik permutation feature importance (PFI), mencapai RMSE sebesar 6.40, MAE sebesar 4.80, dan R^2 sebesar 0.72. Ser et al. [21] mengembangkan model QSPR linier dan non-linier untuk memprediksi efisiensi inhibisi korosi pada 41 senyawa piridin dan kuinolin dengan menggunakan metode GA-ANN, mendapatkan hasil MSE sebesar 16.7 dan RMSE sebesar 8.8.

Untuk penelitian ini, kami mengembangkan model QSPR berbasis ML untuk memprediksi efektivitas senyawa obat sebagai inhibitor korosi. Dengan menerapkan metode ini, kami berupaya untuk meningkatkan kemampuan prediktif model dalam hal inhibisi korosi. Fokus utama kami terletak pada bidang kompleks senyawa obat sebagai inhibitor korosi, di mana kami tidak hanya berusaha meningkatkan akurasi prediksi tetapi juga mendalami pemahaman tentang mekanisme dasar yang mengatur inhibisi korosi. Kolaborasi antara pemodelan komputasi dan validasi eksperimen ini memiliki potensi besar untuk membawa perubahan signifikan dalam bidang ilmu korosi, memberikan wawasan dan solusi baru untuk mengatasi tantangan terkait korosi di berbagai industri.

2. METODE PENELITIAN



Gambar 1. Alur Penelitian

Dalam penelitian ini, metodologi yang digunakan untuk studi eksperimental diilustrasikan dalam Gambar 1. Proses dimulai dengan pengumpulan dataset, yang digunakan untuk melatih model regresi. Data yang

dikumpulkan kemudian diproses, termasuk analisis data eksploratif, pembagian data menjadi set pelatihan dan pengujian, serta normalisasi data. Data yang telah diproses selanjutnya digunakan untuk melatih lima model regresi: Regresi Linear, Support Vector Machine, K-Nearest Neighbor, Random Forest, dan XGBoost. Setelah pelatihan, setiap model dievaluasi menggunakan metrik seperti MSE, RMSE, MAE, dan R2. Berdasarkan evaluasi metrik, dua model dengan kinerja terbaik dari tahap pelatihan awal dipilih untuk penyetelan hiperparameter, yaitu penyesuaian parameter model untuk lebih mengoptimalkan kinerjanya. Alur kerja ini memastikan bahwa model regresi yang dihasilkan adalah yang paling optimal untuk tugas prediksi yang diberikan.

2.1 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berasal dari penelitian yang dilakukan oleh Beltran-Perez et al. pada tahun 2022 [1]. Dataset ini terdiri dari 14 fitur dan 1 target. Fitur-fitur tersebut mencakup berbagai karakteristik molekuler seperti berat molekul (MW), konstanta disosiasi asam (pKa), koefisien partisi oktanol-air (log P), kelarutan dalam air (log S), area permukaan polar (PSA), polarizabilitas (α), energi orbital molekul terisi tertinggi (E-HOMO), energi orbital molekul tidak terisi terendah (E-LUMO), energi ionisasi (I), afinitas elektron (A), keelektronegatifan (χ), elektrofilisitas (ω), kekerasan (η), dan fraksi elektron yang dibagi (ΔN). Sedangkan targetnya adalah efisiensi inhibisi korosi (IE (%)), yang merupakan kinerja inhibisi korosi dari inhibitor.

Berat molekul mencerminkan ukuran senyawa obat, sedangkan konstanta disosiasi asam mengukur kekuatan asam dalam larutan. Log P menunjukkan kelarutan senyawa dalam lingkungan polar dan non-polar, sementara Log S memberikan informasi tentang kelarutan senyawa dalam air, dan PSA memberikan informasi tentang polaritas molekul, yang dapat memengaruhi kelarutannya dalam air. Polarisabilitas molekul, yang dipengaruhi oleh distribusi kerapatan elektron dan kapasitas molekul untuk distorsi dalam kerapatan elektronnya, juga berperan penting. Kemampuan transfer elektron diukur dengan nilai E-HOMO dan E-LUMO, di mana E-HOMO yang tinggi menunjukkan kemampuan donasi elektron yang baik dan E-LUMO menunjukkan penerimaan elektron. Energi ionisasi menunjukkan jumlah energi yang diperlukan untuk melepaskan elektron terluar dari atom, sedangkan afinitas elektron menunjukkan jumlah energi yang dilepaskan untuk menangkap satu mol elektron. Keelektronegatifan mencerminkan kemampuan atom untuk menarik elektron; pusat yang lebih elektronegatif diprediksi akan menunjukkan efisiensi inhibisi korosi yang lebih baik. Kekerasan global adalah parameter yang mencerminkan resistensi molekul terhadap transfer muatan. Terakhir, fraksi elektron yang dibagi memberikan ukuran interaksi elektronik antara logam dan molekul inhibitor. Semakin banyak elektron yang dibagi, semakin efektif inhibitor dalam mencegah korosi pada permukaan logam. Secara keseluruhan, efisiensi inhibisi korosi bergantung pada bagaimana molekul inhibitor berinteraksi dengan permukaan logam [16], [18]-[23].

