ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

MODEL HIBRID GENETIC-XGBOOST DAN PRINCIPAL COMPONENT ANALYSIS PADA SEGMENTASI DAN PERAMALAN PASAR

Rimbun Siringoringo✉, Resianta Perangin-angin, Jamaluddin
Universitas Methodist Indonesia, Medan, Indonesia
Email: rimbun.ringo@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol5No2.pp97-103

ABSTRACT

Extreme Gradient Boosting (XGBoost) is a popular boosting algorithm based on decision trees. XGBoost is the best in the boosting group. XGBoost has excellente convergence. On the other hand, XGBoost is a Hyper parameterized model. Determining the value of each parameter is classified as difficult, resulting in the results obtained being trapped in the local optimum situation. Determining the value of each parameter manually, of course, takes a lot of time. In this study, a Genetic Algorithm (GA) is applied to find the optimal value of the XGBoost hyperparameter on the market segmentation problem. The evaluation of the model is based on the ROC curve. Test result. The ROC test results for several SVM, Logistic Regression, and Genetic-XGBoost models are 0.89; 0.98; 0.99. The results show that the Genetic-XGBoost model can be applied to market segmentation and forecasting.

Keyword: Genetic Algorithm, Market Forecasting, Extreme Gradient Boosting, Market Segmentation, Principal Component Analysis.

ABSTRAK

Extreme Gradient Boosting atau XGBoost adalah algoritma *boosting* populer yang didasarkan pada pohon keputusan. XGBoost merupakan yang terbaik pada kelompok *boostin.* XGBoost memiliki konvergensi yang sangat baik. Disisi lain, XGBoost merupakan model yang Hyper parameter. Penentuan nilai setiap parameter tergolong sulit mengakibatkan hasil yang diperoleh terjebak pada situasi *local optimum*. Penentuan nilai setiap parameter secara manual tentu saja menghabiskan waktu yang tidak sedikit. Pada penelitian ini, *Genetic Algoritm* (GA) diterapkan untuk pencarian nilai optimal *hyperparameter* XGBoost pada masalah segmentasi pasar. Evaluasi model didasarkan kurva ROC. Hasil pengujian. Hasil pengujian ROC untuk beberapa model SVM, Logistic Regression, dan Genetic-XGBoost adalah sebesar 0,89; 0,98; 0,99. Hasil penelitian menunjukkan bahwa model Genetic-XGBoost dapat diterapkan pada segmentasi dan peramalan pasar.

Kata Kunci: Algoritma Genetika, Peramalan Pasar, Extreme Gradient Boosting, Segmentasi Pasar, Principal Component Analysis.

PENDAHULUAN

Indonesia menempati urutan pertama dengan jumlah minimarket terbanyak di Asia Tenggara. Berdasarkan data yang di publikasikan pada (Mulya, Si, Hermawan, & Evienia, 2019), terdapat 43.826 minimarket pada tahun 2019. Terdapat beberapa faktor yang mempengaruhi kondisi tersebut, yaitu pertambahan usia muda kelas menengah (Aribawa, 2016) serta aktivitas masyarakat yang sibuk. Dengan meningkatnya penetrasi internet, masyarakat kini lebih banyak mendapatkan informasi penawaran merek dan produk tertentu. Oleh karena itu diperlukan metode penawaran yang tepat sesuai dengan selera lokal dan perilaku konsumen agar kelangsungan bisnis berjalan dengan lancar.

TINJAUAN PUSTAKA

Penelitian Terkait

Penerapan sains data dan sistem-sistem cerdas dapat membantu segmentasi pasar dan konsumen. Beberapa studi bidang segmentasi pasar melakukan pendekatan pada dua pendekatan utama yaitu pendekatan *time series* (Anggraeni, Andri, Sumaryanto, & Mahananto, 2017), dan *machine learning* (Chatzis, Siakoulis, Petropoulos, Stavroulakis, & Vlachogiannakis, 2018). Metode *time series* melibatkan berbagai model diantaranya adalah *exponential smooting* dan ARIMA (Joyce & Br, 2020). Hasil penelitian menunjukan bahwa model tersebut dapat memprediksi tren volume sampah. Disisi lain, metode-metode di atas tidak teralu memuaskan

