Vol. 9 No.2 Tahun 2023, pp 172-182
Jurnal Teknologi dan Manajemen Informatika
http:// http://jurnal.unmer.ac.id/index.php/jtmi
P-ISSN: 1693-6604 E-ISSN: 2580-8044

Implementasi Fungsi Polinomial pada Algoritma *Gradient Boosting Regressor:* Studi Regresi pada Dataset Obat-Obatan Kadaluarsa sebagai Material Antikorosi

Nicholaus Verdhy Putranto[1], Muhamad Akrom[2], Gustina Alfa Trinapradika[3]

[1]Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Indonesia
[2,3] Research Center for Materials Informatics, Universitas Dian Nuswantoro, Indonesia

Article Info

Article History
Received: 29-09-2023
Revides : 20-10-2023
Accepted: 06-11-2023

Keywords
Corrosion Inhibitor;
Machine Learning;
Polynomial Function;
Gradient Boosting Regressor

Corresponding Author
Muhamad Akrom,
Universitas Dian Nuswantoro,
Tel. +6285859000875
Email :
m.akrom@dsn.dinus.ac.id

ABSTRACT

Corrosion is an electrochemical process between the metal surface and a corrosive environment that can lead to significant losses in various industries, especially in the oil and gas sector. Experimental studies are conducted to evaluate the performance of corrosion inhibitors and available resources. In this research, a machine learning (ML) approach is employed to assess the effectiveness of expired drug compounds as corrosion inhibitors. The primary challenge in machine learning is obtaining a highly accurate model to ensure that predictions are relevant to the properties of the tested materials. Therefore, the polynomial function is tested in the gradient-boosting regressor (GBR) algorithm to enhance the accuracy of the developed ML model. The results indicate that the implementation of the polynomial function in the GBR algorithm can improve the accuracy of the prediction model based on R2 and RMSE metrics.

PENDAHULUAN

Korosi merupakan proses elektrokimia antara permukaan logam dengan lingkungan. Korosi yang dapat menimbulkan kerugian besar di berbagai bidang indrusti kususnya pada minyak dan gas [1]–[3]. Masalah ini sebagian disebabkan oleh parameter operasi yang terkait dengan kimia air, atribut aliran, sifat dari material yang ada dan aktifitas dari mikrobiologi [4]. Korosi logam hasil dari reaksi elektrokimia redoks yang ditandakan dengan bagian yang disebut anodik yang bertanggung jawab atas oksidasi logam, dan bagian katodik yang membuat reduksi molekul oksigen dan air dari lingkungan yang ada [5], [6].

Korosi memiliki dampak ekonomi yang signifikan, karena membutuhkan biaya tinggi untuk perbaikan dan penggantian material yang rusak [7], [8]. Korosi ini patut diselidiki, karena masalah korosi ini memakan besar total biaya Perusahaan terutama penghasil minyak dan gas setiap tahunnya di dunia [9]. Biaya untuk penanganan korosi sendiri sekitar US$ 2,5 triliun per tahun [10]. Untuk mencegah terjadinya peristiwa korosi tersebut, kita dapat menghambat dengan inhibitor [7], [11].

Senyawa farmasi memiliki sifat kimia tertentu yang mampu menghambat atau mengurangi laju korosi pada material tertentu [7]. Mereka dapat membentuk lapisan pelindung pada permukaan material, mencegah reaksi korosif antara material dan lingkungan [4].

Penggunaan senyawa farmasi sebagai *inhibitor* korosi memiliki beberapa keuntungan, termasuk keberlanjutan lingkungan dan *biodegradabilitas* yang baik [7].

Machine learning (ML) merupakan cabang dari kecerdasan buatan yang memungkinkan sistem komputer belajar dari data dan pengalaman untuk melakukan tugas-tugas tertentu [12]. ML mampu memecahkan permasalahan dalam eksplorasi material anti korosi [13]. Dalam konteks prediksi dan analisis korosi, metode ML digunakan untuk mengembangkan model prediktif yang dapat memperkirakan tingkat korosi pada material berdasarkan atribut-atribut yang relevan [10]. Algoritma ML seperti *Gradient Boosting Regressor* (GBR) telah terbukti efektif dalam memodelkan hubungan kompleks antara atribut-atribut dan tingkat korosi [14], [15]. Algoritma tersebut mampu menghasilkan akurasi dengan nilai *coefficient of determination* (R^2) dan *root mean square error* (RMSE) masing-masing sebesar 0.90 dan 0.02 [16]. Pada penelitian lain, model GBR diketahui lebih unggul dibandingkan model *support vector regression* (SVR) dan *k-nearest neighbor* (KNN) [17].