2.2 Pemrosesan Data

Dataset ini terdiri dari 260 data poin, dengan sebagian besar mengandung nilai yang hilang. Untuk mengatasi masalah ini, data poin dengan nilai yang hilang dihapus, sehingga tersisa 78 data poin yang bersih. Data kemudian dibagi menjadi dua subset, yaitu untuk pelatihan dan pengujian model dengan rasio pembagian 80:20. Sebanyak 80% dari data digunakan untuk pelatihan model, sementara 20% sisanya digunakan untuk mengevaluasi kinerja model yang telah dilatih. Pembagian data ini penting untuk memvalidasi kinerja model pada data yang belum pernah dilihat sebelumnya dan mencegah overfitting. Selanjutnya, normalisasi data dilakukan menggunakan metode skala min-max. Pendekatan ini menyesuaikan setiap fitur dalam dataset agar nilainya berada dalam rentang tertentu, biasanya antara 0 hingga 1 atau -1 hingga 1. Persamaannya dapat dilihat pada persamaan (1).

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Di mana x menunjukkan nilai awal dari fitur, x_{min} mewakili nilai minimumnya, x_{max} menunjukkan nilai maksimumnya, dan x_{new} adalah nilai fitur yang telah dinormalisasi setelah transformasi [28].

2.3 Model Machine Learning

Teknik-teknik ML sangat ideal untuk membuat model prediktif ketika dataset yang cukup besar tersedia, hasilnya bergantung pada beberapa variabel, dan belum ada pemahaman mekanistik yang jelas tentang bagaimana variabel input berkaitan dengan hasilnya [29]. Teknik ini menggunakan berbagai algoritma yang belajar untuk melakukan tugas, seperti memprediksi hasil eksperimen, dengan dilatih menggunakan data dari eksperimen sebelumnya. Ketika algoritma ML dilatih menggunakan data yang mencakup hasil dari eksperimen, pendekatan ini disebut pembelajaran terawasi. Beberapa algoritma ML supervised yang populer adalah Regresi Linear, SVM, KNN, Random Forest, dan XGBoost.

Penelitian ini menggunakan lima algoritma ML yang berbeda untuk menganalisis data dan mendapatkan wawasan. Pemilihan algoritma-algoritma ini didasarkan pada keunggulan masing-masing. Regresi linear digunakan untuk memodelkan hubungan linear antara variabel prediktor dan variabel target numerik, menawarkan interpretasi yang langsung dari asosiasi tersebut [30]. Support Vector Machines (SVM), sebuah teknik pembelajaran terawasi yang serbaguna, mampu menangani tugas klasifikasi dan regresi, sangat cocok untuk skenario data dengan dimensi tinggi [31]. Metode K-Nearest Neighbors (KNN), sebuah teknik non-parametrik, digunakan untuk mengklasifikasikan atau memprediksi variabel target berdasarkan kedekatan data input terhadap tetangga terdekatnya dalam ruang fitur. Random Forest, sebuah pendekatan pembelajaran ensemble yang membangun beberapa pohon keputusan dan menggabungkan output-outputnya, diterapkan untuk memanfaatkan kekokohan dan skalabilitas teknik ini [32]. Selain itu, XGBoost, sebuah implementasi gradient boosting yang sangat dioptimalkan dan efisien, digunakan untuk memanfaatkan kekuatan metode pembelajaran ensemble ini dalam meningkatkan kinerja prediktif secara iteratif [33]. Parameter-parameter yang digunakan untuk melatih model-model untuk setiap algoritma dapat ditemukan di Tabel 1.