terutama pada kasus-kasus *non linier* (H. Wang, Lei, Zhang, Zhou, & Peng, 2019). Model-model *machine learning* yang juga populer pada masalah segmentasi adalah *artificial neural network* (ANN) (Pourdaryaei et al., 2019), *support vector regression* (SVR) (Li, Wang, Cheng, & Bai, 2020), dan metode *Convolutional Neural Network* (CNN). Disisi lain, metode-metode di atas sering terjebak pada optimum lokal dan *overfitting* (Cai et al., 2020). XGBoost adalah motode *machine learning* yang sangat populer. Salah satu kelebihan XGBoost adalah tingkat konvergensi dan generalisasi yang sangat baik (Jiang, Tong, Yin, & Xiong, 2019), (Y. Wang & Guo, 2020). Penerapan XGBoost pada klasifikasi menghasilkan tingkat akurasi yang tinggi (Wu, Guo, & Wang, 2020).

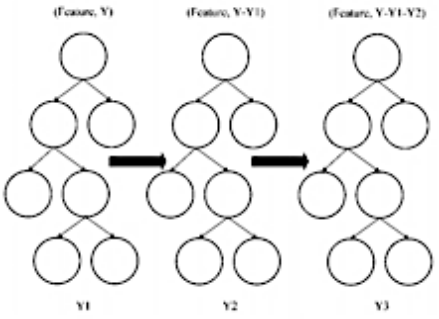
Meskipun XGBoost terkenal sebagai model yang *robust*, banyak studi mengemukakan beberapa kelemahan metode ini. XGBoost merupakan model yang memiliki banyak Parameter (*Hyper parameter*). Penentuan nilai setiap parameter tergolong sulit mengakibatkan hasil yang diperoleh terjebak pada situasi *local optimum* (Mai, Sheng, Shi, & Liao, 2021). Penentuan nilai setiap parameter secara manual tentu saja menghabiskan waktu yang tidak sedikit.

Pendekatan populer yang diterapkan pada penentua nilai parameter XGBoost adalah *Grid Search* dan *Random Search*, *Bayesian optimization* (Y. Xia, Liu, Li, & Liu, 2017), (Putatunda & Rama, 2019). Meskipun metode-metode tersebut mudah untuk di implementasikan, namun termasuk *comptational expensive* (Putatunda & Rama, 2019). *Genetic Algorithm* (GA) merupakan solusi penyelesaian masalah optimasi *hyper parameter* yang sangat populer, diantaranya adalah (X. Xia, Jiang, Zhou, Li, & Wang, 2018) pada penyelesaian *hyper parameter* apada aplikasi polusi air, dan penyelesaian *feedforward neural networks* (Boonthanawat & Boonyasiriwat, 2021).

Pada penelitian ini, GA diterapkan untuk pencarian nilai optimal *hyper parameter* XGBoost pada masalah segmentasi pasar.

XGBoost

XGBoost merupakan adalah algoritma *boosting* yang didasarkan pada pohon keputusan atau pohon regresi sebagaimana ditampilkan pada gambar 1 (Jiang et al., 2019). Dengan XGBoost, nilai *error* dapat direduksi dengan efektif dari pohon pertama (*feature*, Y) ke pohon selanjutnya (*feature*, |Y-Y1|).



Gambar 1. Pohon Regresi **XGBoost**

Principal Component Analysis (PCA)

PCA merupakan metode reduksi fitur data dengan mereduksi variabel-variabel data menjadi lebih kecil. Jika ada data berbentuk matrik *X pada persamaan (1)* dengan sejumlah *n atribut* × *m record* data (Zhang, Zhang, & Wu, 2020). Pada persamaan tersebut, $t_i \in R^n$ adalah nilai vektor dan *m adalah banyak verktor yang terbentuk*