Implementasi fungsi *polynomial* terbukti meningkatkan kinerja algoritma NuSVR secara signifikan [18]. Oleh karena itu, pada penelitian ini mengimplementasikan fungsi *polynomial* untuk meningkatkan akurasi model dari algoritma GBR. Dalam memprediksi nilai efisiensi penghambatan korosi/*corrosion inhibiton efficiency* (CIE) dari senyawa *drugs* [19]. GBR dipilih karena terbukti lebih baik dari algoritma lainnya seperti SVR dan KNN [14]–[17].

METODE

Dataset

Pada penelitian ini, menggunakan dataset yang dipublikasikan yang berisikan 250 senyawa *drugs* dengan 10 deskriptor molekuler sebagai fitur [19]. Fitur -Fitur tersebut adalah *molecular weight* (MW), *acid dissociation constant* (pKa), *octanol-water partition coefficient* (log P), *water solubility* (log S), *polar surface area* (PSA), *polarizability* (α), *energy of highest occupied molecular orbital* (E-HOMO), *energy of lowest unoccupied molecular orbital* (E-LUMO), *electrophilicity* (ω), dan *the fraction electron shared* (ΔN).

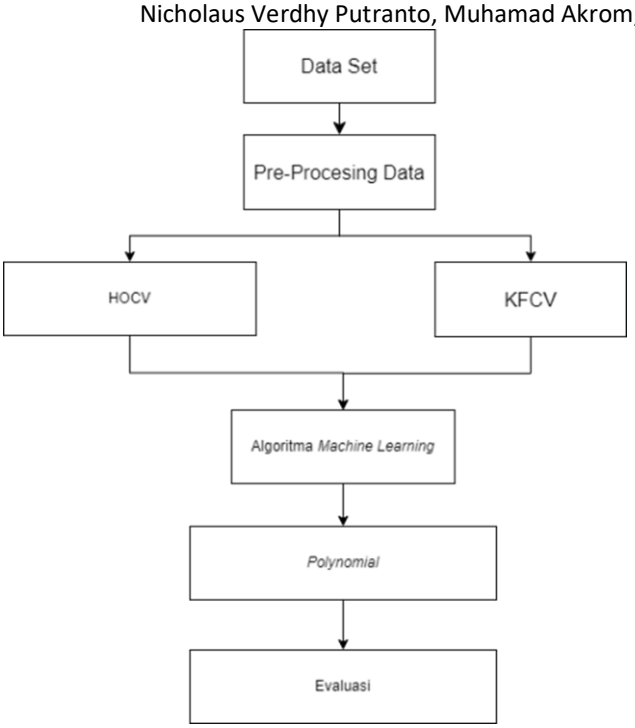
Preprosesing

Pada tahap preprosesing, dilakukan *data cleaning* untuk mengatasi *missing value* dan outlier. Selanjutnya, normalisasi data dilakukan untuk menghindari sensitifitas terhadap data melimpah atau terbatas sehingga menghilangkan perbedaan skala antar fitur dengan teknik *MinMaxScaler* [10], [12], [18].

Model ML

ML merupakan pendekatan statistik yang bertujuan untuk mengungkapkan korelasi antara variabel independen (fitur) dan dependen (target) [20]. Tahap awal penelitian ini adalah membandingkan kinerja model GBR pada variasi *split* data train dan test menggunakan teknik *cross-validation HoldOut* (HOCV) dan *K-Fold* (KFCV) [21], [22]. Pada HOCV, data displit dengan perbandingan rasio 80:20, 70:30, dan 60:40, sedangkan pada KFCV, data displit dengan variasi k = 3, 5, dan 10. Teknik ini dapat mengatasi masalah bias dan varian pada data [17].

Untuk menangani sensivitas terhadap ukuran data, teknik normalisasi *MinMaxScaler* digunakan [19]. Untuk memvalidasi keunggulan GBR, kami membandingkannya dengan model lain, yaitu NuSVR dan *random forest* (RF) [18], [23]. Tahap selanjutnya adalah mengimplementasikan fungsi polynomial pada model GBR untuk meningkatkan akurasi model prediksi [18], [24]. Kinerja model diukur oleh metrik R^2 dan RMSE [25]. Ilustrasi model ML yang dikembangkan dapat dilihat pada Gambar 1.