Tabel 1. Parameter Pelatihan Awal

Model	Parameter Name	Value
Linear Regression	Default	Default
SVM	Kernel	Linear
	C	100
	Epsilon	0.1
KNN	N neighbors	100
Random Forest	Random seed	42
	N Estimator	100
XGboost	Objective	reg:squarederror
	Random seed	42
	N Estimator	100

Setelah pelatihan awal menggunakan parameter-parameter awal, model-model yang dihasilkan dievaluasi menggunakan berbagai metrik. Dua model terbaik kemudian dipilih untuk penyempurnaan lebih lanjut guna mencapai hasil yang lebih baik. Penyetelan hiperparameter dilakukan pada model-model teratas ini menggunakan GridSearchCV untuk meningkatkan kinerjanya. Grid Search adalah metode untuk menyetel hiperparameter dalam model ML. Prosesnya melibatkan pencarian secara menyeluruh melalui semua kombinasi yang mungkin dari hiperparameter, melatih model untuk setiap kombinasi, dan membandingkan skor untuk mengidentifikasi yang terbaik. Grid Search biasanya diperluas dengan validasi silang, melatih model pada beberapa lipatan data yang berbeda dengan berbagai kombinasi hiperparameter untuk mencapai hasil yang lebih akurat [34]. Parameter-parameter yang digunakan untuk penyetelan hiperparameter ditampilkan di Tabel 2.

Table 2. Hyperparameter tuning for Random Forest and XGBoost

Model	Parameter Name	Value
Random Forest	n_estimators	[100, 300, 500]
	max_features	['log2', 'sqrt']
	max_depth	[None, 10, 30, 50]
	min_samples_split	[2, 10, 20]
	min_samples_leaf	[1, 4, 8]
	bootstrap	[True, False]
XGBoost	n_estimators	[100, 200, 300]
	max_depth	[3, 4, 5]
	learning_rate	[0.05, 0.1, 0.2]
	min_child_weight	[1, 3, 5]
	gamma	[0, 0.1, 0.2]
	subsample	[0.6, 0.8, 1.0]
	colsample_bytree	[0.6, 0.8, 1.0]

2.4 Evaluasi Metrik

Untuk menilai kinerja model, metrik evaluasi seperti Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan R-squared (R^2) digunakan untuk menentukan kesesuaian model dengan data. R^2 mengukur sejauh mana variabel dependen dipengaruhi oleh variabel independen, menunjukkan proporsi variabilitas. MSE dan MAE menilai deviasi antara nilai aktual dan nilai yang diprediksi. Berbeda dengan MSE, yang menghitung perbedaan kuadrat, MAE menghitung perbedaan absolut. Sementara itu, RMSE adalah akar kuadrat dari MSE, berfungsi untuk menstandarisasi satuan pengukuran [35]. Persamaan untuk metrik evaluasi ini dapat dilihat pada persamaan 2-5.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)^2} \quad (2)$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (4)$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i| \quad (5)$$

Di mana X_i adalah nilai aktual, Y_i adalah nilai yang diprediksi, $\bar{Y}$ adalah rata-rata dari nilai aktual, dan m adalah jumlah sampel.

3. HASIL DAN PEMBAHASAN

Table 3. Hasil Evaluasi

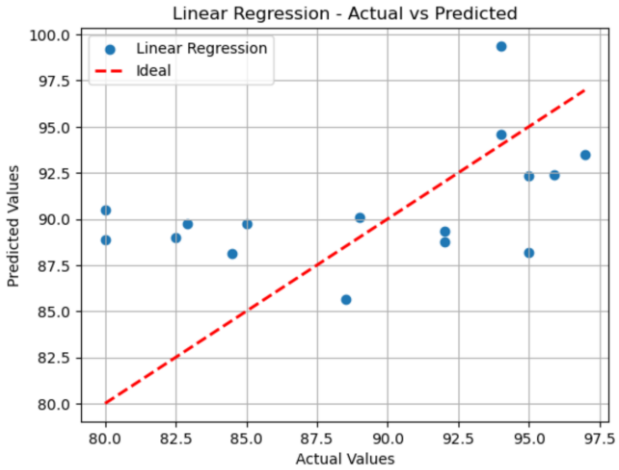
Model	Metric Evaluation			
	MSE	RMSE	MAE	R^2
Linear Regression	28.02	5.29	4.58	0.15
SVM	51.98	7.21	5.64	-0.57
KNN	29.93	5.47	4.59	0.09

Model	Metric Evaluation			
	MSE	RMSE	MAE	R^2
Random Forest	9.78	3.12	2.78	0.70
XGBoost	19.9	4.46	2.49	0.39
C. Beltran-Perez et al. [1]	49.47	7.03	-	-