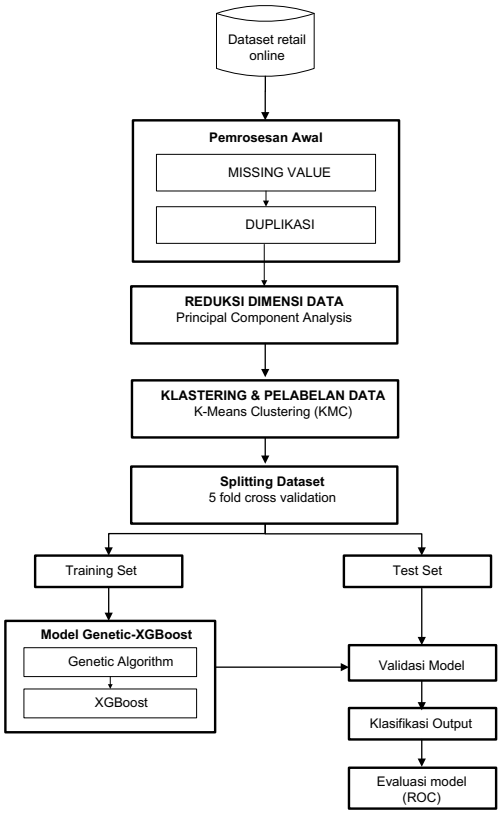
$$X = t_{1P_1^T} + t_{2P_2^T} + \cdots + t_m P_{m+}^T = TP^T \quad (1)$$

Nilai vektor *X* merupakan *principal component* untuk matrik *X*. Jika elemen-elemen minor dihilangkan maka hanya tersisa mejnadi sejumlah *a elemen principal* yang diterapkan, matrik *X* dapat dijelaskan dalam bentuk persamaan (2).

$$X = \sum_{i=1}^{a} t_{iP_i^T} + \sum_{i=a+1}^{m} \quad (2)$$

METODOLOGI

Gambaran umum prosedur dan tahapan penyelesaian masalah segmentasi berbasis Genetic-XGBoost digambarkan pada gambar 2.



Gambar 2. Model Genetic-XGBoost

Data

Sumber data penelitian ini menggunakan *online ritel data* dari *UCI Machine Learning Dataset* (https://archive.ics.uci.edu/ml/datasets/online+retail) Dataset tersebut terdiri dari 525.461 *record* dan 8 fitur data

Tabel 1. Deskripsi Fitur Dataset

Atribut	Unit	Deskripsi
Invoice_No	Nominal	Nomor transaksi
Stock_Code	Nominal	Kode produk
Description	Nominal	Nama produk
Quantity	Numeric	Banyak produk per transaksi
Invoice_Date	Numeric	Tanggal transaksi
Unit_Price	Numeric	Harga produk
Customer_ID	Nominal	Nomor unik setiap pelanggan
Country	Nominal	Negara pelanggan

Pemrosesan awal

Data awal yang digunakan yaitu *online retail* dataset terdapat *missing value*, data *outlier* dan duplikasi di data. Penanganan ke tiga aspek data ini bertujuan agar akurasi klasifikasi dapat dimaksimalkan. Selanjutnya adalah transformasi data awal menjadi data yang diberi label atau klaster. Model data yang diterapkan untuk penelitian ini sesuai dengan skema pada gambar 3. Fitur data dengan dimensi *mn* dengan *m* adalah label atau klaster data, dan *n* merupakan banyak fitur data. Pelabelan data didasarkan pada hasil klaster menggunakan *K-Means Clustering* (KMC).

	fitur 2	*fitur 3*	...	*fitur n*
target 1			...	
target 2			...	
target 3			...	
...	...	...	...	...
target m			...	

Gambar 3. Model Transformasi Data Dimensi *m x n*

Genetic Algorithm (GA)

Fungsi *fitness*

Fungsi *fitness* untuk GA ditentukan berdasarkan skor *F1* sebagaimana di rumuskan pada persamaan (3).

$$F1 = \frac{TP}{TP+{}^{1}/_{2}(FP+FN)} \quad (3)$$

Dengan TP adalah *True Positive*, FP adalah *False Positive*, dan FN adalah *False Negative.* Pengkodean nilai setiap parameter XGBoost ditentukan secara acak *uniform* adalah sebagai berikut:

Tabel 2. Format Data Parameter XGBoost

No	Parameter	Nilai (min, max), step
1	*Learning Rate*	(0.01, 1), 2
2	*N Estimators*	(10, 1500), 25
3	*Max Depth*	(1, 10), 1
4	*Min Child Weight*	(0.01, 10.0), 2
5	*Gamma Value*	(0.01, 10.0), 2
6	*Sub Sample*	(0.01, 1.0), 2
7	*Col Sample By Tree*	(0.01, 1.0), 2

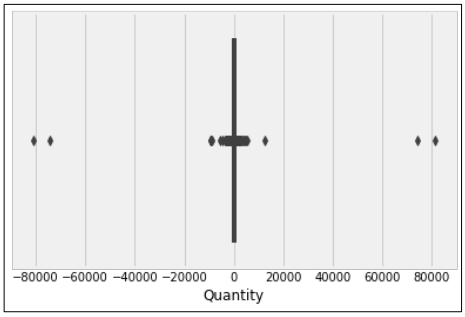
Skema pengkodean GA membetuk matriks numerik berdimensi 8x8 yaitu 8 generasi dengan setiap generasi terdiri dari 8 *parent*.

HASIL DAN PEMBAHASAN

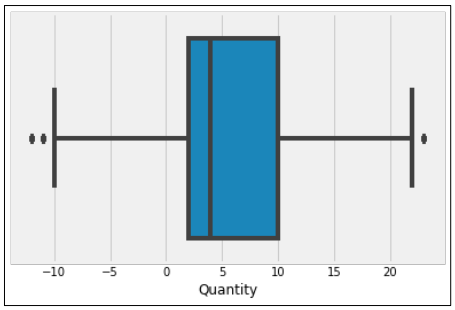
Pemrosesan awal

Dengan metode *Inter Quartile Range* (IQR), *outlier* data pada Atribut *UnitPrice* dan *Quantity* (gambar 4) direduksi sehingga menyajikan data yang lebih baik sebagaimana disajikan melalui gambar 5.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