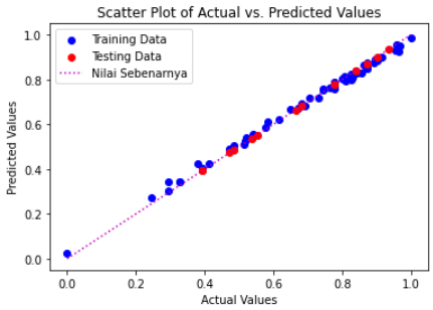
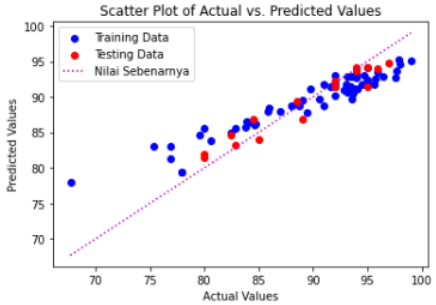
Gambar 1. Pengembangan model ML

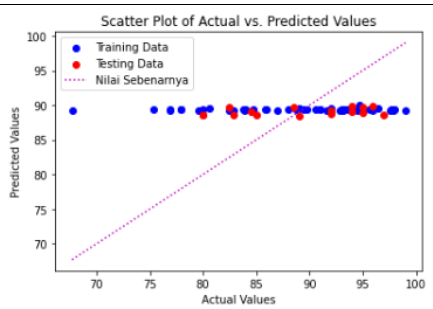
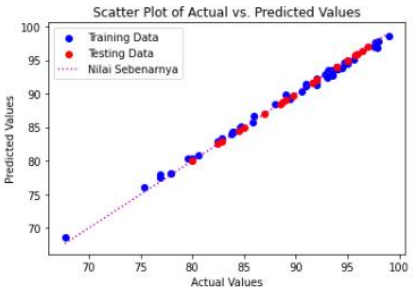
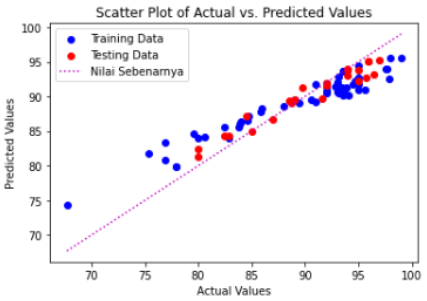
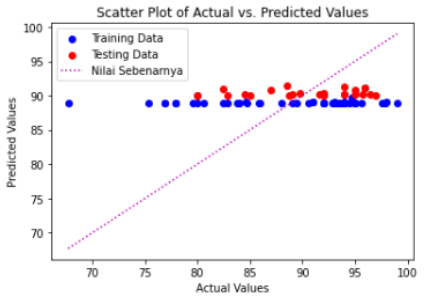
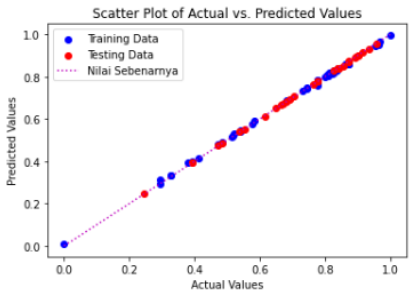
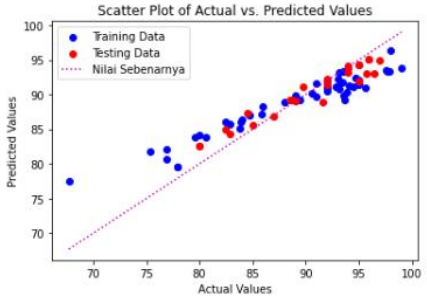
HASIL DAN PEMBAHASAN

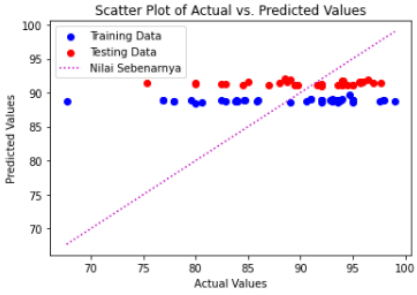
Hasil

Kinerja prediksi model GBR, RF, dan NuSVR berdasarkan metrik evaluasi R^2 dan RMSE serta distribusi data poin disajikan pada Tabel 1 dan Tabel 2. Pada Tabel 1, disajikan informasi mengenai hasil dari model yang diuji dengan menggunakan variasi split data (train-test) menggunakan teknik HOCV dengan rasio 80:20, 70:30, dan 60:40. Sedangkan pada Tabel 2 menyajikan split data menggunakan teknik KFCV.