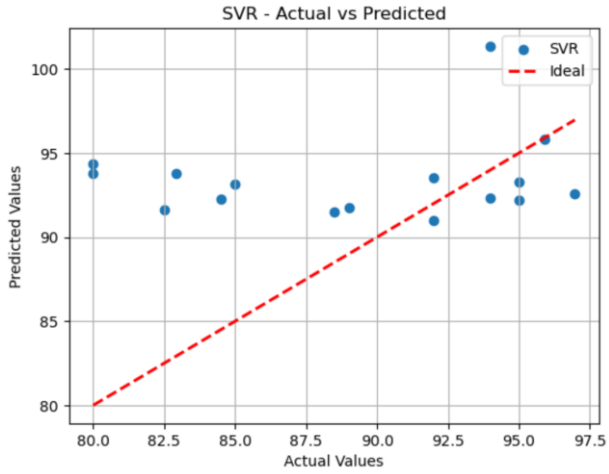
Berdasarkan tabel hasil evaluasi, kinerja berbagai algoritma ML dalam memprediksi variabel target dibandingkan. Tabel tersebut menyajikan metrik evaluasi seperti MSE, RMSE, MAE, dan R^2 untuk setiap model. Detail ini dapat dilihat di Tabel 3. Model Regresi Linear cukup baik dengan MSE sebesar 28.02, RMSE sebesar 5.29, MAE sebesar 4.58, dan R^2 sebesar 0.15. Namun, model ini masih memiliki kesalahan yang relatif besar dibandingkan dengan model lainnya. Sementara itu, model SVM menunjukkan kinerja buruk, dengan nilai MSE dan RMSE yang tinggi yaitu 51.98 dan 7.21, serta nilai R^2 negatif (-0.57), menunjukkan bahwa model ini tidak mampu melakukan prediksi dengan baik.

Model KNN menunjukkan kinerja yang mirip dengan Regresi Linear, dengan MSE sebesar 29.93, RMSE sebesar 5.47, MAE sebesar 4.59, dan R^2 sebesar 0.09. Model ini menunjukkan sedikit kesalahan yang lebih tinggi dan nilai R^2 yang lebih rendah dibandingkan dengan Regresi Linear. Di sisi lain, model Random Forest menunjukkan kinerja yang sangat baik, dengan MSE sebesar 9.78, RMSE sebesar 3.12, MAE sebesar 2.78, dan R^2 sebesar 0.70, menunjukkan bahwa Random Forest dapat melakukan prediksi dengan tingkat akurasi yang tinggi. Model XGBoost juga berkinerja baik, dengan MSE sebesar 19.9, RMSE sebesar 4.46, MAE sebesar 2.49, dan R^2 sebesar 0.39. Meskipun tidak sebaik Random Forest, XGBoost masih memberikan hasil yang cukup akurat. Sebaliknya, penelitian oleh C. Beltran-Perez et al. [1] menunjukkan hasil yang buruk, dengan MSE sebesar 49.47 dan RMSE sebesar 7.03. Sayangnya, nilai MAE dan R^2 tidak disediakan, sehingga sulit untuk dibandingkan langsung dengan model lainnya.

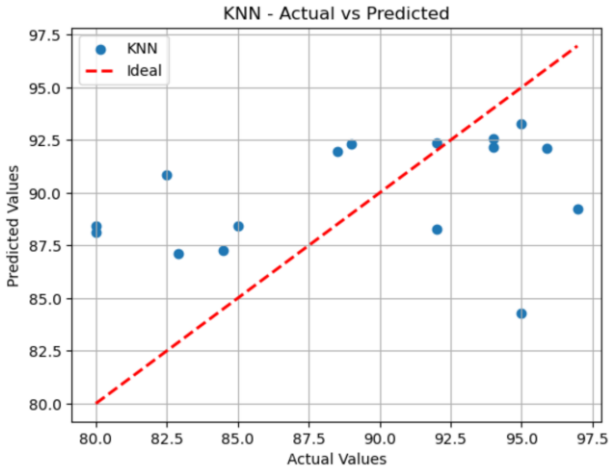
Dari hasil ini, dapat disimpulkan bahwa model Random Forest menawarkan kinerja prediksi terbaik, dengan kesalahan terendah dan nilai R^2 tertinggi. Ini menunjukkan bahwa Random Forest dapat menangkap pola data dengan lebih efektif dibandingkan dengan model lainnya. Model XGBoost juga berkinerja baik, dengan kesalahan yang relatif rendah dan nilai R^2 positif, meskipun Random Forest masih unggul dalam hal akurasi prediksi. Model Regresi Linear dan KNN menunjukkan kinerja yang serupa, dengan kesalahan yang relatif tinggi dan nilai R^2 yang rendah, menunjukkan bahwa keduanya kurang efektif dalam memprediksi variabel target dalam kasus ini. Model SVM menunjukkan kinerja terburuk, dengan kesalahan yang sangat tinggi dan nilai R^2 negatif, menunjukkan bahwa model ini tidak cocok untuk data ini. Secara keseluruhan, pemilihan model yang tepat sangat penting untuk meningkatkan akurasi prediksi, dan dalam kasus ini, Random Forest adalah pilihan terbaik berdasarkan metrik evaluasi yang disajikan. Hasil kurva regresi dapat dilihat dalam Gambar 2-6.