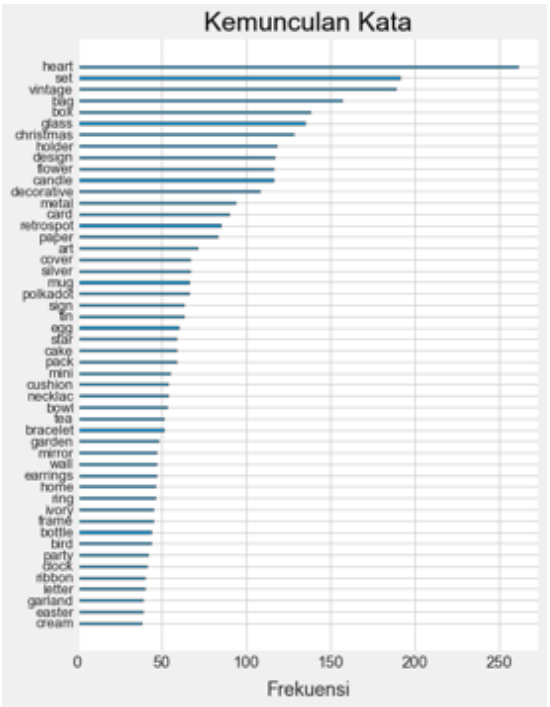


Gambar 4. Visualisasi *Boxplot Outlier* Data



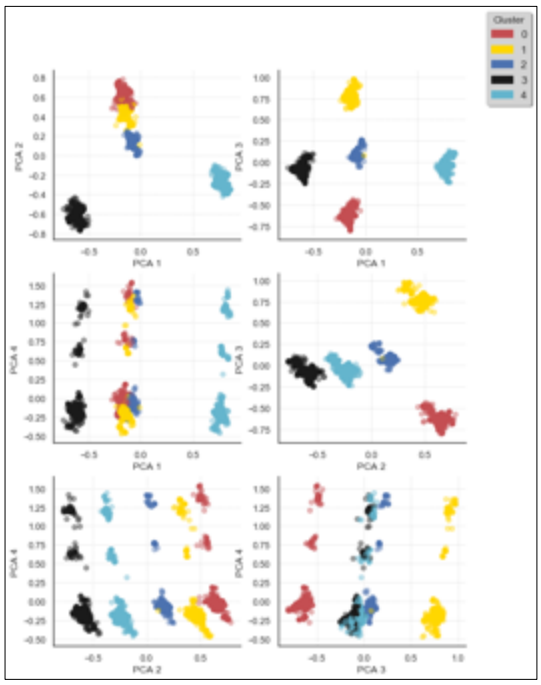
Gambar 5. *Boxplot* Tanpa *Outier*

Visualisasi kemunculan setiap *keyword* sesuai dengan grafik pada gambar 5. *Keyword* produk pada dataset ada sebanyak 1404 *term*.



Gambar 6. Frekuensi kemunculan *keyword*

Penerapan PCA menghasilkan kondisi klaster yang baik yang terbukti dari visualisasi klaster yang terpisah antara klaster yang satu dengan yang lainnya sebagaimana ditampilkan pada gambar 7.



Gambar 7. Visualisasi klaster

Hasil Transformasi dataset ditampilkan pada tabel 3. Banyak fitur data adalah 5 fitur yaitu X={k_0, k_1, k_2, k_3, k_4} dan banyak target label adalah 11 klaster yaitu Y={0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

Tabel 3. Data Hasil Klaster Pelanggan

No	k_0	k_1	k_2	k_3	k_4	cluster
1	-1.782	-0.076	0.058	-0.226	-0.243	3
2	0.652	0.988	-2.501	-0.416	-0.605	0
3	-0.550	0.527	0.902	-0.734	0.708	3
4	-1.445	1.457	-2.071	1.138	-1.771	2
5	3.706	-0.280	0.165	-0.936	-0.376	1
6	-0.187	1.208	2.139	-0.592	-1.132	6
7	-0.008	-0.073	-0.645	0.149	-0.362	3
8	1.404	-0.213	-0.256	-0.513	-0.340	1
9	2.701	0.630	0.762	0.197	0.020	1
10	2.309	0.470	-0.374	0.097	-0.619	1

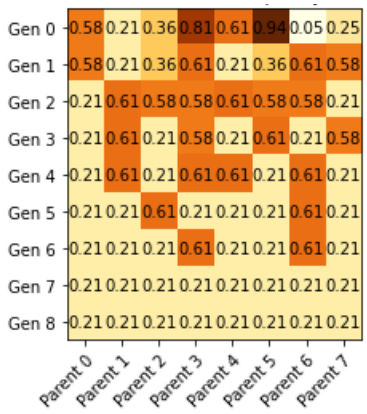
Tabel 4. Segmentasi Data Produk Dan Konsumen

Klaster (Y)	k_0	k_1	k_2	k_3	k_4	Jlh konsumen
4	**59,642**	10,137	13,483	9,576	3,600	289
6	7,607	**58,220**	10,604	14,327	3,085	351
3	9,458	12,801	**56,535**	13,132	3,146	339
5	4,186	9,700	5,981	**66,924**	3,484	266
0	5,050	7,873	10,049	9,714	**63,942**	48
10	10,581	19,333	16,134	**23,173**	22,485	353
1	11,175	23,420	20,501	**29,760**	3,620	1020
7	11,802	24,783	20,632	**28,266**	7,061	523
2	14,619	21,817	21,580	**26,775**	8,968	76
9	17,949	23,625	20,197	22,942	8,706	184
8	22,583	20,810	16,815	24,416	7,045	7

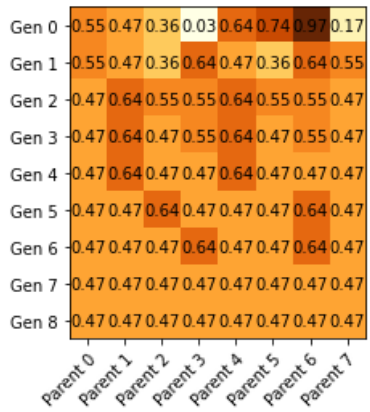
ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Optimasi Hyperparameter XGBoost dengan GA