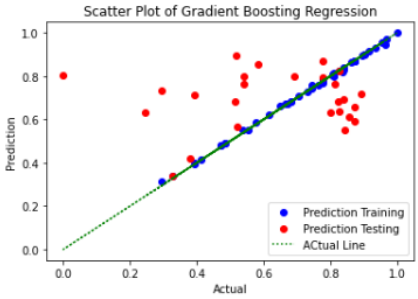
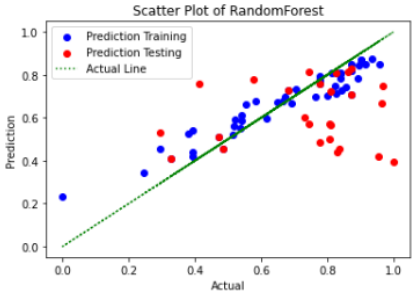
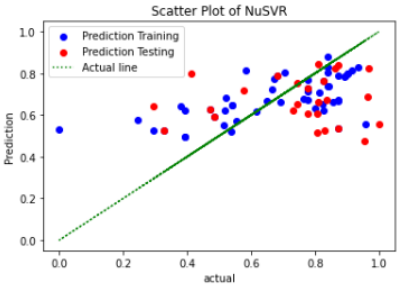
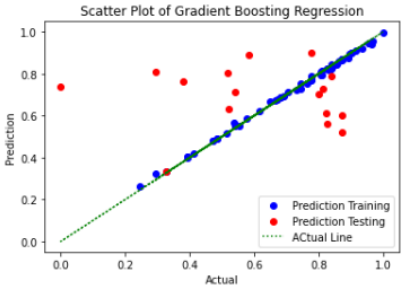
Tabel 1. Kinerja model prediksi dengan HOCV

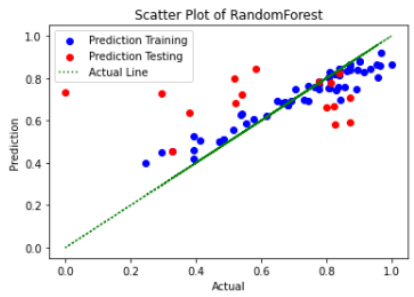
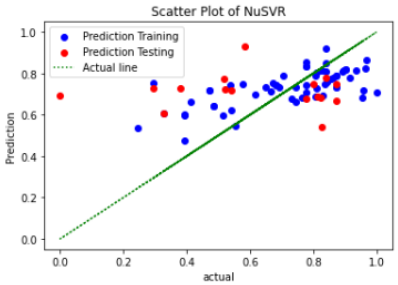
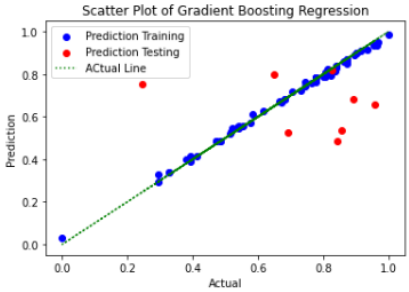
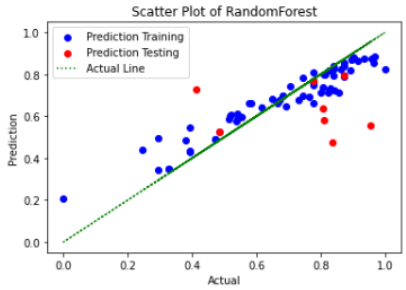
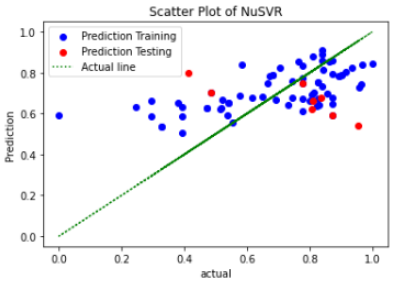
No	HOCV	Model	R^2	RMSE	Distribusi Data
1.	80:20	GBR	0.9932	0.017	
		RF	0.7911	3.055	

		NuSVR	0.0020	6.679	
2.	70:30	GBR	0.9952	0.015	
		RF	0.8231	2.925	
		NuSVR	0.0036	6.942	
3.	60:40	GBR	0.9988	0.008	
		RF	0.7222	3.666	

		NuSVR	0.0111	6.931	

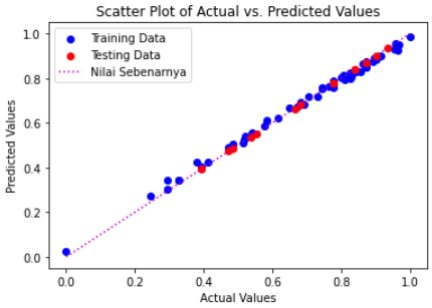
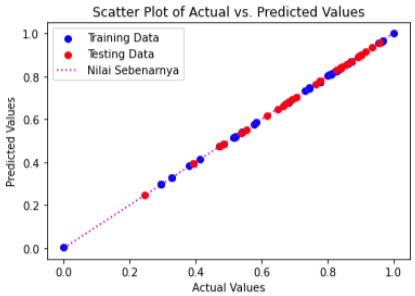
Tabel 2. Kinerja Model Prediksi dengan KFCV