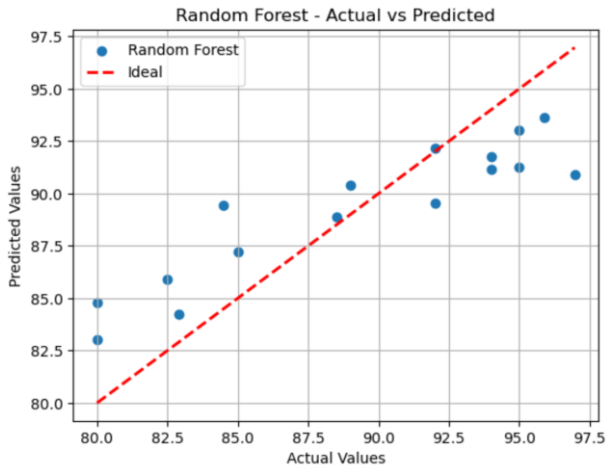
Gambar 2. Kurva Regresi Linear Regression



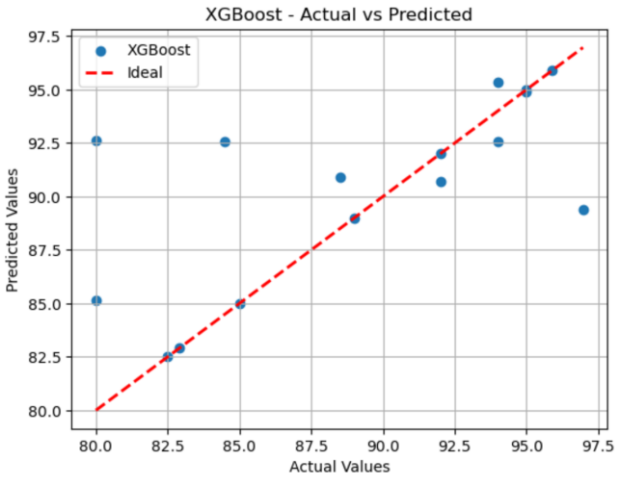
Gambar 3. Kurva Regresi SVM



Gambar 4. Kurva Regresi KNN



Gambar 5. Kurva Regresi Random Forest

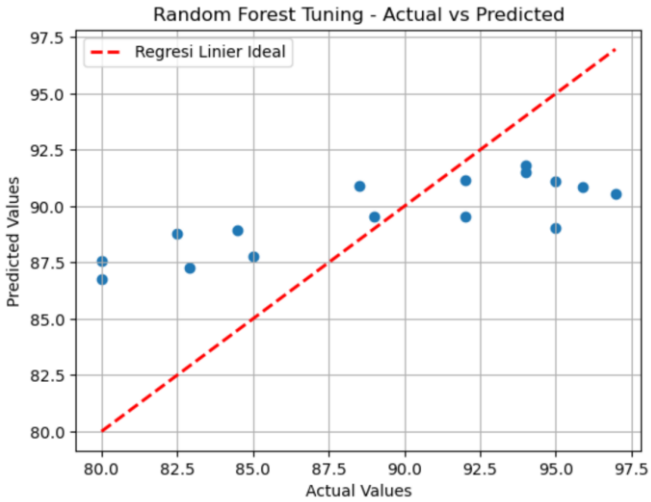


Gambar 6. Kurva Regresi XGBoost

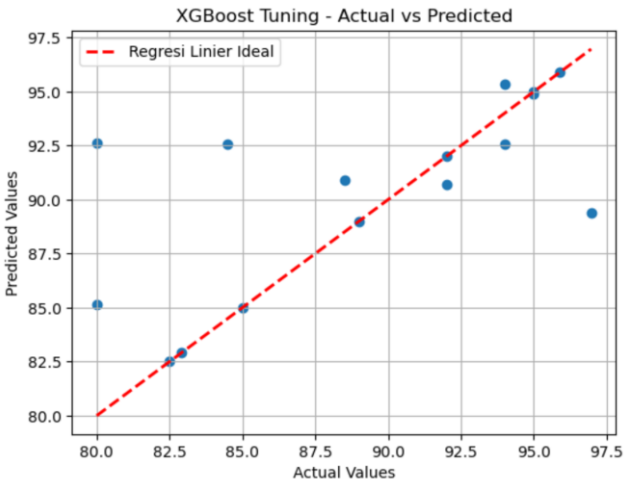
Tabel 4. Hasil Evaluasi Setelah Hiperparameter Tuning