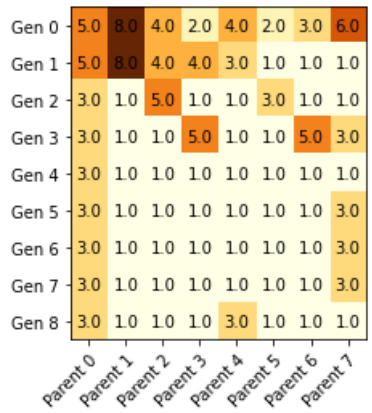
Pada gambar berikut di tampilkan proses pencarian nilai parameter XGBoost dengan GA. Pada gambar 8 berikut ditampilkan proses pencarian nilai Parameter *colsample*_bytree, *learning_rate,* dan *max_depth*.



(a)



(b)



(c)

Gambar 8. Visualisasi pencarian nilai parameter XGBoost dengan GA. (a). Parameter colsample_bytree, (b). Parameter learning_rate, (c). Parameter max_depth

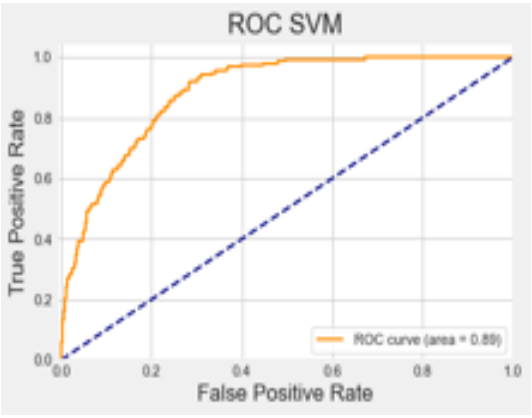
Hasil akhir Parameter *XGBoost* ditampilkan pada tabel 5 berikut.

Tabel 5. Hyper Parameter XGBoost

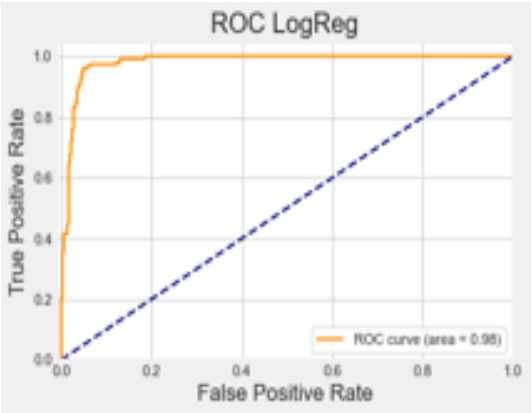
Hyperparameter	Nilai
max_depth	3
min_child_weight	4,27
gamma	2,56
learning_rate	0,47
sub_sample	0,47
colsample_bytree	0,21
n_estimator	79

Evaluasi Model Dengan Grafik ROC

Hasil perbandingan grafik ROC ditampilkan pada gambar 9 sampai gamabr 11 berikut. ROC model SVM, LogReg, dan *Genetic*-XGBoost masing-masing 0,89; 0,98; 0,99



Gambar 9. Grafik ROC SVM



Gambar 10. Grafik ROC LogReg

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Gambar 11. Grafik ROC Genetic-XGBoost

Berdasarkan grafik ROC tersebut, Genetic-XGBoost memiliki nilai ROC yang lebih baik dibanding dua model yang lain.

KESIMPULAN

Penelitian ini melakukan optimasi *hyper parameter* XGBoost menggunakan GA dan untuk mereduksi dimensi dataset diterapkan PCA. Berdasarkan hasil penelitian yang telah uraikan pada pembahasan di atas disimpulkan bahwa GA dapat menentukan nilai *hyper parmeter* XGBoost dengan baik

DAFTAR PUSTAKA

Anggraeni, W., Andri, K. B., Sumaryanto, & Mahananto, F. (2017). The Performance of ARIMAX Model and Vector Autoregressive (VAR) Model in Forecasting Strategic Commodity Price in Indonesia. *Procedia Computer Science*, *124*, 189–196. https://doi.org/https://doi.org/10.1016/j.procs.2017.12.146

Aribawa, D. (2016). E-commerce strategic business environment analysis in Indonesia. *International Journal of Economics and Financial Issues*, *6*(6Special Issue), 130–134.