No	Fold	Model	R^2	RMSE	Distribusi Data
1	3	GBR	0.9985	0.007	
		RF	0.8590	0.079	
		NuSVR	0.3840	0.166	
2	5	GBR	0.9970	0.010	

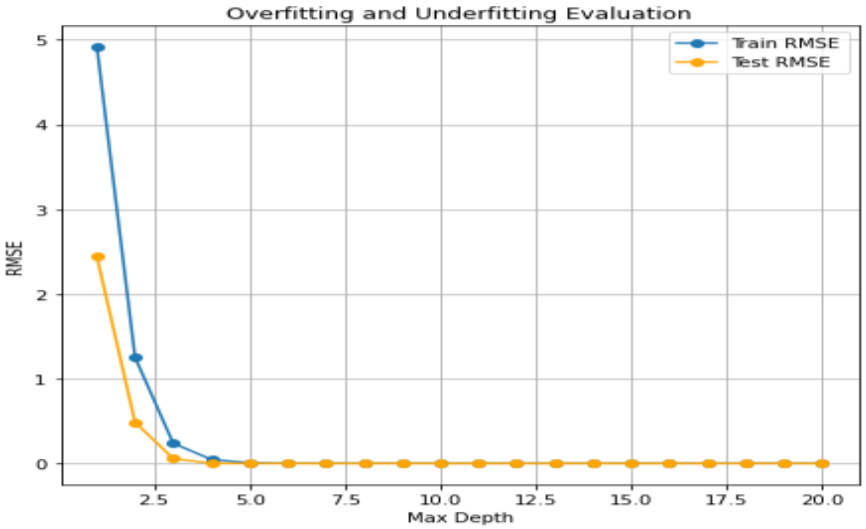
		RF	0.8690	0.068	
		NuSVR	0.4110	0.144	
3	10	GBR	0.9950	0.013	
		RF	0.8660	0.076	
		NuSVR	0.3700	0.166	

Pada Tabel 3, informasi yang disajikan didalamnya adalah menampilkan perbandingan kinerja model GBR sebelum dan sesudah implementasi fungsi *polynomial*. Hal tersebut berdasarkan metrik R^2 dan RMSE serta distribusi data poin.

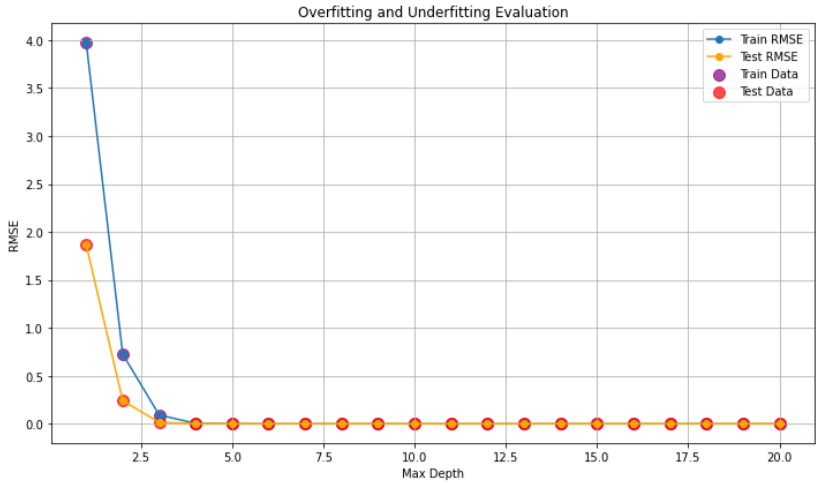
Tabel 3. Perbandingan Kinerja Model GBR Sebelum Dan Setelah Implementasi Fungsi Polynomial

Model	*Training*		*Testing*		Distribusi Data
	R^2	RMSE	R^2	RMSE	
GBR	0.9988	0.008	0.9998	0.002	
GBR-Polynomial	0.9998	0.002	0.9999	0.0003	

Gambar 2 dan Gambar 3 menyajikan kurva pembelajaran model GBR untuk mengevaluasi sejauh mana model mampu melakukan generalisasi data.