Model	Metric Evaluation			
	MSE	RMSE	MAE	R^2
Random Forest	9.78	3.12	2.78	0.70
XGBoost	19.9	4.46	2.49	0.39
RF Tuning	20.6	4.54	4.02	0.37
XGBoost Tuning	8.33	2.88	1.83	0.74

Setelah penyetelan hiperparameter, seperti yang ditunjukkan dalam Tabel 4, hasil evaluasi menunjukkan perubahan kinerja pada dua model terbaik, Random Forest dan XGBoost. Kinerja Random Forest sedikit menurun, dengan MSE yang lebih tinggi (20.6) dibandingkan sebelum penyetelan (9.78). Nilai RMSE dan MAE juga meningkat, dan nilai R^2 turun dari 0.70 menjadi 0.37. Hal ini menunjukkan bahwa penyetelan hiperparameter tidak menghasilkan peningkatan yang signifikan untuk Random Forest dan justru mengurangi kinerja model. Di sisi lain, XGBoost menunjukkan peningkatan kinerja yang signifikan setelah penyetelan. Nilai MSE menurun drastis dari 19.9 menjadi 8.33, RMSE turun dari 4.46 menjadi 2.88, dan MAE menurun dari 2.49 menjadi 1.8. Nilai R^2 juga meningkat dari 0.39 menjadi 0.74. Hal ini menunjukkan bahwa penyetelan hiperparameter untuk XGBoost sangat efektif dalam meningkatkan akurasi prediksi. Hasil kurva regresi dapat dilihat dalam Gambar 7 dan 8.



Gambar 7. Kurva Regresi Random Forest Setelah Tuning



Gambar 8. Kurva Regresi XGBoost Setelah Tuning

Berdasarkan hasil penyetelan, dapat disimpulkan bahwa XGBoost dengan penyetelan hiperparameter memberikan kinerja terbaik dengan kesalahan terendah dan nilai R^2 tertinggi dibandingkan dengan semua model lainnya. Sementara itu, penyetelan hyperparameter pada Random Forest tidak menghasilkan hasil yang optimal dan justru mengurangi kinerja model. Oleh karena itu, XGBoost dengan penyetelan hiperparameter menjadi pilihan terbaik untuk memprediksi variabel target dalam kasus ini.

4. KESIMPULAN

Penelitian ini menggunakan berbagai algoritma ML untuk menganalisis dan memprediksi efisiensi inhibisi korosi oleh senyawa obat. Melalui tahap pelatihan awal, model-model dievaluasi menggunakan berbagai metrik untuk mengidentifikasi yang terbaik. Dua model teratas kemudian dilakukan penyetelan hyperparameter menggunakan GridSearchCV untuk meningkatkan kinerjanya lebih lanjut. Hasil menunjukkan bahwa model XGBoost, setelah disetel dengan baik, secara signifikan melampaui kinerja algoritma lainnya. Model ini menunjukkan tingkat kesalahan terendah dan nilai R^2 tertinggi, menjadikannya model paling efektif untuk memprediksi variabel target dalam konteks ini. Implementasi XGBoost yang efisien dalam gradient boosting mampu meningkatkan akurasi prediksi dengan meminimalkan kesalahan secara iteratif.

Di sisi lain, penyetelan hiperparameter pada model Random Forest tidak memberikan peningkatan kinerja yang signifikan. Bahkan, hal ini menyebabkan penurunan akurasi prediksi. Hal ini menunjukkan bahwa meskipun Random Forest adalah teknik yang kuat dan dapat diskalakan, kinerjanya tidak selalu meningkat melalui penyetelan hiperparameter, terutama dalam konteks dari studi ini. Temuan ini menegaskan pentingnya memilih model ML yang tepat dan mengoptimalkannya untuk mencapai kinerja prediktif terbaik. Dalam konteks memprediksi inhibisi korosi oleh senyawa obat, XGBoost menonjol sebagai algoritma yang paling sesuai dan dapat diandalkan. Penelitian masa depan harus berfokus pada penyempurnaan lebih lanjut dari model-model ini dan mengeksplorasi algoritma tambahan untuk terus meningkatkan akurasi dan keandalan prediksi. Studi ini memberikan dasar yang kuat untuk menggunakan teknik ML dalam bidang ilmu korosi, menyoroti potensi untuk kemajuan signifikan melalui penerapan metode-metode ini.

UCAPAN TERIMA KASIH

Penelitian ini didukung oleh Research Center of Materials Informatics Universitas Dian Nuswantoro. Kami mengucapkan terima kasih kepada rekan-rekan dari Universitas Dian Nuswantoro yang telah memberikan wawasan dan keahlian yang sangat membantu dalam penelitian ini. Meskipun mungkin tidak sepenuhnya setuju dengan semua interpretasi atau kesimpulan dari makalah ini, kontribusi mereka telah sangat berpengaruh dalam membentuk hasil penelitian ini.