Boonthanawat, C., & Boonyasiriwat, C. (2021). Finding optimal hyperparameters of feedforward neural networks for solving differential equations using a genetic algorithm. *Journal of Physics: Conference Series*, *1719*(1), 12033. https://doi.org/10.1088/1742-6596/1719/1/012033

Cai, R., Xie, S., Wang, B., Yang, R., Xu, D., & He, Y. (2020). Wind Speed Forecasting Based on Extreme Gradient Boosting. *IEEE Access*, *8*, 175063–175069. https://doi.org/10.1109/access.2020.3025967

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*, *112*, 353–371. https://doi.org/https://doi.org/10.1016/j.eswa.2018.06.032

Jiang, Y., Tong, G., Yin, H., & Xiong, N. (2019). A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters. *IEEE Access*, *7*, 118310–118321. https://doi.org/10.1109/access.2019.2936454

Joyce, H. M. V, & Br, M. (2020). *Modeling and Forecasting of COVID-19 Confirmed Cases in Indonesia Using ARIMA and Exponential Smoothing*. *198*(Issat), 253–258.

Li, C., Wang, X., Cheng, Z., & Bai, Y. (2020). Forecasting Bus Passenger Flows by Using a Clustering-Based Support Vector Regression Approach. *IEEE Access*, *8*, 19717–19725. https://doi.org/10.1109/ACCESS.2020.2967867

Mai, Y., Sheng, Z., Shi, H., & Liao, Q. (2021). Using Improved XGBoost Algorithm to Obtain Modified Atmospheric Refractive Index. *International Journal of Antennas and Propagation*, *2021*, 5506599. https://doi.org/10.1155/2021/5506599

Mulya, A. S., Si, M., Hermawan, F., & Evienia, B. P. (2019). Feasibility analysis of business; Case study in Indonesia minimarket. *International Journal of Recent Technology and Engineering*, *8*(2 Special Issue 4), 790–795. https://doi.org/10.35940/ijrte.B1159.0782S419

Pourdaryaei, A., Mokhlis, H., Illias, H. A., Kaboli, S. H. A., Ahmad, S., & Ang, S. P. (2019). Hybrid ANN and Artificial Cooperative Search Algorithm to Forecast Short-Term Electricity Price in De-Regulated Electricity Market. *IEEE Access*, *7*, 125369–125386. https://doi.org/10.1109/ACCESS.2019.2938842

Putatunda, S., & Rama, K. (2019). A Modified Bayesian Optimization based Hyper-Parameter Tuning Approach for Extreme Gradient Boosting. *2019 Fifteenth International Conference on Information Processing (ICINPRO)*, 1–6. https://doi.org/10.1109/ICInPro47689.2019.909 2025

Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2019). A review of deep learning for renewable energy forecasting. *Energy Conversion and Management*, *198*, 111799. https://doi.org/https://doi.org/10.1016/j.enconman.2019.111799

Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Communications*, *17*(3), 205–221. https://doi.org/10.23919/JCC.2020.03.017

Wu, D., Guo, P., & Wang, P. (2020). Malware Detection based on Cascading XGBoost and Cost Sensitive. *2020 International Conference on Computer Communication and Network*

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Security (CCNS), 201–205. https://doi.org/10.1109/CCNS50731.2020.00051

Xia, X., Jiang, S., Zhou, N., Li, X., & Wang, L. (2018). Genetic algorithm hyper-parameter optimization using Taguchi design for groundwater pollution source identification. *Water Supply*, *19*(1), 137–146. https://doi.org/10.2166/ws.2018.059

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/https://doi.org/10.1016/j.eswa.2017.02.017

Zhang, Y., Zhang, B., & Wu, Z. (2020). Multi-Model Modeling of CFB Boiler Bed Temperature System Based on Principal Component Analysis. *IEEE Access*, *8*, 389–399. https://doi.org/10.1109/ACCESS.2019.2961414