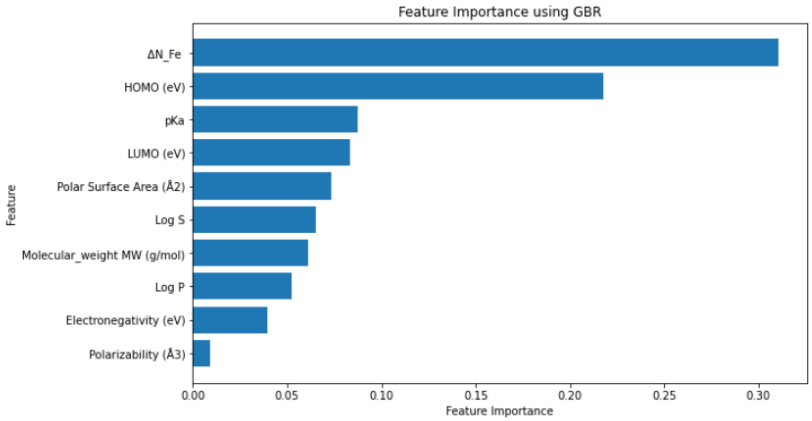


Gambar 2. Kurva Pembelajaran Model GBR (HOCV, rasio 60:40)



Gambar 3. Kurva Pembelajaran Model GBR-Polynomial

Profil *feature importance* (fitur penting) disajikan pada Gambar 4. Ini dapat membantu menganalisis sejauh mana fitur-fitur dapat berpengaruh terhadap kinerja model.



Gambar 4. Analisis Fitur Penting

Pembahasan

Berdasarkan Tabel 1 dan Tabel 2, diketahui bahwa model GBR menunjukkan keunggulan performa prediksi dibandingkan model RF. Sedangkan SVR untuk penggunaan teknik HOCV dan KFCV. Berdasarkan metrik R^2 dan RMSE. GBR memiliki nilai R^2 yang paling tinggi dan RMSE yang terendah. Model yang baik adalah yang memiliki R^2 tinggi dan RMSE rendah [17] [18]. Pada teknik split HOCV (Tabel 1), rasio 60:40 menjadi model GBR terbaik dengan R^2 dan RMSE masing-masing sebesar 0.9988 dan 0.008. Sedangkan pada teknik split KFCV (Tabel 2), nilai k = 3 menjadi model GBR terbaik dengan R^2 dan RMSE masing-masing sebesar 0.9985 dan 0.007. Secara keseluruhan, GBR dengan HOCV (rasio 60:40) merupakan model terbaik. Hal itu dapat dikonfirmasi pada visualisasi distribusi data poin yang cenderung paling mendekati *fitting line* (garis aktual). Keunggulan model GBR ini juga terlihat dari kemampuannya menangani bias dan varian. Sedangkan untuk mengantisipasi kemungkinan terjadinya *overfitting* maupun *underfitting* (Gambar 2), dimana model mampu melakukan pembelajaran secara optimal dan stabil [26].

Selanjutnya kami mengkombinasikan fungsi *polynomial* pada model GBR terbaik untuk melihat pengaruhnya terhadap kinerja prediksi. Hasilnya dapat dilihat pada Tabel 3. Diketahui bahwa implementasi fungsi *polynomial* pada model GBR dapat meningkatkan akurasi model

dimana nilai R^2 meningkat dari 0.9988 menjadi 0.9998 untuk data *training*. Dari 0.9998 menjadi 0.9999 untuk data *testing.* Serta nilai RMSE menurun dari 0.008 menjadi 0.002 untuk data training dan dari 0.002 menjadi 0.0003 untuk data *testing*. Hasil tersebut juga didukung oleh plot distribusi data poin prediksi yang mengikuti pola *fitting line*. Ini menunjukkan bahwa fungsi *polynomial* dapat membantu prediksi model GBR menjadi lebih akurat dalam memprediksi nilai CIE senyawa *drugs* yang diuji sehingga relevan dengan karakteristik sebenarnya [18]. Kombinasi fungsi *polynomial* pada model GBR juga tetap mempertahankan kemampuan model dalam menangani bias dan varian. Hal tersebut didukung oleh kurva pembelajaran pada Gambar 3, dimana model tetap optimal dan stabil selama proses pembelajaran.

Analisis terhadap fitur penting pada model GBR dilakukan untuk mengukur sejauh mana setiap fitur dapat berpengaruh besar terhadap kemampuan model dalam melakukan kinerjanya [27], [28]. Dari hasil analisis fitur penting (Gambar 4), dapat diamati bahwa deskriptor *the fraction electron shared* (ΔN) dan energi HOMO tampil sebagai fitur yang paling bertanggungjawab terhadap performa model. Namun demikian, seluruh fitur menunjukkan korelasi positif terhadap target (CIE) yang diprediksi, hal ini tentu saja sangat mendukung kinerja model prediksi. Deskriptor molekuler menjadi faktor penting dalam pemodelan berbasis ML sebagai fitur *input* [17], [29], [30].