REFERENSI

[1] C. Beltran-Perez *et al.*, "A General Use QSAR-ARX Model to Predict the Corrosion Inhibition Efficiency of Drugs in Terms of Quantum Mechanical Descriptors and Experimental Comparison for Lidocaine," *Int J Mol Sci*, 2022, doi: 10.3390/ijms23095086.

[2] M. Akrom *et al.*, "DFT and microkinetic investigation of oxygen reduction reaction on corrosion inhibition mechanism of iron surface by Syzygium Aromaticum extract," *Appl Surf Sci*, 2023, doi: 10.1016/j.apsusc.2022.156319.

[3] M. Akrom, "Investigation Of Natural Extracts As Green Corrosion Inhibitors In Steel Using Density Functional Theory," *Jurnal Teori dan Aplikasi Fisika*, 2022, doi: 10.23960/jtaf.v10i1.2927.

[4] M. Tampubolon, R. G. Gultom, L. Siagian, P. Lumbangaol, and C. Manurung, "Laju Korosi Pada Baja Karbon Sedang Akibat Proses Pencelupan Pada Larutan Asam Sulfat (H2SO4) dan Asam Klorida (HCl) dengan Waktu Bervariasi," *SPROCKET JOURNAL OF MECHANICAL ENGINEERING*, 2020, doi: 10.36655/sproket.v2i1.294.

[5] M. Akrom, S. Rustad, A. G. Saputro, A. Ramelan, F. Fathurrahman, and H. K. Dipojono, "A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds," *Mater Today Commun*, 2023, doi: 10.1016/j.mtcomm.2023.106402.

[6] E. Bowman *et al.*, "International Measures of Prevention, Application, and Economics of Corrosion Technologies Study. NACE International. Available at: http://impact.nace.org/documents/Nace-International-Report.pdf," 2016.

[7] L. Mawaddah, M. R. Rosyid, A. P. Santosa, and M. Akrom, "Optimizing Quantum Neural Networks for Predicting the Effectiveness of Drug Compounds as Corrosion Inhibitors," *Technology and Science (BITS)*, vol. 6, no. 1, 2024, doi: 10.47065/bits.v6i1.5318.

[8] M. Akrom *et al.*, "Artificial Intelligence Berbasis QSPR Dalam Kajian Inhibitor Korosi," *JoMMiT : Jurnal Multi Media dan IT*, 2023, doi: 10.46961/jommit.v7i1.721.

[9] M. Akrom, S. Rustad, and H. K. Dipojono, "A machine learning approach to predict the efficiency of corrosion inhibition by natural product-based organic inhibitors," *Phys Scr*, 2024, doi: 10.1088/1402-4896/ad28a9.

[10] Q. Wang *et al.*, "Application of Biomass Corrosion Inhibitors in Metal Corrosion Control: A Review," *Molecules*. 2023. doi: 10.3390/molecules28062832.

[11] M. Akrom, S. Rustad, and H. K. Dipojono, "SMILES-based machine learning enables the prediction of corrosion inhibition capacity," *MRS Commun*, Apr. 2024, doi: 10.1557/s43579-024-00551-6.

[12] M. Akrom, "Investigation of Syzygium Aromaticum and Nicotiana Tabacum Extracts as Corrosion Inhibitor," *Science Tech: Jurnal Ilmu Pengetahuan dan Teknologi*, vol. 8, no. 1, pp. 42–48, Feb. 2022, doi: 10.30738/st.vol8.no1.a11775.

[13] N. Vaszilcsin, V. Ordodi, and A. Borza, "Corrosion inhibitors from expired drugs," *Int J Pharm*, 2012, doi: 10.1016/j.ijpharm.2012.04.015.

[14] A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," *MRS Communications*. 2019. doi: 10.1557/mrc.2019.73.

[15] M. Akrom, S. Rustad, and H. Kresno Dipojono, "Prediction of Anti-Corrosion performance of new triazole derivatives via Machine learning," *Comput Theor Chem*, vol. 1236, Jun. 2024, doi: 10.1016/j.comptc.2024.114599.

[16] A. H. Alamri and N. Alhazmi, "Development of data driven machine learning models for the prediction and design of pyrimidine corrosion inhibitors," *Journal of Saudi Chemical Society*, vol. 26, no. 6, p. 101536, Nov. 2022, doi: 10.1016/j.jscs.2022.101536.