SIMPULAN DAN SARAN

Untuk memprediksi efisiensi penghambtan korosi oleh senyawa obat-obatan kadaluarsa (*drugs*), penulis mengusulkan teknik ML berbasis QSPR yang unggul. Berdasarkan nilai R^2 dan RMSE, model GBR muncul sebagai model yang terbaik dibandingkan RF dan SVR. Implementasi fungsi *polynomial* pada model GBR ditemukan dapat meningkatkan akurasi model secara signifikan. Secara keseluruhan, penelitian ini memberikan sudut pandang lain dalam meningkatkan akurasi model ML untuk memprediksi kemampuan penghambatan korosi oleh senyawa *drugs*. Menjadi penting untuk dipertimbangkan pada penelitian selanjutnya, apakah model GBR dan implementasi fungsi *polynomial* juga menunjukkan kinerja yang relevan terhadap dataset dari senyawa lainnya sebagai inhibitor korosi. Selain itu, penting juga diimplementasikan teknik lain untuk meningkatkan akurasi model prediksi, seperti penerapan teknik *virtual sample generator* (VSG) misalnya.

DAFTAR PUSTAKA

[1] T. Sutojo, S. Rustad, M. Akrom, A. Syukur, G. F. Shidik, and H. K. Dipojono, "A machine learning approach for corrosion small datasets," *Npj Mater. Degrad.*, vol. 7, no. 1, Art. no. 1, Mar. 2023, doi: 10.1038/s41529-023-00336-7.

[2] M. Akrom, "INVESTIGATION OF NATURAL EXTRACTS AS GREEN CORROSION INHIBITORS IN STEEL USING DENSITY FUNCTIONAL THEORY," *J. Teori Dan Apl. Fis.*, vol. 10, no. 1, Art. no. 1, Jan. 2022, doi: 10.23960/jtaf.v10i1.2927.

[3] M. A. Quraishi, D. S. Chauhan, and V. S. Saji, "Heterocyclic biomolecules as green corrosion inhibitors," *J. Mol. Liq.*, vol. 341, p. 117265, Nov. 2021, doi: 10.1016/j.molliq.2021.117265.

[4] C. I. Ossai, "A Data-Driven Machine Learning Approach for Corrosion Risk Assessment—A Comparative Study," *Big Data Cogn. Comput.*, vol. 3, no. 2, Art. no. 2, Jun. 2019, doi: 10.3390/bdcc3020028.

[5] T. L. P. Galvão, G. Novell-Leruth, A. Kuznetsova, J. Tedim, and J. R. B. Gomes, "Elucidating Structure–Property Relationships in Aluminum Alloy Corrosion Inhibitors by Machine Learning," *J. Phys. Chem. C*, vol. 124, no. 10, pp. 5624–5635, Mar. 2020, doi: 10.1021/acs.jpcc.9b09538.

[6] 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.*, vol. 615, p. 156319, Apr. 2023, doi: 10.1016/j.apsusc.2022.156319.

[7] P. D. N. H. S. M.T S. T., S. M.T S. T., and S. H. M.T S. T., *Pengantar Inhibitor Korosi Alami*. Deepublish, 2021.

[8] M. A. Deyab, Q. Mohsen, and L. Guo, "Aesculus hippocastanum seeds extract as eco-friendly corrosion inhibitor for desalination plants: Experimental and theoretical studies," *J. Mol. Liq.*, vol. 361, p. 119594, Sep. 2022, doi: 10.1016/j.molliq.2022.119594.

[9] M. Finšgar and J. Jackson, "Application of corrosion inhibitors for steels in acidic media for the oil and gas industry: A review," *Corros. Sci.*, vol. 86, pp. 17–41, Sep. 2014, doi: 10.1016/j.corsci.2014.04.044.

[10] M. Akrom *et al.*, "Artificial Intelligence Berbasis QSPR Dalam Kajian Inhibitor Korosi," *JoMMiT J. Multi Media Dan IT*, vol. 7, no. 1, pp. 015–020, Jul. 2023, doi: 10.46961/jommit.v7i1.721.

[11] S. Tanwer and S. K. Shukla, "Recent advances in the applicability of drugs as corrosion inhibitor on metal surface: A review," *Curr. Res. Green Sustain. Chem.*, vol. 5, p. 100227, Jan. 2022, doi: 10.1016/j.crgsc.2021.100227.

[12] 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.*, vol. 35, p. 106402, Jun. 2023, doi: 10.1016/j.mtcomm.2023.106402.