[17] M. Liu and W. Li, "Prediction and analysis of corrosion rate of 3C steel using interpretable machine learning methods," *Mater Today Commun*, vol. 35, p. 106408, Jun. 2023, doi: 10.1016/j.mtcomm.2023.106408.

[18] T. W. Quadri *et al.*, "Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors," *Mater Today Commun*, vol. 30, p. 103163, 2022, doi: https://doi.org/10.1016/j.mtcomm.2022.103163.

[19] Y. Liu *et al.*, "A Machine Learning-Based QSAR Model for Benzimidazole Derivatives as Corrosion Inhibitors by Incorporating Comprehensive Feature Selection," *Interdiscip Sci*, vol. 11, Jul. 2019, doi: 10.1007/s12539-019-00346-7.

[20] T. H. Pham, P. K. Le, and D. N. Son, "A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques," *RSC Adv*, vol. 14, no. 16, pp. 11157–11168, Apr. 2024, doi: 10.1039/d4ra02159b.

[21] C. T. Ser, P. Žuvela, and M. W. Wong, "Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships," *Appl Surf Sci*, vol. 512, May 2020, doi: 10.1016/j.apsusc.2020.145612.

[22] I. B. Obot and S. A. Umoren, "Experimental, DFT and QSAR models for the discovery of new pyrazines corrosion inhibitors for steel in oilfield acidizing environment," *Int J Electrochem Sci*, vol. 15, no. 9, pp. 9066–9080, Sep. 2020, doi: 10.20964/2020.09.72.

[23] A. H. Radhi, "HOMO-LUMO Energies and Geometrical Structures Effecton Corrosion Inhibition for Organic Compounds Predict by DFT and PM3 Methods," *NeuroQuantology*, vol. 18, no. 1, pp. 37–45, Jan. 2020, doi: 10.14704/nq.2020.18.1.NQ20105.

[24] X. Chen, Y. Chen, J. Cui, Y. Li, Y. Liang, and G. Cao, "Molecular dynamics simulation and DFT calculation of 'green' scale and corrosion inhibitor," *Comput Mater Sci*, vol. 188, p. 110229, Feb. 2021, doi: 10.1016/j.commatsci.2020.110229.

[25] S. Hadisaputra, A. D. Irham, A. A. Purwoko, E. Junaidi, and A. Hakim, "Development of QSPR models for furan derivatives as corrosion inhibitors for mild steel," *Int J Electrochem Sci*, vol. 18, no. 8, p. 100207, Aug. 2023, doi: 10.1016/j.ijoes.2023.100207.

[26] R. L. Camacho-Mendoza, L. Feria, L. Á. Zárate-Hernández, J. G. Alvarado-Rodríguez, and J. Cruz-Borbolla, "New QSPR model for prediction of corrosion inhibition using conceptual density functional theory," *J Mol Model*, vol. 28, no. 8, p. 238, Aug. 2022, doi: 10.1007/s00894-022-05240-6.

[27] I. B. Obot and N. O. Obi-Egbedi, "Theoretical study of benzimidazole and its derivatives and their potential activity as corrosion inhibitors," *Corros Sci*, vol. 52, no. 2, pp. 657–660, Feb. 2010, doi: 10.1016/j.corsci.2009.10.017.

[28] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl Soft Comput*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[29] M. Aghaaminiha *et al.*, "Machine learning modeling of time-dependent corrosion rates of carbon steel in presence of corrosion inhibitors," *Corros Sci*, vol. 193, p. 109904, Dec. 2021, doi: 10.1016/j.corsci.2021.109904.

[30] E. Jumin *et al.*, "Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, pp. 713–725, Jan. 2020, doi: 10.1080/19942060.2020.1758792.

[31] F. G. Altin, İ. Budak, and F. Özcan, "Predicting the amount of medical waste using kernel-based SVM and deep learning methods for a private hospital in Turkey," *Sustain Chem Pharm*, vol. 33, p. 101060, Jun. 2023, doi: 10.1016/j.scp.2023.101060.

[32] A. Huang, R. Xu, Y. Chen, and M. Guo, "Research on multi-label user classification of social media based on ML-KNN algorithm," *Technol Forecast Soc Change*, vol. 188, p. 122271, Mar. 2023, doi: 10.1016/j.techfore.2022.122271.

[33] S. Ben Jabeur, S. Mefteh-Wali, and J.-L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Ann Oper Res*, vol. 334, no. 1–3, pp. 679–699, Mar. 2024, doi: 10.1007/s10479-021-04187-w.

[34] C. G. Siji George and B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," *International Journal of Advanced Computer Science and Applications*, 2020, doi: 10.14569/IJACSA.2020.0110920.

[35] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.