[13] A. H. Alamri, "Application of machine learning to stress corrosion cracking risk assessment," *Egypt. J. Pet.*, vol. 31, no. 4, pp. 11–21, Dec. 2022, doi: 10.1016/j.ejpe.2022.09.001.

[14] D. Xia, S. Song, J. Wang, J. Shi, H. Bi, and Z. Gao, "Determination of corrosion types from electrochemical noise by phase space reconstruction theory," *Electrochem. Commun.*, vol. 15, no. 1, pp. 88–92, Feb. 2012, doi: 10.1016/j.elecom.2011.11.032.

[15] L. Yan, Y. Diao, Z. Lang, and K. Gao, "Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach," *Sci. Technol. Adv. Mater.*, vol. 21, no. 1, pp. 359–370, Jan. 2020, doi: 10.1080/14686996.2020.1746196.

[16] Y. Diao, L. Yan, and K. Gao, "Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features," *Mater. Des.*, vol. 198, p. 109326, Jan. 2021, doi: 10.1016/j.matdes.2020.109326.

[17] M. Akrom and T. Sutojo, "Investigation of QSPR-Based Machine Learning Models in Pyrimidine Corrosion Inhibitors," *Eksergi*, vol. 20, no. 2, Art. no. 2, Jul. 2023, doi: 10.31315/e.v20i2.9864.

[18] S. Budi, M. Akrom, G. A. Trisnapradika, T. Sutojo, and W. A. E. Prabowo, "Optimization of Polynomial Functions on the NuSVR Algorithm Based on Machine Learning: Case Studies on Regression Datasets," *Sci. J. Inform.*, vol. 10, no. 2, Art. no. 2, May 2023, doi: 10.15294/sji.v10i2.43929.

[19] 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.*, vol. 23, no. 9, Art. no. 9, Jan. 2022, doi: 10.3390/ijms23095086.

[20] H. D. Panduwinata, S. Suyitno, and M. N. Huda, "Model Regresi Weibull Pada Data Kontinu yang Diklasifikasikan," *EKSPONENSIAL*, vol. 13, no. 2, Art. no. 2, Nov. 2022.

[21] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *J. Am. Stat. Assoc.*, vol. 0, no. 0, pp. 1–12, Apr. 2023, doi: 10.1080/01621459.2023.2197686.

[22] J. J. Eertink, M. W. Heymans, G. J. C. Zwezerijnen, J. M. Zijlstra, H. C. W. de Vet, and R. Boellaard, "External validation: a simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients," *EJNMMI Res.*, vol. 12, no. 1, p. 58, Sep. 2022, doi: 10.1186/s13550-022-00931-w.

[23] Y. Ji *et al.*, "Random forest incorporating ab-initio calculations for corrosion rate prediction with small sample Al alloys data," *Npj Mater. Degrad.*, vol. 6, no. 1, Art. no. 1, Oct. 2022, doi: 10.1038/s41529-022-00295-5.

[24] V. V. Ramalingam and P. Ramasamy, "Modelling Corrosion Behavior of Friction Stir Processed Aluminium Alloy 5083 Using Polynomial: Radial Basis Function," *Trans. Indian Inst. Met.*, vol. 70, no. 10, pp. 2575–2589, Dec. 2017, doi: 10.1007/s12666-017-1110-1.

[25] M. Akrom, S. Rustad, A. G. Saputro, and H. K. Dipojono, "Data-driven investigation to model the corrosion inhibition efficiency of Pyrimidine-Pyrazole hybrid corrosion inhibitors," *Comput. Theor. Chem.*, vol. 1229, p. 114307, Nov. 2023, doi: 10.1016/j.comptc.2023.114307.

[26] M. Akrom, S. Rustad, and H. Kresno Dipojono, "Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors," *Results Chem.*, vol. 6, p. 101126, Dec. 2023, doi: 10.1016/j.rechem.2023.101126.

[27] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," *Brief. Bioinform.*, vol. 23, no. 1, p. bbab354, Jan. 2022, doi: 10.1093/bib/bbab354.

[28] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.

[29] Putra et al., 2019. Study on building machine learning model to predict biodegradable-ready materials | AIP Conference Proceedings | AIP Publishing. Available: https://pubs.aip.org/aip/acp/article-abstract/2088/1/060003/797803/Study-on-building-machine-learning-model-to

[30] M. Akrom, "DFT Investigation of Syzygium Aromaticum and Nicotiana Tabacum Extracts as Corrosion Inhibitor," *Sci. Tech J. Ilmu Pengetah. Dan Teknol.*, vol. 8, no. 1, Art. no. 1, Feb. 2022, doi: 10.30738/st.vol8.no1.a11775.